

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

Nguyễn Tiến Thanh

**TRÍCH CHỌN QUAN HỆ THỰC THỂ TRÊN
WIKIPEDIA TIẾNG VIỆT DỰA VÀO
CÂY PHÂN TÍCH CÚ PHÁP**

KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ thông tin

HÀ NỘI - 2010

**ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

Nguyễn Tiến Thanh

**TRÍCH CHỌN QUAN HỆ THỰC THỂ TRÊN
WIKIPEDIA TIẾNG VIỆT DỰA VÀO
CÂY PHÂN TÍCH CÚ PHÁP**

KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ thông tin

Cán bộ hướng dẫn: PGS.TS. Hà Quang Thụy

Cán bộ đồng hướng dẫn: ThS. Nguyễn Thu Trang

HÀ NỘI - 2010

LỜI CẢM ƠN

Lời đầu tiên, tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới PGS.TS Hà Quang Thụy, ThS. Nguyễn Thu Trang và CN. Trần Nam Khánh đã tận tình hướng dẫn tôi trong suốt quá trình thực hiện khoá luận tốt nghiệp.

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

Tôi cũng xin gửi lời cảm ơn tới ThS. Trần Mai Vũ và các anh chị, các bạn sinh viên tại phòng thí nghiệm KT-Sislab đã giúp tôi rất nhiều trong việc thu thập và xử lý dữ liệu. Tôi xin gửi lời cảm ơn tới các bạn trong lớp K51CA và K51CHTTT đã ủng hộ khuyến khích tôi trong suốt quá trình học tập tại trường.

Cuối cùng, tôi muốn được gửi lời cảm ơn 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 !

Hà Nội, ngày 21 tháng 05 năm 2010

Sinh viên

Nguyễn Tiến Thanh

Tóm tắt

Trích chọn *quan hệ ngữ nghĩa* (gọi tắt là “*quan hệ*”) được xem là bài toán cơ bản của xử lý ngôn ngữ tự nhiên nhận được sự quan tâm rất lớn từ các nhà nghiên cứu, các hội nghị lớn trên thế giới[1, 9, 41]. Tại Việt Nam, bài toán này vẫn đặt ra rất nhiều thách thức do tính phức tạp của ngôn ngữ tiếng Việt và sự không đầy đủ của các tài nguyên ngôn ngữ học.

Trên cơ sở phân tích ưu và nhược điểm của các phương pháp trích chọn quan hệ, khóa luận áp dụng phương pháp trích chọn quan hệ dựa trên đặc trưng để giải quyết bài toán này. Các đặc trưng biểu thị quan hệ được trích chọn dựa trên cây phân tích cú pháp tiếng Việt, sau đó được đưa vào bộ phân lớp SVM tìm được loại quan hệ tương ứng, từ đó trích chọn được các thể hiện của quan hệ. Hơn nữa, nhằm giảm công sức cho giai đoạn xây dựng tập dữ liệu học, khóa luận khai thác tính giàu cấu trúc của dữ liệu trên Wikipedia tiếng Việt để xây dựng tập dữ liệu học bán tự động.

Kết quả thực nghiệm trên một số loại quan hệ ban đầu cho thấy mô hình trích chọn của hệ thống cho độ đo F_1 đạt trung bình 86,4%. Điều này khẳng định mô hình là khả quan, có khả năng ứng dụng trong thực tế.

MỤC LỤC

Lời cảm ơn	i
Tóm tắt	ii
Mục lục	iii
Danh sách các bảng.....	v
Danh sách các hình vẽ.....	vi
Danh sách các từ viết tắt.....	vii
Mở đầu	1
Chương 1. Khái quát về bài toán trích chọn ngữ nghĩa	3
1.1. Quan hệ ngữ nghĩa	3
1.1.1. Khái niệm	3
1.1.2. Phân loại quan hệ ngữ nghĩa	3
1.2. Bài toán trích chọn quan hệ ngữ nghĩa	7
1.3. Ứng dụng	8
Tóm tắt chương một	9
Chương 2. Một số hướng tiếp cận trích chọn quan hệ ngữ nghĩa.....	10
2.1. Học không giám sát trích chọn quan hệ.....	10
2.2. Học có giám sát trích chọn quan hệ	13
2.2.1. Phương pháp Link grammar.....	13
2.2.2. Phương pháp trích chọn dựa trên các đặc trưng.....	16
2.2.3. Phương pháp trích chọn dựa trên hàm nhân	21
2.3. Học bán giám sát trích chọn quan hệ.....	24
2.3.1. Phương pháp DIRPE.....	24
2.3.2. Phương pháp Snowball	27
2.4. Nhận xét.....	29
Tóm tắt chương hai.....	29
Chương 3. Mô hình trích chọn quan hệ trên Wikipedia tiếng Việt dựa vào cây phân tích cú pháp.....	30
3.1. Đặc trưng của Wikipedia.....	30
3.1.1. Thực thể trong Wikipedia	30
3.1.2. Infobox	31
3.1.3. Mục phân loại.....	31
3.2. Cây phân tích cú pháp tiếng Việt.....	32
3.2.1. Phân tích cú pháp.....	32

3.2.2.	Một số thành phần cơ bản của cây phân tích cú pháp tiếng Việt.....	32
3.3.	Mô hình trích chọn quan hệ dựa trên cây phân tích cú pháp trên Wikipedia tiếng Việt.....	33
3.3.1.	Phát biểu bài toán.....	33
3.3.2.	Ý tưởng giải quyết bài toán.....	33
3.3.3.	Xây dựng tập dữ liệu học.....	34
3.3.4.	Mô hình hệ thống trích chọn quan hệ.....	36
	Tổng kết chương ba.....	40
Chương 4.	Thực nghiệm và đánh giá kết quả.....	41
4.1.	Môi trường thực nghiệm.....	41
4.1.1.	Câu hình phần cứng.....	41
4.1.2.	Công cụ phần mềm.....	41
4.2.	Dữ liệu thực nghiệm.....	42
4.3.	Thực nghiệm.....	42
4.3.1.	Mô tả cài đặt chương trình.....	42
4.3.2.	Xây dựng tập dữ liệu học dựa trên Wikipedia tiếng Việt.....	42
4.3.3.	Sinh vector đặc trưng.....	45
4.3.4.	Bộ phân lớp SVM.....	47
4.4.	Đánh giá.....	48
4.4.1.	Đánh giá hệ thống.....	48
4.4.2.	Phương pháp đánh giá.....	49
4.4.3.	Kết quả kiểm thử.....	49
4.5.	Nhận xét.....	51
Kết luận	52
Phục lục	53
Tài liệu tham khảo.....		56

Danh sách các bảng

Bảng 1-1 : 15 quan hệ trong Wordnet	4
Bảng 1-2: 22 loại quan hệ ngữ nghĩa theo Roxana Girju	5
Bảng 2-1: Đường đi ngắn nhất.....	23
Bảng 2-2: Một số đặc trưng thu được từ đường đi phụ thuộc	23
Bảng 3-1: Các thuộc tính của vector đặc trưng.....	39
Bảng 4-1: Cấu hình phần cứng.....	41
Bảng 4-2: Danh sách các phần mềm sử dụng	41
Bảng 4-3 : Các giá trị đánh giá hệ thống phân lớp.....	49
Bảng 5-1: Bảng các nhãn được sử dụng trong cây phân tích cú pháp	53

Danh sách các hình vẽ

Hình 1: Ví dụ về đường liên kết (1)	14
Hình 2: Ví dụ về đường liên kết (2)	14
Hình 3: Ví dụ về <i>mẫu</i>	14
Hình 4: Ví dụ về cặp thực thể sinh bởi quá trình <i>khớp mẫu</i>	14
Hình 5: Ví dụ về cây phân tích cú pháp.....	21
Hình 6: Các đặc trưng thu được từ cây phân tích cú pháp	21
Hình 7: Minh họa đồ thị phụ thuộc	22
Hình 8: Các quan hệ mẫu trích chọn được.....	26
Hình 9: Kiến trúc của hệ thống Snowball.....	27
Hình 10: Ví dụ về cây phân tích cú pháp tiếng Việt	32
Hình 11: Quá trình xây dựng tập dữ liệu học	34
Hình 12: Cấu trúc biểu diễn của thông tin của infobox.....	35
Hình 13: Mô hình trích chọn quan hệ trên Wikipedia.....	36
Hình 14: Cây con biểu diễn quan hệ “thành_lập”	38
Hình 15: Ví dụ về tìm kiếm trên Wikipedia	44
Hình 16 : Bảng thống kê dữ liệu học của quan hệ “ngày sinh”	48
Hình 17: Kết quả kiểm thử đối với quan hệ “năm thành lập”	50
Hình 18: Kết quả kiểm thử đối với quan hệ “hiệu trưởng”	50
Hình 19: Kết quả kiểm thử đối với quan hệ “ngày sinh”	51
Hình 20: So sánh kết quả trung bình của ba quan hệ	51

Danh sách các từ viết tắt

Từ hoặc cụm từ	Viết tắt
A Library for Support Vector Machines	LibSVM
Dual Iterative Pattern Relation Expansion	DIPRE
Support vector machine	SVM
Wikipedia	Wiki

Mở đầu

Trích chọn *quan hệ ngữ nghĩa* (hay *quan hệ*) được xem là bài toán cơ bản của xử lý ngôn ngữ tự nhiên, thực hiện nhiệm vụ trích chọn quan hệ giữa các khái niệm về mặt ngữ nghĩa hoặc dựa vào quan hệ xác định trước nhằm tìm kiếm những thông tin phục vụ cho quá trình xử lý khác. Trích chọn quan hệ được ứng dụng nhiều cho các bài toán như: xây dựng Ontology[15, 16, 19, 22], hệ thống hỏi đáp [22,29], phát hiện ảnh qua đoạn văn bản [11], tìm mối liên hệ giữa bệnh-genes [27],... Vì thế, trích chọn quan hệ không những 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: Coling/ACL, Senseval,... mà còn là một phần trong các dự án quan trọng mang tầm cỡ quốc tế trong lĩnh vực khai phá dữ liệu như: ACE (Automatic Content Extraction), DARPA EELD (Evidence Extraction and Link Discovery), ARDA-AQUAINT (Question Answering for Intelligence), ARDA NIMD (Novel Intelligence from Massive Data).

Tại Việt Nam, bài toán này vẫn đặt ra rất nhiều thách thức do tính phức tạp của ngôn ngữ tiếng Việt và sự không đầy đủ của các tài nguyên ngôn ngữ học. Trên cơ sở phân tích các phương pháp trích chọn quan hệ, khóa luận đã đưa ra mô hình học có giám sát *trích chọn quan hệ thực thể dựa vào cây phân tích cú pháp* trên miền dữ liệu Wikipedia tiếng Việt. Kết quả thực nghiệm bước đầu cho thấy mô hình là khả quan và có khả năng ứng dụng tốt.

Nội dung của khóa luận được bố cục gồm có 4 chương:

Chương 1: Giới thiệu khái quát về bài toán trích chọn quan hệ ngữ nghĩa cũng như các khái niệm liên quan.

Chương 2: Giới thiệu các phương pháp tiếp cận giải quyết bài toán trích chọn quan hệ. Với mỗi phương pháp học máy: có giám sát, không giám sát và bán giám sát, khóa luận giới thiệu một số mô hình tiêu biểu. Đây là cơ sở phương pháp luận quan trọng để khóa luận đưa ra mô hình áp dụng đối với bài toán trích chọn quan hệ trên miền dữ liệu Wikipedia tiếng Việt.

Chương 3: Trên cơ sở phân tích ưu và nhược điểm của các phương pháp được trình bày ở chương 2, khóa luận đã lựa chọn phương pháp trích chọn quan hệ dựa trên đặc trưng theo tiếp cận học có giám sát để giải quyết bài toán này. Các đặc trưng của quan hệ được trích chọn dựa trên cây phân tích cú pháp tiếng Việt, sau đó được đưa vào bộ phân lớp sử dụng thuật toán SVM, tìm được loại quan hệ tương

ứng, từ đó trích chọn được các thể hiện của quan hệ. Hơn nữa, để giảm công sức cho giai đoạn xây dựng tập dữ liệu học, các đặc trưng biểu diễn dữ liệu giàu cấu trúc trên Wikipedia tiếng Việt đã được sử dụng. Nội dung chính của chương này trình bày các đặc trưng của Wikipedia, cây phân tích cú pháp tiếng Việt và đề xuất một mô hình trích chọn quan hệ dựa trên cây phân tích cú pháp.

Chương 4: Thực nghiệm, kết quả và đánh giá. Tiến hành thực nghiệm việc xây dựng tập dữ liệu học, thực nghiệm trích chọn quan hệ sử dụng bộ phân lớp SVM.

Phần kết luận và định hướng phát triển khoá luận: Tóm lược những nội dung chính đạt được của khóa luận đồng thời cũng chỉ ra những điểm cần khắc phục và đưa ra những định hướng nghiên cứu trong thời gian sắp tới.

Chương 1. Khái quát về bài toán trích chọn ngữ nghĩa

Nội dung chính của khóa luận là đề xuất một mô hình trích chọn quan hệ thực thể dựa trên cây phân tích cú pháp trên miền dữ liệu Wikipedia tiếng Việt. Chương này sẽ giới thiệu các khái niệm về quan hệ ngữ nghĩa, bài toán trích chọn quan hệ ngữ nghĩa và những ứng dụng của bài toán này. Đây là cơ sở lý thuyết quan trọng cho việc xác định mục tiêu cũng như phạm vi giải quyết của mô hình đề xuất.

1.1. Quan hệ ngữ nghĩa

1.1.1. Khái niệm

Xác định quan hệ ngữ nghĩa (semantic relation) là một lĩnh vực nghĩa nhận được nhiều sự quan tâm từ các nhà nghiên cứu về ngôn ngữ học cũng như xử lý ngôn ngữ tự nhiên. Có nhiều định nghĩa về quan hệ ngữ nghĩa đã được đưa ra. Theo nghĩa hẹp, Birger Hjørland [42] đã định nghĩa quan hệ ngữ nghĩa:

“Quan hệ ngữ nghĩa 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 từ.”

Ví dụ: Ta có câu “**Trường Đại học Công nghệ** được Thủ tướng chính phủ quyết định thành lập **ngày 25 tháng 5 năm 2004**.” Khi đó, ta nói: (“*Trường Đại học Công nghệ*”, “*ngày 25 tháng 5 năm 2004*”) có quan hệ ngữ nghĩa là “*ngày thành lập*”.

Trong khóa luận này, trong trường hợp không gây nhầm lẫn, khái niệm *quan hệ ngữ nghĩa* được gọi tắt là *quan hệ*.

Việc xác định quan hệ giữa các khái niệm là một vấn đề quan trọng trong tìm kiếm thông tin. Điều này sẽ làm tăng tính ngữ nghĩa cho câu hay tập tài liệu. Đồng thời, khi tìm kiếm một thông tin nào đó, ta có thể nhận 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 quan hệ và tìm hiểu các phương pháp để xác định được các quan hệ đó.

1.1.2. Phân loại quan hệ ngữ nghĩa

Quan hệ ngữ nghĩa thể hiện quan hệ giữa các khái niệm và được biểu diễn dưới dạng cấu trúc phân cấp thông qua các quan hệ. Trong [17], Iris Hendrickx và cộng sự đã tổng kết và chỉ ra rằng phân loại quan hệ ngữ nghĩa là rất đa dạng, phụ thuộc vào những đặc trưng ngữ nghĩa cũng như mục đích và đối tượng tiếp cận. Mục này sẽ giới thiệu hai hệ thống phân loại quan hệ ngữ nghĩa được sử dụng khá

phổ biến trong bài toán trích chọn quan hệ đó là WordNet và hệ thống phân loại của Girju.

WordNet [16, 39] là một từ điển trực tuyến trong Tiếng Anh, được phát triển bởi các nhà từ điển học thuộc trường đại học Princeton (Mỹ). WordNet bao gồm 100.000 khái niệm bao gồm danh từ, động từ, tính từ, phó từ liên kết với nhau thông qua 15 quan hệ (được mô tả trong bảng 1-1)

Bảng 1-1 : 15 quan hệ trong Wordnet

STT	Quan hệ ngữ nghĩa	Các khái niệm được liên kết bởi quan hệ ngữ nghĩa	Ví dụ
1.	Hypernymy (is - a)	Danh từ - Danh từ Động từ - Động từ	Cat is-a feline Manufacture is-a make
2.	Hyponymy (reverse is-a)	Danh từ - Danh từ Động từ - Động từ	Feline reverse is-a cat Manufacture reverse is-a mak
3.	Is-part- of	Danh từ - Danh từ	Leg is-part-of table
4.	Has-part	Danh từ - Danh từ	Table has-part leg
5.	Is-member-of	Danh từ - Danh từ	UK is-member-of NATO
6.	Has-member	Danh từ - Danh từ	NATO has-member UK
7.	Is-suff-of	Danh từ - Danh từ	Carbon is-stuff-of coal
8.	Has-stuff	Danh từ - Danh từ	Coal has-stuff carbon
9.	Cause-to	Động từ - Động từ	To develop cause-to to grow
10.	Entail	Động từ - Động từ	To snore entail to sleep
11.	Atribute	Tính từ - Danh từ	Hot attribute temperature
12.	Synonymy (synset)	Danh từ - Danh từ Động từ - Động từ Tính từ - Tính từ Phó từ - Phó từ	Car synonym automobile To notice synonym to observe Happy synonym content Mainly synonym primarily

13.	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
14.	Similarity	Tính từ - Tính từ	Abridge similarity shorten
15.	See-also	Động từ - Động từ Tính từ - Tính từ	Touch see-also touch down Inadequate see-also insatisfactory

Thông thường, người ta hay sử dụng WordNet vào việc tìm kiếm các quan hệ ngữ nghĩa. Đồng thời, dựa vào các quan hệ này, một từ trong WordNet có thể tìm được các liên hệ với các khái niệm khác.

Roxana Girju [10] đã đưa ra hệ thống các quan hệ ngữ nghĩa gồm 22 loại như trong bảng 1-2, trong đó một số quan hệ ngữ nghĩa quan trọng thường được dùng để thể hiện quan hệ giữa các khái niệm như: hyponymy/ hypernymy (is - a), meronymy/holonym (part - whole), đồng nghĩa (synonymy) và trái nghĩa (antonymy).

Bảng 1-2: 22 loại quan hệ ngữ nghĩa theo Roxana Girju

STT	Quan hệ ngữ nghĩa	Mô tả	Ví dụ
1.	HYPERNYMY (IS-A)	Một thực thể/ sự kiện/ trạng thái là lớp con của một thực thể/ sự kiện/ trạng thái khác	daisy flower; large company, such as Microsoft
2.	PART-WHOLE (MERONYMY)	Một thực thể/ sự kiện/ trạng thái là một bộ phận của thực thể/ sự kiện/ trạng thái khác	door knob; the door of the car
3.	CAUSE	Một sự kiện/trạng thái là nguyên nhân cho một sự kiện/trạng thái khác xảy ra	malaria mosquitos; “death by hunger”; “The earthquake

			generated a big Tsunami”
4.	INSTRUMENT	Một thực thể được sử dụng như là một phương tiện/công cụ	pump drainage; He broke the box with a hammer.
5.	MAKE / PRODUCE	Một thực thể tạo ra/ sản xuất ra một thực thể khác	honey bees; GM makes cars
6.	KINSHIP (thân thích)	Một thực thể có liên quan tới thực thể khác bởi quan hệ huyết thống, hôn nhân	boy’s sister; Mary has a daughter
7.	POSSESSION (sở hữu)	Một thực thể sở hữu thực thể khác	family estate; the girl has a new car.
8.	SOURCE / FROM	Xuất xứ của thực thể	olive oil
9.	PURPOSE	Một trạng thái hay dành động là kết quả từ một trạng thái hay sự kiện khác	migraine drug; He was quiet in order not to disturb her.
10.	LOCATION/SPACE	quan hệ đặc biệt giữa hai thực thể hoặc giữa thực thể và sự kiện	field mouse; I left the keys in the car
11.	TEMPORAL	Thời gian liên quan tới một sự kiện	5-O’ clock tea; the store opens at 9 am
12.	EXPERIENCER	Cảm giác hay trạng thái của một thực thể	desire for chocolate; Mary’s fear.
13.	MEANS	Phương tiện mà một sự kiện được thực hiện	bus service; I go to school by bus.
14.	MANNER	Cách thức mà một sự kiện xảy ra	hard-working immigrants; performance with

			passion
15.	TOPIC	Một đối tượng là đặc trưng của đối tượng khác	they argued about politics
16.	BENEFICIARY	Một thực thể hưởng lợi ích từ một trạng thái hay sự kiện	customer service; I wrote Mary a letter.
17.	PROPERTY	Thuộc tính của một thực thể/sự kiện hay trạng thái	red rose; the juice has a funny color.
18.	THEME	Một thực thể được mô tả theo/ trong một hành động hay sự kiện khác	music lover
19.	AGENT	Tác nhân thực hiện hành động	the investigation of the police
20.	DEPICTION-DEPICTED	Một thực thể được biểu diễn trong một thực thể khác	the picture of the girl
21.	TYPE	Một từ hay khái niệm là kiểu của một từ hay hay khái niệm khác	member state; framework law
22.	MEASURE	Một thực thể biểu diễn số lượng của một thực thể/sự kiện nào đó	70-km distance; The jacket costs \$60; a cup of sugar

1.2. Bài toán trích chọn quan hệ ngữ nghĩa

Theo [9, 36, 41], trích chọn quan hệ được xem là một bộ phận quan trọng của trích chọn thông tin. Tập các câu hay các văn khi xem xét ở mức trừu tượng cao thì đây chính là tập hợp các khái niệm, các thực thể và quan hệ giữa chúng. Các thực thể hay khái niệm được thể hiện dưới dạng các từ hay cụm từ. Quan hệ ngữ nghĩa giữa chúng được ẩn trong các liên kết giữa các khái niệm hay thực thể này. Việc phát hiện ra các quan hệ này có ý nghĩa rất quan trọng trong các bài toán xử lý ngôn ngữ tự nhiên.

Roxana Girju [10] đã phát biểu bài toán trích chọn quan hệ ngữ nghĩa 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 quan hệ ngữ nghĩa giữa chúng”

Một ví dụ về trích chọn quan hệ ngữ nghĩa được Roxana Girju [10] đưa ra như sau:

Cho một đoạn văn bản với các thực thể/khái niệm được gán nhãn:

[Saturday’s snowfall]_{TEMP} topped **[a record in Hartford, Connecticut]**_{LOC} with **[the total of 12/5 inches]**_{MEASURE}, **[the weather service]**_{TOPIC} said. The storm claimed its fatality Thursday when **[a car driven by a [college student]**_{PART-WHOLE}_{THEME} skidded on **[an interstate overpass]**_{LOC} in **[the mountains of Virginia]**_{LOC/PART-WHOLE} and hit **[a concrete barrier]**_{PART-WHOLE}, police said.

Khi đó, hệ thống trích chọn quan hệ ngữ nghĩa sẽ cho kết quả là các quan hệ có thể có giữa các thực thể/khái niệm này, cụ thể như sau:

TEMP (Saturday, snowfall)	LOC (mountains, Virginia)
PART-WHOLE/LOC (mountains, Virginia)	LOC (Hartford Connecticut, record)
PART-WHOLE (concrete, barrier)	LOC (interstate, overpass)
PART-WHOLE (student, college)	TOPIC (weather, service)
THEME (car, driven by a college student)	MEASURE(total, 12.5 inches)

1.3. Ứng dụng

Trích chọn quan hệ ngữ nghĩa được ứng dụng trong nhiều lĩnh vực khác nhau. Lĩnh vực đầu tiên phải nhắc tới là việc xây dựng cơ sở tri thức mà điển hình là xây dựng Ontology – thành phần nhân của Web ngữ nghĩa. Trong khi những lợi ích mà Web ngữ nghĩa đem lại là rất lớn thì việc xây dựng các ontology một cách thủ công lại hết sức khó khăn. Giải pháp cho vấn đề này chính là kỹ thuật trích chọn thông tin nói chung và trích chọn quan hệ nói riêng để tự động hóa một phần quá trình xây dựng các ontology. Đã có nhiều các nghiên cứu liên quan tới vấn đề này như [15, 16, 19, 22]

Trích chọn mối quan hệ ngữ nghĩa cũng được sử dụng nhiều trong các hệ thống hỏi đáp. Một số hệ thống hỏi đáp đã được xây dựng dựa vào việc trích xuất tự động các từ, khái niệm và mối quan hệ. Chẳng hạn Kim và cộng sự [22] cũng đưa ra

một hệ thống hỏi đáp OntotrileQA sử dụng kỹ thuật trích chọn quan hệ ngữ nghĩa cho các thực thể trên ontoloty đã được gán nhãn bằng tay.

Ngoài ra, trích chọn quan hệ còn có ứng dụng trong các lĩnh vực xử lý ảnh như phát hiện ảnh qua đoạn văn bản (text-to-image generation) [11] . Trích chọn quan hệ cũng là một công cụ đặc lực tron lĩnh vực công nghệ sinh học như tìm quan hệ bệnh tật - Genes, ảnh hưởng qua lại giữa protein-protein (Protein-Protein interaction)[27]...

Tóm tắt chương một

Trong chương này, khoá luận đã giới thiệu khái quát các khái niệm liên quan tới bài toán trích chọn quan hệ ngữ nghĩa, một số loại quan hệ ngữ nghĩa và những ứng dụng nổi bật. Trong chương tiếp theo, khoá luận sẽ tập trung làm rõ các phương pháp diễn hình mô hình hóa bài toán trích chọn quan hệ ngữ nghĩa và cách giải quyết tương ứng.

Chương 2. Một số hướng tiếp cận trích chọn quan hệ ngữ nghĩa

Trích chọn quan hệ được xem là một phần quan trọng của trích chọn thông tin [9], nhận được sự quan tâm ngày càng nhiều hơn của cộng đồng xử lý ngôn ngữ tự nhiên và học máy. Các tiếp cận giải quyết bài toán hiện nay tập trung vào sử dụng các phương pháp học máy để tiến hành trích chọn tự động. Cả ba loại học máy là học không giám sát, học có giám sát và học bán giám sát đều thể hiện được những ưu điểm riêng của mình.

Hơn nữa, trong các nghiên cứu gần đây [8, 12, 13, 17, 21], cây phân tích cú pháp của câu được xem là một thông tin quan trọng cho trích chọn quan hệ. Do đó, trong chương này, với mỗi phương pháp học máy, khóa luận sẽ giới thiệu một số mô hình tiêu biểu. Đây là cơ sở phương pháp luận quan trọng để khóa luận đưa ra mô hình áp dụng đối với bài toán trích chọn quan hệ trên miền dữ liệu Wikipedia tiếng Việt.

2.1. Học không giám sát trích chọn quan hệ

Học không giám sát có bản chất là sử dụng các thuật toán phân cụm các quan hệ để mô hình hóa. Có nhiều cách khác nhau [1, 7, 12, 18] để biểu diễn quan hệ giữa hai thực thể/khái niệm, trong đó phổ biến nhất là biểu diễn quan hệ này dưới dạng vector đặc trưng. Vấn đề cốt lõi là làm thế nào để lựa chọn được các đặc trưng tốt và hiệu quả. Một giải pháp đã được Jinxiu Chen và cộng sự [18] đưa ra dựa trên ý tưởng xây dựng hàm Entropy để xếp hạng các đặc trưng, từ đó, đưa một thuật toán lựa chọn được đặc trưng và số cụm tối ưu nhất. Cụ thể như sau:

Đầu tiên, Jinxiu Chen và cộng sự đưa ra một số khái niệm:

Gọi $P = \{p_1, p_2, \dots, p_N\}$ là tập tất cả các vector *ngữ cảnh* mà đồng thời xuất hiện cặp thực thể E_1 và E_2 . Ở đây, *ngữ cảnh* bao gồm tất cả các từ xuất hiện trước, ở giữa và sau cặp thực thể.

Gọi $W = \{w_1, w_2, \dots, w_M\}$ là tập các đặc trưng, bao gồm tất cả các từ xuất hiện trong P .

Giả sử, p_n ($1 \leq n \leq N$) thuộc không gian đặc trưng W (chiều của W là M). Độ tương đồng giữa vector p_i và p_j được cho bởi công thức:

$$S_{i,j} = \exp(-\alpha * D_{i,j}) \text{ trong đó:}$$

- $D_{i,j}$ là độ đo Oclit giữa p_i và p_j ,

- $\alpha = -\frac{\ln 0.5}{D}$ là hằng số dương thu được bằng thực nghiệm
- \bar{D} là khoảng cách trung bình giữa các p_i

Khi đó, *entropy* của tập dữ liệu P với N điểm dữ liệu được định nghĩa là:

$$E = -\sum_{i=1}^N \sum_{j=1}^N (S_{i,j} \log S_{i,j} + (1 - S_{i,j}) \log(1 - S_{i,j})) \quad (2.1)$$

Sau đó, để lựa chọn một tập con các đặc trưng quan trọng từ W , các đặc trưng được xếp hạng theo độ quan trọng của chúng theo cụm. Hàm xếp hạng các đặc trưng dựa trên một giả thiết rằng “*một đặc trưng là không quan trọng nếu nó xuất hiện trong tập dữ liệu có thể tách rời*” [18]. Độ quan trọng của mỗi đặc trưng $I(w_k)$ được xác định bởi *entropy* của tập dữ liệu sau khi loại bỏ đi đặc trưng w_k .

Dựa trên nhận xét rằng: “*một đặc trưng là kém quan trọng nhất nếu sau khi loại bỏ nó đi sẽ làm cho E đạt giá trị nhỏ nhất*”, các đặc trưng được sắp xếp theo độ quan trọng của chúng, ta thu được tập $W_r = \{f_1, \dots, f_M\}$.

Khi đó, việc tìm tập con đặc trưng tốt nhất F sẽ trở thành bài toán tìm kiếm trên không gian $\{(f_1, \dots, f_k), 1 \leq k \leq M\}$: tức là tìm $F_k = \arg \max_{F \subseteq W_r} \{criterion(F, k)\}$

Gọi P^μ là tập con các cặp thực thể được lấy mẫu từ tập các cặp thực thể đầy đủ P . Kích thước của P^μ là αN (với $\alpha = 0.9$)

Gọi C (hay C^μ) là ma trận kết nối có kích thước $|P| * |P|$ (hay $|P^\mu| * |P^\mu|$) dựa trên các kết quả phân cụm tương ứng từ P (hay P^μ) trong đó:

$$c_{ij} = \begin{cases} 1 & \text{nếu như cặp thực thể } p_i \text{ và } p_j \text{ nằm trong cùng một cụm} \\ 0 & \text{trong trường hợp ngược lại} \end{cases}$$

Khi đó, độ ổn định $M(C^\mu, C)$ (là độ nhất quán giữa kết quả phân cụm trên C^μ và C) sẽ được tính theo công thức:

$$M(C^\mu, C) = \frac{\sum_{i,j} 1\{C_{i,j}^\mu = C_{i,j} = 1, p_i \in P^\mu, p_j \in P^\mu\}}{\sum_{i,j} 1\{C_{i,j} = 1, p_i \in P^\mu, p_j \in P^\mu\}} \quad (2.2)$$

Tuy nhiên, vì $M(C^\mu, C)$ có chiều hướng giảm khi số cụm k tăng nên để tránh trường hợp giá trị k nhỏ sẽ được lựa chọn làm số cụm, biến ngẫu nhiên độc lập ρ_k

được sử dụng để chuẩn hóa $M(C^\mu, C)$. Biến ngẫu nhiên độc lập này có được bằng cách với mỗi giá trị k , thực hiện q lần việc tách dữ liệu vào k cụm một cách ngẫu nhiên. Khi đó, hàm mục tiêu $M(C_{F,k}^\mu, C_{F,k})$ sẽ được tính theo công thức (2.2) và:

$$M_{F,k}^{norm} = \frac{1}{q} \sum_{i=1}^q M(C_{F,k}^{\mu_i}, C_{F,k}) - \frac{1}{q} \sum_{i=1}^q M(C_{F,\rho_k}^{\mu_i}, C_{F,\rho_k}) \quad (2.3)$$

Hàm này được thực hiện theo 8 bước sau:

Hàm: criterion(F, k, P, q)

Đầu vào: tập con đặc trưng F , số cụm k , tập các cặp thực thể P và tần xuất lấy mẫu q

Đầu ra: Điểm đánh giá chất lượng của F và k

Xử lý:

1. Thực hiện thuật toán k-means với k cụm theo như input trên các tập các cặp P^F
2. Khởi tạo ma trận kết nối $C_{F,k}$ dựa trên kết quả phân cụm ở trên
3. Sử dụng biến độc lập ngẫu nhiên ρ_k để gán nhãn cho từng cặp trong P^F
4. Khởi tạo ma trận kết nối C_{F,ρ_k} cho tất cả các P^F
5. Khởi tạo q tập con của tập các cặp thực thể đầy đủ bằng cách lựa chọn ngẫu nhiên αN trong số N cặp ban đầu ($0 \leq \alpha \leq 1$)
6. Với mỗi tập con, thực hiện phân cụm như trong các bước 2, 3, 4 và cho ra kết quả $C_{F,k}^\mu, C_{F,\rho_k}$
7. Tính $M_{F,k}$ để đánh giá chất lượng của k thông qua công thức 2.3
8. Trả về kết quả $M_{F,k}$

Cuối cùng, mô hình thuật toán lựa chọn (Model Selection Algorithm) cho trích chọn quan hệ:

Đầu vào: Tập dữ liệu D với các thực thể được gán nhãn (E_1, E_2)

Đầu ra: Tập con các đặc trưng và số lượng kiểu quan hệ (Model Order)

Xử lý:

1. Tìm tất cả các *ngữ cảnh* của tất cả các cặp thực thể có trong tập D. Tập ngữ cảnh này đặt tên là P
2. Xếp hạng các đặc trưng dựa theo công thức (2.1)
3. Tính khoảng (K_l, K_h) : số các cụm quan hệ có thể có (thấp nhất tới cao nhất)
4. Thiết lập giá trị ước lượng số kiểu quan hệ $k = K_l$
5. Lựa chọn các đặc trưng theo thuật toán $\text{criterion}(F, k, P, q)$
6. Lưu giữ giá trị \hat{F}_k, k và điểm số chất lượng tương ứng là $M_{F,k}$
7. Nếu $k < K_h$ thì quay lại bước 5, không thì sang bước 8
8. Lựa chọn k và tập con đặc trưng \hat{F}_k có giá trị lớn nhất trong các giá trị $M_{F,k}$

2.2. Học có giám sát trích chọn quan hệ

Bài toán trích chọn quan hệ ngữ nghĩa giữa hai thực thể cũng được giải quyết bằng cách coi đây là bài toán phân lớp sử dụng phương pháp học máy. Các thể hiện của quan hệ được chuyển sang các một tập các đặc trưng f_1, f_2, \dots, f_N , tạo nên một vector đặc trưng N chiều. Trong quá trình học, các thuật toán phân lớp được áp dụng đối với các thực thể đầu vào để xác định lớp quan hệ của nó, từ đó trích chọn được quan hệ có thể có.

Theo G. Zhou và M. Zhang [32], các mô hình có thể được chia làm ba nội dung chính: Phương pháp dựa trên mô hình sinh, dựa vào hàm nhân (tree kernel) và phương pháp tiếp cận dựa vào đặc trưng.

2.2.1. Phương pháp Link grammar

Phương pháp này được các nhà nghiên cứu thuộc học viện Mac-Planck đưa ra năm 2006. Về nguyên tắc, có thể trích chọn được bất cứ quan hệ nào. Hệ thống đã thực nghiệm trên 3 quan hệ: *birthdate*, *synonymy*, *instanceOf*.

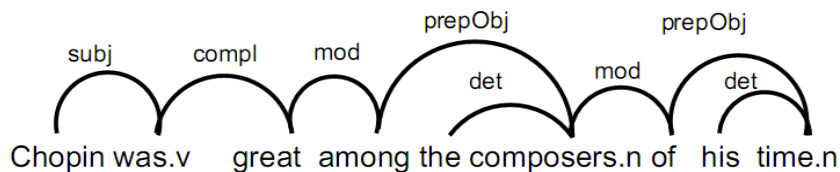
Trong phương pháp này đã sử dụng một số các khái niệm cơ bản về *linkgrammar* [12, 40] như sau:

Mỗi *đường liên kết* (linkage) là một đồ thị phẳng vô hướng, trong đó:

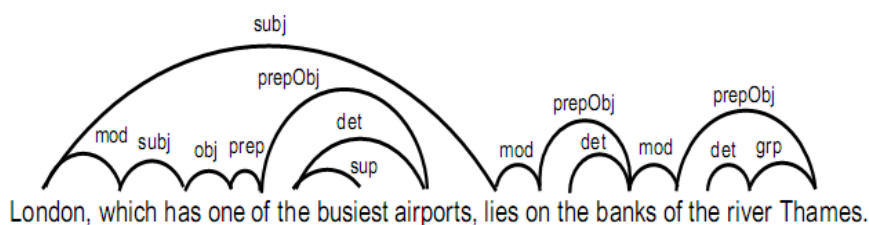
- Các nút của đồ thị này là các từ của câu.
- Cung nối giữa các nút gọi là *kết nối* (link).
- Nhãn của các cung này gọi là *loại kết nối* (connectors) – lấy từ một tập hữu hạn các kí hiệu.

Link grammar là một tập các luật quy định một từ sẽ kết nối với từ đứng sau hoặc trước nó bởi *loại kết nối* nào: $\langle \text{word} - \text{connectors} \rangle$ hoặc $\langle \text{connectors} - \text{word} \rangle$. Ví dụ: từ “was” trong hình 1 sẽ có $\langle \text{subj_link} - \text{“was”} \rangle$ và $\langle \text{“was”} - \text{compl_link} \rangle$

Mỗi *đường liên kết* của một câu được sinh ra bởi *link grammar*.



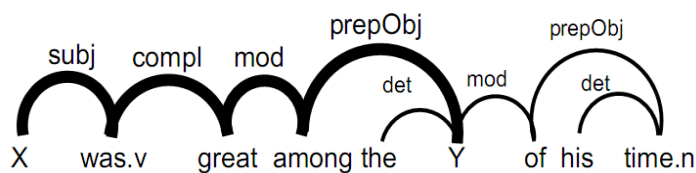
Hình 1: Ví dụ về đường liên kết (1)



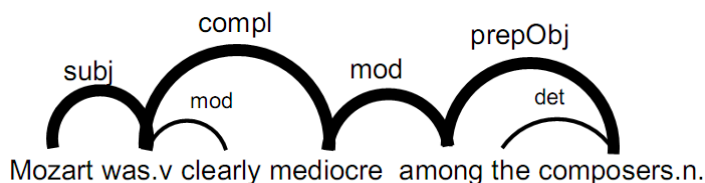
Hình 2: Ví dụ về đường liên kết (2)

Một *đường liên kết* biểu diễn một quan hệ R nếu câu mà *đường liên kết* mô tả chứa cặp thực thể nằm trong quan hệ R . Ví dụ: trong hình 2, thể hiện quan hệ *sở hữu*: “London” has an “airports”

Một *mẫu* là một *đường liên kết* mà trong đó hai từ (cụm từ) có thể được thay thế bởi một kí hiệu đại diện (placeholder). Ví dụ: trong hình 1, thay “Chopin” bởi X và “composers” bởi Y , ta được một mẫu như ở hình 3.



Hình 3: Ví dụ về *mẫu*



Hình 4: Ví dụ về cặp thực thể sinh bởi quá trình *khớp mẫu*

Đường đi ngắn nhất (duy nhất) từ một kí tự đại diện này tới kí tự đại diện kia được gọi là một *cầu* (bridge). (Đường in đậm trong hình 3). *Cầu* này không bao gồm các kí tự đại diện.

Một *mẫu* được gọi là *khớp* với một *đường liên kết* nếu *cầu* của *mẫu* xuất hiện trong *đường liên kết* (cho phép các danh từ hay tính từ là khác nhau)

Khi một *mẫu* khớp với một *đường liên kết*, ta nói *mẫu sinh ra* một cặp từ (cụm từ). Cặp từ này nằm ở vị trí của các kí tự đại diện tương ứng giữa link và mẫu. Ví dụ: ở hình ..., cặp “Mozart” và “composers” xuất hiện trong đường liên kết, nằm tương ứng với các kí tự đại diện **X** và **Y** trong mẫu ở hình 4. Ta nói, *mẫu sinh ra* cặp thực thể <“Mozart” - “composers”>.

Để tiến hành việc học, Fabian M. Suchanek và cộng sự [15] đã tiến hành phân loại các cặp từ, chia chúng làm 3 loại sau:

- Một cặp có thể là một **ví dụ** (example) cho quan hệ đích. Ví dụ: với quan hệ *birthdate*, các **ví dụ** là một danh sách *tên người* và *ngày sinh* của họ

<Frederic Chopin - 1810>

<Wolfgang Amadeus Mozart - 1756>

- Một cặp có thể là một **phản ví dụ** (counterExample) – là các cặp không thể nằm trong một quan hệ. Ví dụ, với quan hệ *birthdate*, các **phản ví dụ** có thể được suy diễn từ **ví dụ**. Nếu <“Chopin” - “1810”> là một **ví dụ** thì <“Chopin” - “2000”> hiển nhiên một **phản ví dụ**.
- Một cặp có thể là một **ứng viên** (candidate) có thể có cho quan hệ đích. Ví dụ, với quan hệ *birthdate*, chỉ các cặp có dạng <Tên riêng người – ngày> mới có thể là **ứng viên**.
- Một cặp có thể không thuộc vào 1 trong 3 loại trên.

Dựa trên các khai niệm này, hệ thống trích chọn quan hệ được đưa ra với 3 pha xử lý chính:

Pha 1: Pha nhận dạng (discovery phase): Xác định các mẫu biểu diễn quan hệ đích

- Trong tất cả các câu, tìm các *đường liên kết* mà các cặp **ví dụ** xuất hiện.
- Thay thế các cặp này bởi các kí tự đại diện → tạo ra các mẫu. Các mẫu thu được lúc này được gọi là *mẫu chắc chắn* (positive patterns)

Ví dụ: Khi có câu “Chopin was born in 1810”, thì mẫu “X was born in Y” sẽ được sinh ra

- Duyệt qua các câu một lần nữa, tìm tất cả các câu có *đường liên kết* khớp với *mẫu chắc chắn* mà các cặp thực thể sinh ra từ quá trình khớp này thuộc **phản ví dụ** thì tiến hành thay thế các cặp này bởi các kí tự đại diện, ta được các mẫu, gọi là *mẫu không chắc chắn* (negative patterns)

Ví dụ: Khi duyệt lại, tìm được câu "*Chopin was born in 2000*", có cặp <X – Y> là <*Chopin* - 2000> thuộc **phản ví dụ** thì mẫu "*X was born in Y*" sẽ được thu sẽ cho vào tập mẫu *mẫu không chắc chắn*

Pha2: Pha học (Training Phase): Tạo ra các *mẫu chắc chắn* nhờ mô hình học máy

- Mô hình học thống kê được áp dụng để học các khái niệm của các *mẫu chắc chắn* từ tập *mẫu chắc chắn* và *mẫu không chắc chắn*.
- Kết quả của pha này là bộ phân lớp cho các mẫu – *mẫu chắc chắn* hay là *mẫu không chắc chắn*.
- Sử dụng thuật toán phân lớp K-người hàng xóm gần nhất (kNN) hoặc SVM

Pha 3: Pha kiểm thử (Testing Phase):

- Với mỗi *đường liên kết*, tạo tất cả các *mẫu* có thể bằng cách thay thế cặp từ (cụm từ) tương ứng bởi các kí tự đại diện.
- Nếu cặp từ này có dạng **ứng viên** và mẫu được phân lớp là *mẫu chắc chắn* thì cặp từ này được chấp nhận như là phân tử mới của quan hệ đích.

2.2.2. Phương pháp trích chọn dựa trên các đặc trưng

Trong phương pháp này, vector đặc trưng thể hiện quan hệ ngữ nghĩa giữa hai thực thể M1 và M2 được xác định từ ngữ cảnh bao quanh các thực thể này. Theo Abdulrahman Almuhareb [4], các vector đặc trưng được chia làm hai loại chính: một là, đặc trưng dựa vào các *từ lân cận* của M1 và M2; hai là, đặc trưng dựa vào *quan hệ về mặt ngữ pháp* của M1 và M2. Nội dung của khóa luận này quan tâm tới loại đặc trưng thứ hai.

Trong loại này, thứ tự xuất hiện của các thực thể cũng được phân biệt, ví dụ M1 – Parent-Of – M2 thì khác với M2 – Parent-Of – M1. Với mỗi cặp thực thể, các thông tin về từ vựng, ngữ pháp và ngữ nghĩa sẽ được sử dụng như là các đặc trưng thể hiện cho quan hệ.

G. Zhou và M. Zang [32] đưa ra 8 loại đặc trưng thường được sử dụng trong phương pháp này:

Đặc trưng về từ: Tùy theo vị trí của từ mà chúng được phân chia làm 4 loại:

- Từ biểu diễn M1 và M2: Trong những từ này, từ trung tâm (head word) được coi là quan trọng hơn và mang nhiều ý nghĩa thông tin hơn. Từ trung tâm của M1(M2) là từ cuối cùng của cụm từ biểu diễn M1 (M2). Trong trường hợp có giới từ nằm trong cụm từ biểu diễn M1 (M2) thì từ trung tâm là từ cuối cùng trước khi gặp giới từ. Ví dụ, với một cụm từ biểu diễn M1 là “University of Michigan” thì từ trung tâm ở đây là “University”.
- Từ nằm giữa M1 và M2: Các từ này được chia làm 3 loại:
 - Từ đầu tiên nằm ở giữa
 - Từ cuối cùng nằm ở giữa
 - Và các từ còn lại
- Từ nằm trước M1 và từ nằm sau M2: chỉ quan tâm tới 2 từ đứng ngay trước M1 và đứng ngay sau M2, được chia làm 2 loại:
 - Từ đầu tiên đứng trước M1 và từ đầu tiên đứng sau M2
 - Từ thứ hai đứng trước M1 và từ thứ hai đứng sau M2

Như vậy, đặc trưng về từ sẽ gồm các phần sau:

- WM1: tập các từ trong M1
- HM1: từ trung tâm của M1
- WM2: tập các từ trong M2
- HM2: từ trung tâm của M2
- HM12: kết hợp các từ trung tâm của cả HM1 và HM2
- WBNUL: khi không có từ nào nằm giữa
- WBFL: từ duy nhất nằm giữa khi chỉ có một từ nằm giữa
- WBF: từ đầu tiên nằm giữa khi có ít nhất hai từ nằm giữa M1 và M2
- WBL: từ cuối cùng nằm giữa khi có ít nhất hai từ nằm giữa M1 và M2
- WBO: các từ không phải từ đầu tiên và cuối cùng nằm giữa M1 và M2
- BM1#1: từ đầu tiên nằm trước M1
- BM1#2: từ thứ hai đứng trước M1
- AM2#1: từ đầu tiên đứng sau M2
- AM2#2: từ thứ hai đứng sau M2

Đặc trưng về kiểu thực thể: có 5 loại thực thể được quan tâm là NGƯỜI, TỔ CHỨC, CÔNG TY, ĐỊA DANH và GPE. Đặc trưng này sẽ có các thuộc tính sau:

- ET12: thể hiện kiểu thực thể của M1 và M2
- EST12: thể hiện các kiểu thực thể con của M1 và M2
- EC12: thể hiện lớp thực thể của M1 và M2

Đặc trưng về các bậc có liên quan (mention level): thể hiện các đặc trưng liên quan tới thực thể đang xem xét, ví dụ M1 hoặc M2 có thể là TÊN, DANH TỪ và ĐẠI TỪ... Đặc trưng này bao gồm hai thuộc tính:

- ML12: kết hợp các thông tin liên quan của M1 và M2
- MT12: kết hợp các thông tin của LDC về kiểu của M1 và M2

Đặc trưng về nạp chồng: các thuộc tính của đặc trưng này gồm có

- #MB: số lượng
- #WB: số lượng các từ nằm giữa
- $M1 > M2$ hay $M1 < M2$:

Thông thường, các đặc trưng trùng nhau ở trên là quá phổ biến để có thể tự mình gây ảnh hưởng. Vì vậy, chúng cần được kết hợp thêm với các thuộc tính khác:

- ET12 (hoặc EST12) + $M1 > M2$
- ET12(EST12) + $M1 < M2$
- HM12 + $M1 > M2$
- HM12 + $M1 < M2$

Đặc trưng dựa trên cụm từ: đặc trưng này được đánh giá mang tính then chốt trong các bài toán trích chọn quan hệ. Các phương pháp khác sử dụng thông tin này dựa trên cây phân tích cú pháp, tuy nhiên, trong phương pháp này thì tách bạch việc tạo ra các cụm từ và cây phân tích cú pháp đầy đủ. Ở đây, các cụm từ được trích chọn dựa trên cây phân tích cú pháp. Hầu hết các đặc trưng về cụm từ quan tâm tới từ trung tâm của các cụm nằm giữa M1 và M2. Tương tự như các đặc trưng về từ, đặc trưng về cụm từ được chia làm 3 loại sau:

- Các cụm từ trung tâm nằm giữa M1 và M2 chia làm 3 loại con:
 - Cụm từ đầu tiên nằm giữa M1 và M2
 - Cụm từ cuối cùng nằm giữa M1 và M2

- Cụm từ nằm giữa M1 và M2
- Cụm từ trung tâm nằm trước M1, gồm 2 cụm từ:
 - Cụm từ đầu tiên trước M1
 - Cụm từ thứ hai trước M1
- Cụm từ trung tâm nằm sau M2, gồm 2 cụm từ:
 - Cụm từ đầu tiên sau M2
 - Cụm từ thứ hai sau M2

Như vậy, đặc trưng này gồm có 12 thuộc tính được biểu diễn như sau:

- CPHBNUL: không có cụm từ nào nằm giữa M1 và M2
- CPHBFL: cụm từ trung tâm duy nhất khi chỉ có duy nhất một cụm từ trung tâm
- CPHBF: cụm từ trung tâm đầu tiên nằm giữa nếu có ít nhất hai cụm từ nằm giữa M1 và M2
- CPHBL: cụm từ trung tâm cuối cùng nằm giữa nếu có ít nhất hai cụm từ nằm giữa M1 và M2
- CPHBO: các cụm từ trung tâm khác nằm giữa M1 và M2 (ngoại trừ CPHBF và CPHBL)
- CPHBM1#1: cụm từ trung tâm đầu tiên trước M1
- CPHBM1#2: cụm từ trung tâm thứ hai trước M1
- CPHAM2#1: cụm từ trung tâm đầu tiên sau M2
- CPHAM2#2: cụm từ trung tâm thứ hai sau M2
- CPP: đường nối các nhãn cụm từ trên đường đi từ M1 sang M2
- CPPH: đường nối các nhãn cụm từ trên đường đi từ M1 sang M2 chỉ tính các cụm từ trung tâm (nếu có ít nhất 2 cụm từ nằm giữa)

Đặc trưng cây phụ thuộc: đặc trưng này bao gồm các thông tin về từ, từ loại, nhãn cụm từ của M1 và M2 dựa trên cây phụ thuộc, trích xuất từ cây phân tích cú pháp đầy đủ. Cây phụ thuộc được sinh ra bằng cách sử dụng thông tin về các cụm từ trung tâm dựa vào phân tích cú pháp Collins và liên kết tất cả các thành phần của cụm từ tới từ trung tâm của cụm từ đó. Các cờ đánh dấu thể hiện M1 và M2 có cùng là cụm danh từ, cụm động từ hay cụm giới từ không. Cụ thể, các thuộc tính của đặc trưng này như sau:

- ET1DW1: kết hợp của kiểu thực thể và từ phụ thuộc vào M1
- H1DW1: kết hợp của từ trung tâm và từ phụ thuộc vào M1
- ET2DW2: kết hợp của kiểu thực thể và từ phụ thuộc vào M2
- ET2DW2: kết hợp các từ trung tâm và từ phụ thuộc vào M2
- ET12SameNP: kết hợp ET12 với thông tin M1 và M2 có cùng là cụm danh từ hay không.
- ET12SamePP: kết hợp ET12 với thông tin M1 và M2 có cùng là cụm giới từ hay không.
- ET12SameVP: kết hợp ET12 với thông tin M1 và M2 có cùng là cụm động từ hay không.

Đặc trưng cây phân tích cú pháp: đặc trưng biểu diễn các thông tin có được từ cây phân tích cú pháp đầy đủ, bao gồm các thuộc tính:

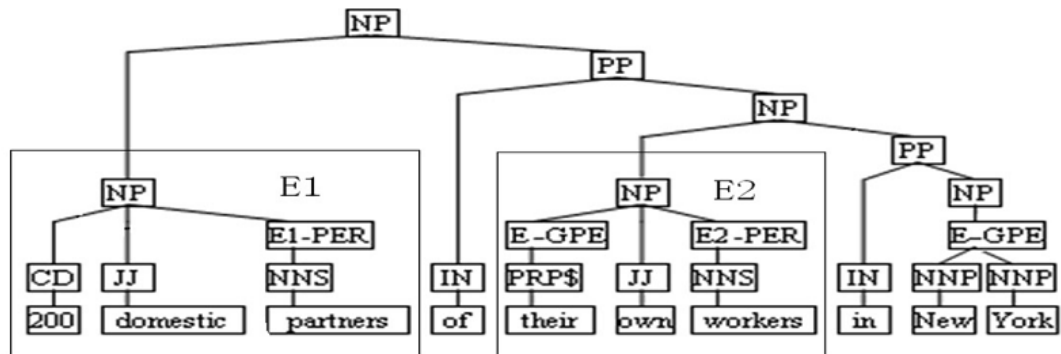
- PTP: đường đi thể hiện các nhãn cụm từ (loại bỏ các trùng lặp) nối M1 và M2 trên cây phân tích cú pháp
- PTPH: đường đi thể hiện các nhãn cụm từ (loại bỏ các trùng lặp) nối M1 và M2 trên cây phân tích cú pháp (chỉ tính các cụm từ trung tâm)

Đặc trưng từ các nguồn tài nguyên giàu ngữ nghĩa: Thông tin ngữ nghĩa từ rất nhiều nguồn tài nguyên như WordNet được sử dụng để phân lớp các từ quan trọng vào các danh sách ngữ nghĩa khác nhau tương ứng với các quan hệ đã được chỉ ra. Các thông tin này rất có ích trong việc giải quyết các trường hợp dữ liệu thô trong trích chọn quan hệ. Các nguồn này bao gồm:

- Danh sách tên các quốc gia: bao gồm các thông tin về tên quốc gia và các tỉnh, thành phố của nó. Có hai thuộc tính được sử dụng để biểu diễn đặc trưng này:
 - ET1 Country: kiểu thực thể của M1 khi M2 là tên của một quốc gia
 - ContryET2: kiểu thực thể của M2 khi M1 là tên của một quốc gia
- Danh sách từ thể hiện các quan hệ trong gia đình : bao gồm 6 loại quan hệ: cha mẹ, ông bà, vợ chồng, anh (chị) em, các quan hệ gia đình khác và quan hệ khác. Có hai thuộc tính được sử dụng để biểu diễn thông tin này, bao gồm:

- ET1SC2: kết hợp kiểu thực thể của M1 và lớp ngữ nghĩa của M2 khi M2 là một kiểu con của quan hệ xã hội
- SC1ET2: kết hợp kiểu thực thể của M2 và lớp ngữ nghĩa của M1 khi tham số đầu tiên là một dạng của quan hệ gia đình

Nanda Kambhatla [21] đã huấn luyện mô hình cực đại hóa Entropy sử dụng các đặc trưng có được từ luồng đặc trưng như mô tả ở trên để tiến hành trích chọn quan hệ.



Hình 5: Ví dụ về cây phân tích cú pháp

2. Entity Type: ET12_PER+PER;
3. Mention Level: ML12_NOMINAL+NOMINAL;
4. Overlap: #MB_0; #WB_1; M1>M2_NO; M1<M2_NO; ET12_PER+PER·M1>M2_NO; ET12_PER+PER·M1<M2_NO; HM12_partners+workers·M1>M2_NO; HM12_partners+workers·M1<M2_NO;
5. Base Phrase Chunking: CPHBFL_of; CPHBM1#1_to; CPHBM1#2_benefits; CPHAM1; CPHAM2#2_New+York; CPP_NP+PP+NP; CPPH_NP+PP(of)+NP;
6. Dependency Tree: ET1DW1_PER+200; ET1DW1_PER+domestic; H1DW1_partners; H1DW1_partners+domestic; ET2DW2_PER+their; ET2DW2_PER+workers; H2DW2_workers+their; H2DW2_workers+own; ET12_PER+PER·SameNP_YES;
7. Parse Tree: PTP_NP+PP+NP; PTPH_NP(partners)+PP+NP;
8. Semantic Resources: ET1SC2_PER+NONRelative; SC1ET2_NONRelative +PER;

Hình 6: Các đặc trưng thu được từ cây phân tích cú pháp

2.2.3. Phương pháp trích chọn dựa trên hàm nhân

Phương pháp này cũng giống phương pháp trích chọn dựa vào đặc trưng ở chỗ cũng biểu diễn quan hệ dưới dạng một vector đặc trưng. Nhưng điểm khác biệt ở cơ bản đối với phương pháp dựa vào đặc trưng là ở chỗ: phương pháp này tập trung vào việc xây dựng hàm nhân thế nào cho hiệu quả khi tiến hành phân lớp sử dụng thuật toán SVM chứ không phải là đặc trưng nào sẽ được lựa chọn.

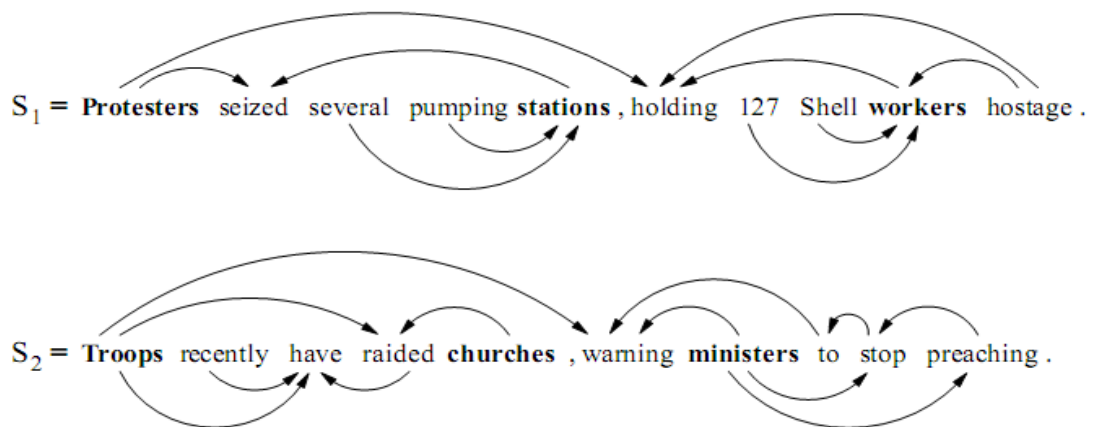
Razvan C. Bunescu và Raymond J. Mooney [8] đã đưa ra một phương pháp trích chọn quan hệ dựa trên quan sát rằng thông tin thể hiện quan hệ giữa hai thực thể có tên trong cùng một câu được biểu diễn bởi đường đi ngắn nhất giữa hai thực thể này trong đồ thị phụ thuộc (dependency graph) [35].

Dựa trên hai giả thiết:

- Các quan hệ được trích chọn được là quan hệ giữa các thực thể nằm trong cùng một câu
- Sự tồn tại hay không tồn tại của một quan hệ thì độc lập với đoạn văn bản trước và sau câu đang xem xét.

Điều này có nghĩa là chỉ trích chọn các quan hệ được mô tả trong câu chứa hai thực thể quan tâm.

Hơn nữa, với một câu được coi là một đồ thị phụ thuộc gồm các nút tương ứng với các từ trong câu, các cung có hướng được nối giữa hai từ phụ thuộc nhau dựa trên chức năng về ngữ pháp: tính từ bổ nghĩa cho danh từ trong cụm danh từ (“several → stations”), danh từ ghép (“pumping → stations”) hay trạng từ bổ nghĩa cho động từ (“recently → raided”) ... như ví dụ trong hình 7.



Hình 7: Minh họa đồ thị phụ thuộc

Trên đồ thị vô hướng thu được từ *đồ thị phụ thuộc* này, ta tìm được đường đi ngắn nhất giữa hai thực thể. Ví dụ một số đường đi ngắn nhất được thể hiện trong bảng 2-1.

Bảng 2-1: Đường đi ngắn nhất

Relation Instance	Shortest Path in Undirected Dependency Graph
S_1 : protesters AT stations	protesters \longrightarrow seized \longleftarrow stations
S_1 : workers AT stations	workers \longrightarrow holding \longleftarrow protesters \longrightarrow seized \longleftarrow stations
S_2 : troops AT churches	troops \longrightarrow raided \longleftarrow churches
S_2 : ministers AT churches	ministers \longrightarrow warning \longleftarrow troops \longrightarrow raided \longleftarrow churches

Đường đi này là dạng biểu diễn cô đọng nhất quan hệ giữa hai thực thể. *Đường đi phụ thuộc* được biểu diễn như là một chuỗi các từ. Dựa trên thông tin về từ loại, các kiểu thực thể... vector đặc trưng sẽ được sinh ra tương ứng với mỗi đường đi phụ thuộc. Ví dụ với đường “**protester** \rightarrow seized \leftarrow **stations**” ở bảng 2-1, ta được:

$$\begin{bmatrix} \text{protester} \\ \text{NNS} \\ \text{Noun} \\ \text{PERSON} \end{bmatrix} \times [\rightarrow] \times \begin{bmatrix} \text{seized} \\ \text{VBD} \\ \text{Verb} \end{bmatrix} \times [\leftarrow] \times \begin{bmatrix} \text{station} \\ \text{NNS} \\ \text{Noun} \\ \text{FACILITY} \end{bmatrix}$$

Khi đó, sẽ có tất cả $48 = (4 \times 1 \times 3 \times 1 \times 4)$ đặc trưng thu được cho đường đi này, ví dụ là:

Bảng 2-2: Một số đặc trưng thu được từ đường đi phụ thuộc

protesters	\rightarrow	seized	\leftarrow	stations
Noun	\rightarrow	Verb	\leftarrow	Noun
PERSON	\rightarrow	seized	\leftarrow	FACILITY
PERSON	\rightarrow	Verb	\leftarrow	FACILITY
... (48 features)				

Hàm nhân mà Razvan C. Bunescu và Raymond J. Mooney [7] đưa ra như sau:

Gọi $\mathbf{x} = x_1 x_2 \dots x_m$ và $\mathbf{y} = y_1 y_2 \dots y_n$ là hai quan hệ, trong đó x_i biểu diễn tập các thông tin ứng với từ nằm ở vị trí thứ i trong quan hệ. Khi đó, hàm nhân là số đặc trưng trùng nhau giữa \mathbf{x} và \mathbf{y} và được tính theo công thức:

$$K(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{nếu } m \neq n \\ \prod_{i=1}^n c(x_i, y_i) & \text{nếu } m = n \end{cases}$$

Trong đó $c(x_i, y_i) = |x_i \cap y_i|$ là số thuộc tính chung tại vị trí thứ i của \mathbf{x} và \mathbf{y}

Ví dụ: với hai thể hiện của quan hệ LOCATED:

1. “his actions in Breko” , và
2. “his arrival in Beijing”.

Ta có đường đi phụ thuộc tương ứng là:

1. “his→actions ← in←Breko”
2. “his→arrival← in←Beijing”

Lúc này:

$\mathbf{x} = [x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ x_7]$ trong đó $x_1 = \{\text{his, PRP, PERSON}\}$, $x_2 = \{\rightarrow\}$, $x_3 = \{\text{actions, NNS, Noun}\}$, $x_4 = \{\leftarrow\}$, $x_5 = \{\text{in, IN}\}$, $x_6 = \{\leftarrow\}$, $x_7 = \{\text{Breko, NNP, Noun, LOCATION}\}$

$\mathbf{y} = [y_1 \ y_2 \ y_3 \ y_4 \ y_5 \ y_6 \ y_7]$, trong đó $y_1 = \{\text{his, PRP, PERSON}\}$, $y_2 = \{\rightarrow\}$, $y_3 = \{\text{arrival, NN, Noun}\}$, $y_4 = \{\leftarrow\}$, $y_5 = \{\text{in, IN}\}$, $y_6 = \{\leftarrow\}$, $y_7 = \{\text{Beijing, NNP, Noun, LOCATION}\}$

Theo công thức trên, hàm nhân $K(\mathbf{x}, \mathbf{y}) = 3*1*1*1*2*1*3 = 18$.

Sử dụng thuật toán SVM với hàm nhân này để tiến hành phân lớp quan hệ, từ đó trích chọn được các quan hệ cần tìm.

2.3. Học bán giám sát trích chọn quan hệ

2.3.1. Phương pháp DIRPE

Vào năm 1998 [7][1], Brin đã giới thiệu một phương pháp học bán giám sát cho việc trích chọn mẫu quan hệ ngữ nghĩa DIRPE. Phương pháp được thử nghiệm với quan hệ “author –book” với tập dữ liệu ban đầu khoảng 5 ví dụ cho quan hệ này. DIRPE mở rộng tập ban đầu thành một danh sách khoảng 15.000 cuốn sách.

Phương pháp DIRPE được mô tả như sau:

Đầu vào: Tập các quan hệ mẫu $S = \{ \langle A_i, B_i \rangle \}$. Ví dụ trong trường hợp trên, tập quan hệ mẫu là $S = \{ \langle \text{author}_i, \text{book}_i \rangle \}$. Tập này được gọi là tập hạt giống.

Đầu ra: Tập các quan hệ R trích chọn được.

Xử lý:

- Tập quan hệ đích R được khởi tạo từ tập hạt giống S.
- Tìm tất cả các câu có chứa đủ các thành phần của tập hạt giống ban đầu.
- Dựa vào tập câu đã tìm được, tiến hành tìm các mẫu quan hệ giữa các thành phần của hạt giống ban đầu. Brin định nghĩa mẫu ban đầu rất đơn giản, bằng việc giữ lại khoảng m kí tự trước thành phần mẫu đầu tiên, gọi là *prefix*; giữ

lại phía sau thành phần thứ hai n kí tự gọi là *suffix*; k kí tự nằm giữa hai thành phần này, gọi là *middle*. Mẫu quan hệ được biểu diễn dưới dạng sau: [order, author, book, prefix, suffix, middle] trong đó, order thể hiện thứ tự xuất hiện của author và book trong một câu. (order = 1 thì author đứng trước book và bằng 0 trong trường hợp còn lại)

- Từ những mẫu mà chưa được gán nhãn ta thu được một tập hạt giống $\langle A', B' \rangle$ mới; thêm hạt giống mới này vào tập hạt giống cho quan hệ đó.
- Quay lại bước 2 để tìm ra những hạt giống và mẫu mới cho tới khi tập

Ví dụ minh họa đối với quan hệ “tác giả - sách” ở trên :

Đầu vào:

- Tập hạt giống ban đầu $S = \{ \langle \text{Arthur Conan Doyle, The Adventures of Sherlock Holmes} \rangle \}$.
- Và một tập các tài liệu bao gồm các hạt giống ban đầu

Xử lý:

- Quan hệ đích R được gán bằng S
- 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 hạt giống ban đầu. Từ tập câu này, tiến hành trích chọn các mẫu quan hệ. (như hình 8).

Từ đó trích chọn 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] ...

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

Hình 8: Các quan hệ mẫu trích chọn được

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 chọn cho:

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

- Việc sinh hạt giống 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 wrote Speckled Band in 1892, that is around 662 years apart which would make the stories*”...

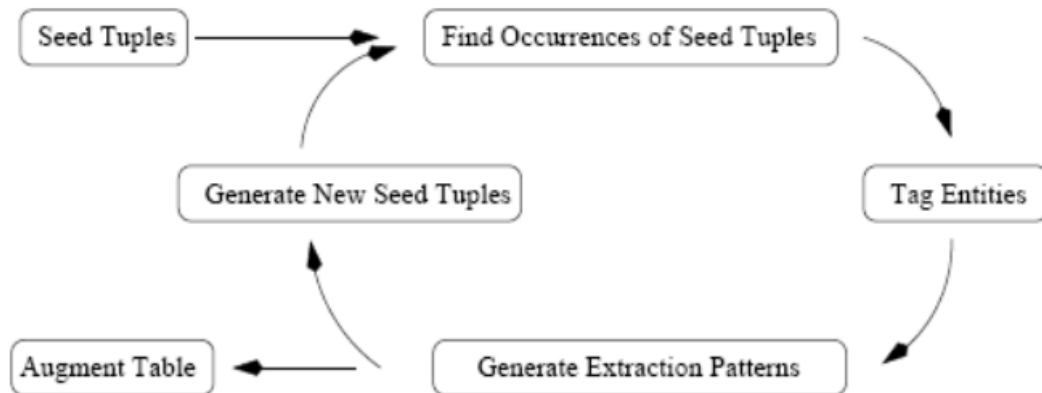
Từ tập câu tìm kiếm được, ta có thể trích xuất ra được những tập hạt giống mới mới: (Arthur Conan Doyle, Speckled Band)

Phương pháp đạt hiệu quả cao trên dữ liệu html cho việc xác định tập mẫu và sinh hạt giống mới. Vì thế, dựa trên ý tưởng của phương pháp DIPRE, vào năm 2000, Agichtein và Gravano đưa ra phương pháp Snowball [14] 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 hạt giống mới được sinh ra và bổ sung thêm việc nhận dạng thực thể. Phương pháp này được trình bày chi tiết hơn ở phần tiếp theo.

2.3.2. Phương pháp Snowball

Snowball [14][1] là hệ thống trích chọn quan hệ mà tập mẫu và tập hạt giống mới được sinh ra được đánh giá chất lượng trong quá trình xử lý. Giải thuật được thực nghiệm trên quan hệ “tổ chức – địa điểm” (“organization – location”). Với tập hạt giống ban đầu như: Microsoft – Redmond, IBM – Armonk, Boeing – Seattle, Intel – Santa Clara.



Hình 9: Kiến trúc của hệ thống Snowball

Kiến trúc cơ bản của Snowball được minh hoạ như hình 9 và được mô tả như sau:

Đầu vào:

- Một tập văn bản D (tập huấn luyện).
- Tập nhân hạt giống ban đầu $S = \{A_i, B_i\}$ gồm các cặp quan hệ mẫu nào đó. Ví dụ cặp quan hệ <Tổ chức – địa điểm> như trình bày ở trên.

Đầu ra: Tập các quan hệ trích chọn được

Xử lý:

Bước 1: Tìm sự xuất hiện của các cặp quan hệ trong dữ liệu

- Với hạt giống $\langle A_i, B_i \rangle$, tiến hành tìm dữ liệu là các câu có chứa cả A_i và B_i . Hệ thống sẽ tiến hành phân tích, chọn lọc và trích chọn các mẫu. Tương tự như DIPRE, một câu khớp với biểu thức “* A_i * B_i *” thì cụm từ đứng trước A_i gọi là prefix, cụm từ đứng giữa A_i và B_i là middle và cụm từ đứng sau B_i gọi là suffix.

Bước 2: Tìm sự xuất hiện của các thực thể trong dữ liệu

- 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 nhằm làm giảm số lượng các mẫu cũng như làm cho mẫu có tính khái quát cao hơn.
- Gọi $(prefix1, middle1, suffix1)$ và $(prefix2, middle2, suffix2)$ là hệ số ngữ cảnh tương ứng với mẫu1 và mẫu2 thì độ tương đồng $Match(mẫu1, mẫu2)$ được xác định 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 hạt giống mới $\langle A', B' \rangle$ nào không. Hạt giống mới $\langle A', B' \rangle$ sẽ nằm một trong các trường hợp sau:
 - *Positive*: Nếu $\langle A', B' \rangle$ đã nằm trong danh sách hạt giống
 - *Negative*: Nếu $\langle A', B' \rangle$ chỉ có đúng một trong hai (A' hoặc B') xuất hiện trong danh sách hạt giống.
 - *Unknown*: Nếu $\langle A', B' \rangle$, cả A' , B' đều không xuất hiện trong danh sách hạt giống. Tập Unknown được xem là tập các hạt giống mới cho vòng lặp sau.

Bước 3: Sinh mẫu mới

- 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:

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

Bước 4: Tìm các hạt giống mới cho vòng lặp tiếp theo

- Với mỗi mẫu trong danh sách top N được chọn sẽ là các cặp trong tập hạt giống 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 hạt giống cho quá trình chọn 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.

2.4. Nhận xét

Cả ba loại học không giám sát, có giám sát và bán giám sát đều thể hiện được những ưu và nhược điểm riêng của mình. Theo Valpola [31], đối với học có giám sát, chất lượng trích chọn của hệ thống trên những miền dữ liệu cụ thể là rất tốt, tuy nhiên chi phí đối với việc xây dựng tập dữ liệu là rất tốn kém, do đó khả năng mở rộng miền ứng dụng là khó khăn. Còn đối với phương pháp học không giám sát cho khả năng học với lượng dữ liệu lớn hơn và tốc độ nhanh tuy nhiên mô hình học lại phức tạp hơn học có giám sát. Trong khi 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. Việc lựa chọn phương pháp nào là tùy thuộc vào từng miền ứng dụng và đặc trưng của bài toán.

Tại Việt Nam, các nghiên cứu và các sản phẩm thiết yếu xử lý văn bản tiếng Việt ra đời [2, 38] cho phép áp dụng nhiều kỹ thuật xử lý hơn để trích chọn quan hệ ngữ nghĩa, chẳng hạn các thông tin về tách từ, nhãn từ loại và đặc biệt là cây phân tích cú pháp. Hơn nữa, dựa trên việc tổng hợp các kết quả nghiên cứu gần đây, G. Zhou và M. Zhang [32] đã khẳng định các rằng phương pháp tiếp cận dựa trên đặc trưng đạt được kết quả tốt hơn.

Đây chính là các lý do vì sao mà khóa luận đã đưa ra mô hình trích chọn quan hệ dựa vào cây phân tích cú pháp theo phương pháp dựa trên đặc trưng.

Tóm tắt chương hai

Trong chương này đã mô tả khái quát các phương pháp giải quyết bài toán trích chọn quan hệ, chỉ ra được những ưu nhược điểm và lý do lựa chọn phương pháp dựa trên đặc trưng để giải quyết bài toán này. Mô hình trích chọn quan hệ của khóa luận này sẽ được trình bày chi tiết trong chương tiếp theo.

Chương 3. Mô hình trích chọn quan hệ trên Wikipedia tiếng Việt dựa vào cây phân tích cú pháp

Trên cơ sở phân tích ưu và nhược điểm của các phương pháp trích chọn quan hệ, khóa luận đã lựa chọn phương pháp học có giám sát trích chọn quan hệ dựa trên đặc trưng để giải quyết bài toán này. Các đặc trưng của quan hệ sẽ được lấy ra dựa trên cây phân tích cú pháp tiếng Việt, sau đó được đưa vào bộ phân lớp sử dụng thuật toán SVM. Hơn nữa, để giảm công sức cho giai đoạn xây dựng tập dữ liệu học, các đặc trưng của dữ liệu trên Wikipedia tiếng Việt đã được sử dụng. Vì vậy, trong chương này, khóa luận trình bày các đặc trưng của Wikipedia, cây phân tích cú pháp tiếng Việt và mô hình đề xuất trích chọn quan hệ trên Wikipedia.

3.1. Đặc trưng của Wikipedia

Wikipedia gọi tắt là **Wiki** (phát âm như "Uy-ki"; từ tiếng Hawaii *wikiwiki*, có nghĩa "nhanh"; cũng được gọi là *công trình mở*), là một loại ứng dụng xây dựng và quản lý các trang thông tin do nhiều người cùng phát triển được đưa ra vào năm 2001 bởi Jimmy Wales và Larry Sanger [24]. Wiki được xây dựng theo nguyên tắc phân tán: Ai cũng có thể chỉnh sửa, thêm mới, bổ sung thông tin lên các trang tin và không ghi lại dấu ấn là ai đã cung cấp thông tin đó. Đây được xem là một “Bách khoa toàn thư” – bộ tra cứu lớn nhất và phổ biến nhất trên Internet hiện nay [23].

Nhờ đặc trưng biểu diễn thông tin rất giàu ngữ nghĩa được thể hiện ở các mẫu định dạng dữ liệu, các liên kết giữa các thực thể trang Wiki và cách phân mục các trang Wiki mà Wikipedia trở thành một đối tượng được quan tâm đặc biệt trong lĩnh vực khai phá dữ liệu và xử lý ngôn ngữ tự nhiên [5, 6, 13, 16, 19, 23].

3.1.1. Thực thể trong Wikipedia

Trên Wiki, một thực thể thường được liên kết tới một trang Wiki mô tả thực thể đó (đôi khi được gọi là thực thể trang Wiki) theo cách: khi một thực thể được tạo ra trên wiki, tác giả tạo ra một liên kết giữa thực thể và trang web Wiki mô tả thực thể đó, đồng thời, với mỗi thực thể xuất hiện trong trang Wiki này, liên kết tới trang Wiki mô tả thực thể đó cũng tạo ra. Đây là một đặc trưng quan trọng của Wiki cho phép dễ dàng xác định các thực thể. Ví dụ sau được trích ra từ trang “Đại học Công nghệ, Đại học Quốc gia Hà Nội” trên Wiki, bao gồm các liên kết tới thực thể “Đại học Quốc gia Hà Nội”, “Nguyễn Văn Hiệu”...

“**Trường Đại học Công nghệ** (tên [tiếng Anh](#): *University of Engineering and Technology* hay *UET*) là một trường đại học thuộc [Đại học Quốc gia Hà Nội](#),

được [Thủ tướng chính phủ](#) quyết định thành lập ngày [25 tháng 5](#) năm [2004](#). Đây là một [mô hình đại học](#) hiện đại. GS. TSKH. Viện sỹ [Nguyễn Văn Hiều](#) là Hiệu trưởng sáng lập trường.”

3.1.2. Infobox

Infobox của một trang Wiki là một bảng được thiết kế theo một mẫu cố định theo quy định của Wikipedia, nằm ở góc trên bên phải của trang, biểu diễn tóm tắt các thông tin về trang wiki đó với nội dung thường là các sự kiện (fact) và các thống kê liên quan [33]. Nội dung của bảng thường được biểu diễn dưới các cặp <thuộc tính – giá trị> [16]. Hình 12 là một ví dụ về infobox của trang Wiki “Trường Đại học Khoa học Tự nhiên”. Các bảng này cho phép trích chọn các thông tin một cách chính xác và nhanh chóng.

3.1.3. Mục phân loại

Wikipedia cũng cung cấp các mục phân loại, cho phép các tác giả phân nhóm và tạo các liên kết tới từ các trang tới các mục phân loại tương ứng. Một trang có thể liên kết tới nhiều mục. Một mục trên Wikipedia có một tên duy nhất. Một mục mới có thể được tạo ra bởi một tác giả tuân theo những khuyến cáo của Wiki trong việc tạo một mục mới và liên kết các trang tới nó. Một vài thuộc tính quan trọng của mục trên Wikipedia gồm có:

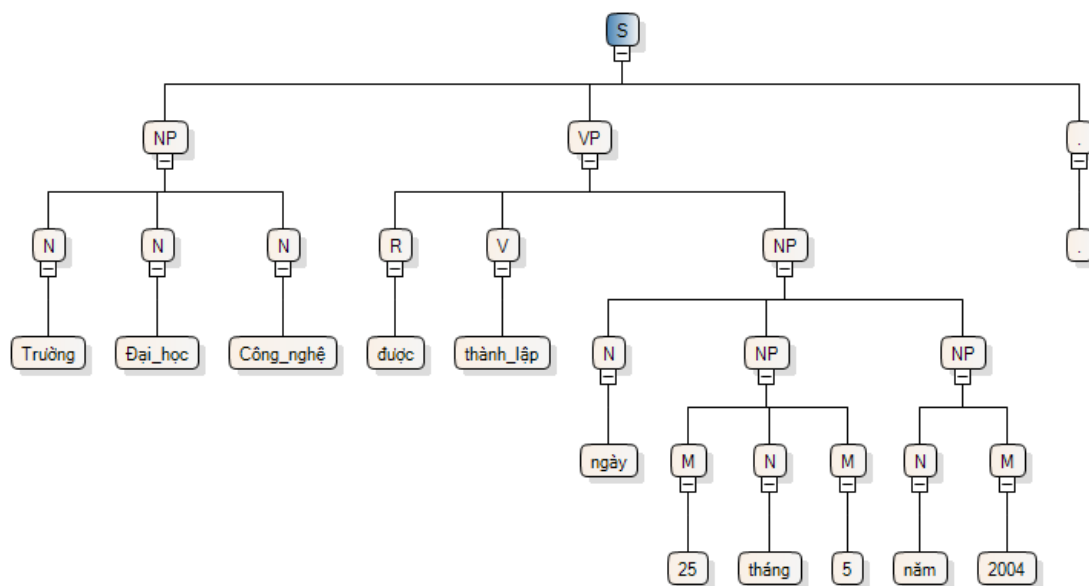
- Một mục có thể có nhiều mục con và nhiều mục cha
- Một mục có thể có chứa rất nhiều trang nhưng cũng có những mục chỉ có một lượng nhỏ các trang.
- Một trang mà thuộc về mục mở rộng thường không thuộc về các mục cha của mục mở rộng đó. Ví dụ trang Spain không thuộc mục “Người châu Âu”
- Quan hệ “mục con của một mục” không phải luôn luôn là quan hệ cha con. Ví dụ, “Bản đồ Châu Âu” là mục con của mục “Châu Âu” nhưng hai mục này không có quan hệ *is-a*
- Có chu trình trong đồ thị biểu diễn các mục.

3.2. Cây phân tích cú pháp tiếng Việt

Trong mục này sẽ trình bày một số các khái niệm và thành phần cơ bản về cây phân tích cú pháp¹, là cơ sở cho biểu diễn các đặc trưng của một quan hệ.

3.2.1. Phân tích cú pháp

Nhận đầu vào là một chuỗi các từ tố (là kết quả của quá trình phân tích từ tố, thông thường đối với xử lý ngôn ngữ là các từ), phân tích cú pháp (parsing hay syntactic analysis) là quá trình phân tích nhằm đưa ra cấu trúc ngữ pháp của chuỗi từ đó dựa vào một văn phạm nào đó. Thông thường cấu trúc ngữ pháp được là ở dạng cây, bởi thông qua dạng này sự phụ thuộc của các thành phần là trực quan. Cây này được gọi là cây phân tích cú pháp.



Hình 10: Ví dụ về cây phân tích cú pháp tiếng Việt

3.2.2. Một số thành phần cơ bản của cây phân tích cú pháp tiếng Việt

Cấu trúc của cây cú pháp như sau:

- Nút gốc thể hiện loại câu (trần thuật, nghi vấn, cảm thán, cầu khiến)
- Các nút lá biểu diễn các từ trong câu
- Nút cha của các nút lá này biểu diễn nhãn từ loại tương ứng của nút con.

¹ KC01.01/06-10: "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt" (VLSP)

- Các nút trung gian còn lại thể hiện chức năng ngữ pháp (cụm danh từ, cụm động từ, bổ ngữ ...)

Ví dụ: Với câu: “*Trường Đại học Công nghệ được thành lập ngày 25 tháng 5 năm 2004.*”, sau khi tiến hành phân tích cú pháp, ta được cây phân tích cú pháp như hình 10. Có 14 nhãn từ loại, 5 nhãn cụm từ và 4 loại nhãn câu được liệt kê và mô tả như trong phụ lục.

3.3. Mô hình trích chọn quan hệ dựa trên cây phân tích cú pháp trên Wikipedia tiếng Việt

3.3.1. Phát biểu bài toán

Bài toán trích chọn quan hệ đã được Roxana Girju [10] phát biểu như ở chương 1, trong trường hợp này có thể được viết lại như sau:

Đầu vào:

- Tập dữ liệu D: tập các trang web trên Wikipedia tiếng Việt
- Tập thực thể $E = \{e_i\} \quad i = \overline{1, n}$ xuất hiện trong D
- Tập các loại quan hệ $\mathcal{R} = \{R_j\} \quad j = \overline{1, m}$

Đầu ra:

- Tất cả các bộ quan hệ (e_{i_1}, R_j, e_{i_2}) với $1 \leq i \leq n, 1 \leq j \leq m$

3.3.2. Ý tưởng giải quyết bài toán

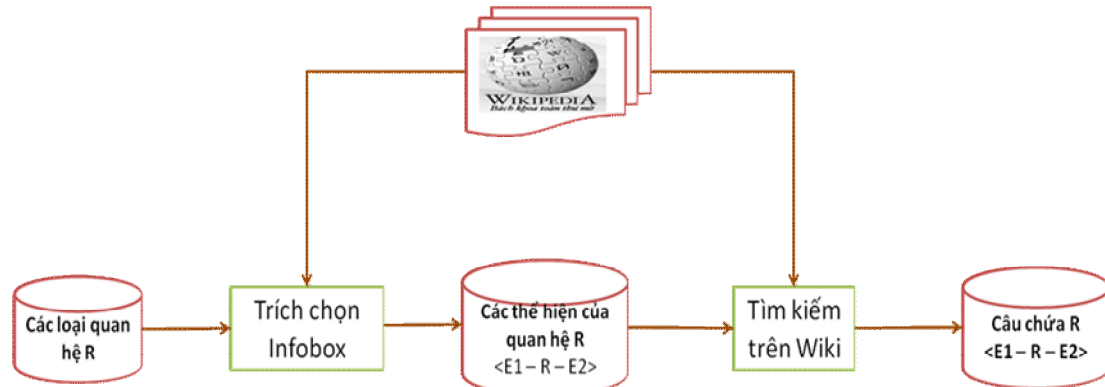
Việc tìm tất cả các bộ quan hệ (e_{i_1}, R_j, e_{i_2}) có thể được tiến hành bằng cách, với mỗi quan hệ $R_j \in \mathcal{R}$, tìm tất cả các cặp thực thể (e_{i_1}, e_{i_2}) thỏa mãn quan hệ R_j này. Như vậy, bài toán bây giờ trở thành: tìm tất cả các thể hiện của một quan hệ R cho trước. Dựa trên giả thiết rằng: “*mỗi thể hiện của 1 quan hệ được mô tả trong một câu*”, ý tưởng giải quyết bài toán được đưa ra như sau:

- Dựa trên cây phân tích cú pháp của câu, biểu diễn các thể hiện của quan hệ dưới dạng *cây quan hệ*. Mỗi cây quan hệ này sẽ tương ứng với một vector đặc trưng.
- Coi mỗi quan hệ R giống như một tập hợp – hay một **lớp** - các *cây quan hệ*. Nhãn của lớp này là tên quan hệ.
- Tiến hành tạo *bộ phân lớp các cây quan hệ*, từ đó trích chọn được thể hiện của quan hệ.

Mô hình trích chọn quan hệ được chia làm 2 pha chính: xây dựng tập dữ liệu học và giai đoạn áp dụng.

3.3.3. Xây dựng tập dữ liệu học

Một trong những nhược điểm của phương pháp học có giám sát là chi phí cho việc xây dựng tập dữ liệu là rất tốn kém. Dựa vào các đặc trưng của Wikipedia, khóa luận đã đưa ra mô hình xây dựng tập dữ liệu học bán tự động, giảm thiểu được nhiều chi phí xây dựng. Mô hình này được mô tả như trong hình 11:



Hình 11: Quá trình xây dựng tập dữ liệu học

a. Trích chọn thông tin trên Infox:

Như đã mô tả ở phần trước, thông tin trên infobox là một dạng biểu diễn có cấu trúc. Điều này cho phép ta trích chọn tự động các thể hiện của một quan hệ. Mỗi cặp <thuộc tính – giá trị> của infobox cho ta một bộ ba quan hệ với thực thể trang wiki có dạng: <Thực_thể_trang_Wiki – Thuộc_tính - Giá_trị>, các loại quan hệ <thuộc tính> và các cặp thực thể cùng nằm trong quan hệ <Thực_thể_trang_Wiki – Giá_trị>. Ví dụ, trong trường hợp hình 12, ta sẽ trích được bộ ba quan hệ, loại quan hệ, cặp thực thể tương ứng là:


<Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội – *Năm thành lập* - 1993>

<*Năm thành lập*>

< Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội – 1993>

b. Tìm kiếm trên Wikipedia

Mục tiêu của xử lý này là tìm ra các câu chứa cả ba thành phần của quan hệ <E1 – R – E2>. Do infobox là bảng thông tin tóm tắt về nội dung của trang nên sẽ gần như luôn tìm được các câu mà thể hiện quan hệ <E1 – R – E2>.

Infobox	Mã html tương ứng
<p>Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội</p>  <p>Tên gọi khác Trường Đại học Đông Dương Trường Đại học Khoa học Trường Đại học Tổng hợp Hà Nội</p> <p>Khẩu hiệu Khẩu hiệu</p> <p>Năm thành lập 1993</p> <p>Loại hình Trường Đại học công lập</p> <p>Giám đốc 1</p> <p>Hiệu trưởng PGS., TS. Bùi Duy Cam</p> <p>Hiệu phó Nguyễn Hữu Dư Nguyễn Hoàng Lương Nguyễn Văn Nội</p> <p>Giáo viên gần 700 ^[1]</p> <p>Học sinh trên 10.000 sinh viên ^[1]</p> <p>Địa chỉ 334 Nguyễn Trãi, Thanh Xuân, Hà Nội, Việt Nam</p> <p>Điện thoại (84) 043-8584615/ 8581419</p> <p>Email dhkhtn@vnn.vn</p> <p>Website http://www.hus.edu.vn</p>	<pre> <table class="infobox" > <tbody> <tr> <td>Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội
</td> <td colspan="2"></td> </tr> <tr> <th>Tên gọi khác</th> <td>Trường Đại học Đông Dương
Trường Đại học Khoa học
Trường Đại học Tổng hợp Hà Nội</td> </tr> <tr> <th>Khẩu hiệu</th> <td>Khẩu hiệu</td> </tr> <tr> <th>Năm thành lập</th> <td>1993</td> </tr> <tr> <th>Loại hình</th> <td>Trường Đại học công lập</td> </tr> <tr> <th>Giám đốc</th> <td>1</td> </tr> <tr> <th>Hiệu trưởng</th> <td>PGS., TS. Bùi Duy Cam</td> </tr> <tr> <th>Hiệu phó</th> <td>Nguyễn Hữu Dư
Nguyễn Hoàng Lương
Nguyễn Văn Nội</td> </tr> <tr> <th>Giáo viên</th> <td>gần 700 ^[1]</td> </tr> <tr> <th>Học sinh</th> <td>trên 10.000 sinh viên ^[1]</td> </tr> <tr> <th>Địa chỉ</th> <td>334 Nguyễn Trãi, Thanh Xuân, Hà Nội, Việt Nam</td> </tr> <tr> <th>Điện thoại</th> <td>(84) 043-8584615/
8581419</td> </tr> <tr> <th>Email</th> <td>dhkhtn@vnn.vn</td> </tr> <tr> <th>Website</th> <td>http://www.hus.edu.vn</td> </tr> </tbody> </table> </pre>

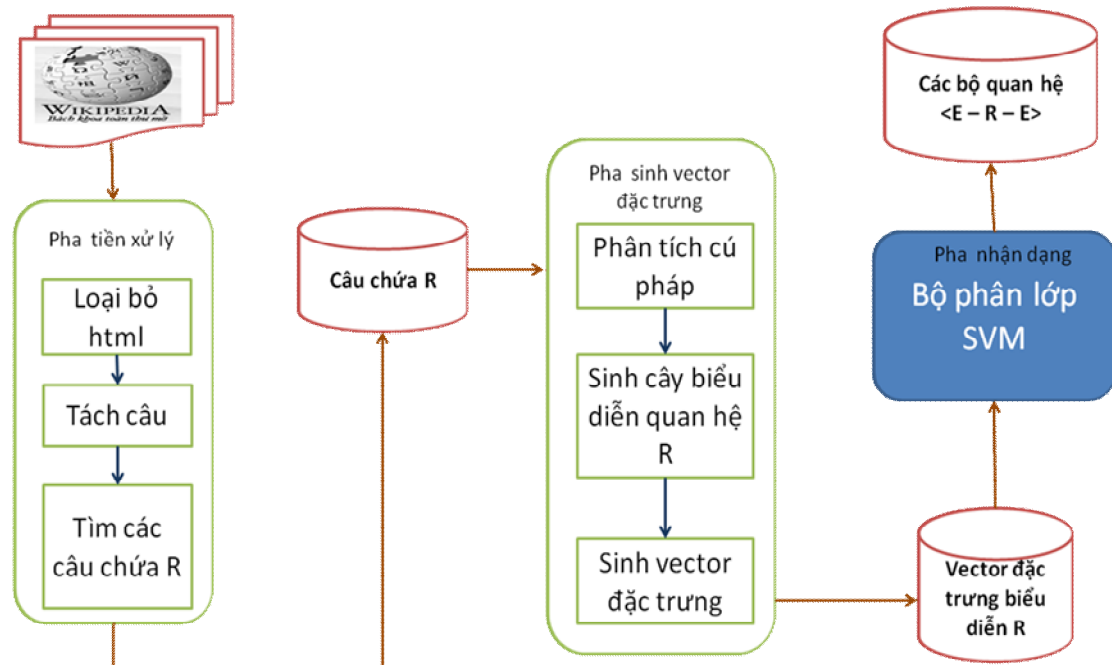
Hình 12: Cấu trúc biểu diễn của thông tin của infobox

Sau khi trích chọn được một tập các câu chứa các bộ quan hệ tương ứng <E1 – R – E2>, tiến hành phân tích cây cú pháp, tìm cây biểu diễn quan hệ này, rồi sinh

ra vector đặc trưng tương ứng. Các vector này sẽ được gán nhãn bằng tay và cho vào huấn luyện bộ phân lớp SVM như được mô tả dưới đây.

3.3.4. Mô hình hệ thống trích chọn quan hệ

Mô hình trích chọn quan hệ gồm có 3 pha chính: tiền xử lý, sinh vector đặc trưng và nhận dạng như được mô tả như trong hình vẽ sau:



Hình 13: Mô hình trích chọn quan hệ trên Wikipedia

Chi tiết về xử lý của từng pha như sau:

3.3.4.1. Pha tiền xử lý

Trong pha này, nhận đầu vào một tập các trang Wikipedia trên một miền ứng dụng quan tâm, sau quá trình xử lý thu được một tập các câu *tiềm năng* thể hiện quan hệ R . Các câu tiềm năng là các câu chứa từ khóa thể hiện quan hệ R đang xem xét.

Lần lượt từng trang sẽ được loại bỏ các thẻ html. Trong quá trình loại bỏ thẻ html thì đánh dấu các liên kết tới các thực thể trang Wiki khác.

Tiến hành tách câu sử dụng bộ công cụ JvnTextpro [43].

Chẳng hạn như trong ví dụ về thực thể trang “Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội”, với quan hệ “*năm thành lập*” các ta sẽ tìm được câu tiềm năng là:

“Trường Đại học Khoa học Tự nhiên thuộc Đại học Quốc gia Hà Nội được thành lập theo nghị định số 97/CP ngày 10/12/1993 của chính phủ”.

Các câu này sẽ được lưu lại, phục vụ cho pha tiếp theo.

3.3.4.2. *Pha sinh vector đặc trưng*

Trong pha này gồm 3 xử lý con:

a. Phân tích cú pháp

Trong pha này, sử dụng Hệ phân tích câu tiếng Việt [38], ta thu được các cây phân tích cú pháp tương ứng với từng câu thu được ở pha một.

b. Sinh cây con biểu diễn quan hệ R

Dựa trên một số nhận xét sau:

- Tiếng Việt là ngôn ngữ có cấu trúc câu dạng “chủ ngữ - vị ngữ - bổ ngữ”, tức có nghĩa là *chủ ngữ* thường đi trước, sau đó tới *vị ngữ* và cuối cùng là *bổ ngữ* [4]. Cấu trúc này tương đương với cấu trúc “subject – verb – object” trong tiếng Anh [34].
- Trong câu, chủ ngữ thường là các danh từ, cụm danh từ.
- Các thực thể hay khái niệm là các danh từ hay cụm danh từ
- Dựa trên liên kết “chủ ngữ - vị ngữ - bổ ngữ”, ta có được liên kết “(cụm) danh từ – (cụm) động từ – (cụm) danh từ” trên cây phân tích cú pháp.

Khi đó, cây con (của cây phân tích cú pháp) có khả năng biểu diễn quan hệ R sẽ có ba thành phần trung tâm là: một cụm từ trung tâm biểu diễn quan hệ R (thông thường là cụm động từ) và hai cụm danh từ biểu diễn hai thực thể tương ứng. Thủ tục sinh các cây này như sau:

Đầu vào: cây phân tích cú pháp có chứa các từ khóa k thể hiện quan hệ R

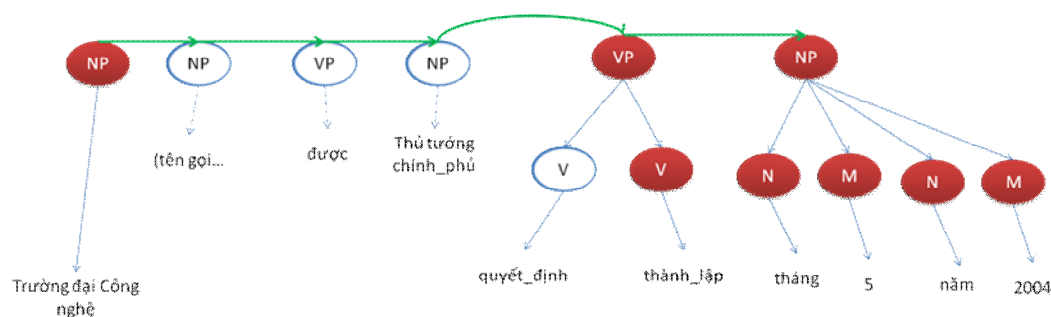
Đầu ra: tất cả các cây con tiềm năng thể hiện quan hệ R

Xử lý:

- i. Tìm nút nhỏ nhất trên cây chứa từ khóa k , gọi là nút K
- ii. Tìm tất cả các cụm danh từ NP thỏa mãn một trong các điều kiện [2]:
 - a. Nhánh NP có độ sâu bằng 1
 - b. Nhánh NP có độ sâu bằng 2 có phần đầu, danh từ trung tâm và phần sau. Trong đó, phần sau là nhánh có nhãn khác PP (cụm giới từ) và khác SBAR (câu)

- c. Nhánh NP có độ sâu bằng 3 chỉ gồm danh từ trung tâm và theo sau là một NP có độ sâu bằng 2
- d. Các nhánh có nhãn QP cũng được xem xét là cụm danh từ chỉ số lượng
- iii. Với từng cặp (NP_i, NP_j) có được từ bước ii, dựa vào cây phân tích cú pháp, tìm đường đi từ NP_i tới NP_j mà đi qua KEY. Đường đi này cho ta cây con tiềm năng biểu diễn R.

Ví dụ với câu “*Trường Đại học Công nghệ (tên gọi tiếng Anh : ...) được thủ tướng chính phủ quyết định **thành lập** ngày 25 tháng 5 năm 2004*” ta lấy được cây con biểu diễn R có dạng:



Hình 14: Cây con biểu diễn quan hệ “thành lập”

c. Sinh vector đặc trưng

Mỗi cây con ở trên tương ứng với một vector đặc trưng. Vector đặc trưng này gồm có 5 đặc trưng sau:

- *Cụm nhân trung tâm*: cụm nhân có nội dung biểu diễn quan hệ R. Trong hình 14, cụm này là VP (nhãn màu đỏ)
- *Cụm nhân thể hiện E_1* : cụm nhân có nội dung biểu diễn thực thể E_1 . Ví dụ: NP ngoài cùng bên trái
- *Cụm nhân thể hiện E_2* : cụm nhân có nội dung biểu diễn thực thể E_2 . Ví dụ: NP ngoài cùng bên phải
- *Đường dẫn nhãn E_i* : đường đi từ cụm nhân biểu diễn E_i tới cụm nhân trung tâm. Trong ví dụ trên: đường dẫn nhãn E_1 và E_2 lần lượt là **NP** -> **NP** -> **VP** -> **NP** -> **VP** và **NP** -> **VP**. Đặc trưng này có 2 thuộc tính:
 - Số nút nằm trung gian khi đi từ nút biểu diễn thực thể E_i tới nút trung tâm
 - Độ dài trung bình của đường đi (Bằng trung bình trọng số của các nút trung gian trên đường đi từ thực thể E_i tới nút trung tâm)
- Trọng số của một nút được xác định như sau:
 - Nút lá có trọng số bằng 1
 - Nút còn lại có trọng số bằng tổng trọng số của các nút con

Như vậy, một vector đặc trưng gồm có 7 thuộc tính, được mô tả chi tiết trong bảng sau:

Bảng 3-1: Các thuộc tính của vector đặc trưng

STT	Tên cụm	Giá trị	Ý nghĩa
1	Cụm nhãn trung tâm	[0,1]	Khả năng nhãn thể hiện quan hệ đang tìm. Giá trị càng cao thì khả năng càng lớn.
2	Cụm nhãn thể hiện E1	[0,1]	Khả năng nhãn thể hiện một thực thể đúng. Giá trị càng cao thì khả năng càng lớn.
3	Cụm nhãn thể hiện E2	[0,1]	Khả năng nhãn thể hiện một thực thể đúng. Giá trị càng cao thì khả năng càng lớn.
4	Đường dẫn nhãn E1	Số nhãn nằm trung gian khi đi từ nhãn biểu diễn thực thể E1 tới nhãn trung tâm	Độ liên quan của thực thể đối với quan hệ, thể hiện qua khoảng cách và thành phần của các nhãn trung gian. Giá trị càng lớn thì độ liên quan càng nhỏ.
5		Độ dài trung bình của đường đi (Bằng trung bình trọng số của các nút trung gian trên đường đi từ thực thể E ₁ tới nút trung tâm)	
6	Đường dẫn nhãn E2	Số nhãn nằm trung gian khi đi từ nhãn biểu diễn thực thể E2 tới nhãn trung tâm	Độ liên quan của thực thể đối với quan hệ, thể hiện qua khoảng cách và thành phần của các nhãn trung gian. Giá trị càng lớn thì độ liên quan càng nhỏ.
7		Độ dài trung bình của đường đi (Bằng trung bình trọng số của các nút trung gian trên đường đi	

		từ thực thể E_2 tới nút trung tâm)	
--	--	--------------------------------------	--

3.3.4.3. Pha nhận dạng

Việc nhận dạng các vector đặc trưng trở thành việc phân lớp nhị phân sử dụng mô hình SVM đã được huấn luyện.

Như đã trình bày ở bước xây dựng tập dữ liệu học, các câu trong bộ dữ liệu học sẽ được phân tích cú pháp, sinh cây con biểu diễn quan hệ R và sinh vector đặc trưng tương ứng như các bước ở trên. Sau đó, các vector này sẽ được gán nhãn bằng tay. Nếu cây con được sinh ra thực sự biểu diễn quan hệ R, vector tương ứng sẽ được gán nhãn $c1$ ngược lại sẽ được gán nhãn $c0$. Tiến hành huấn luyện mô hình SVM với tập dữ liệu học này ta được bộ phân lớp SVM cho quan hệ R.

Các vector đặc trưng của các cây con tiềm năng sẽ được phân lớp bởi bộ phân lớp này. Từ các vector nhận giá trị $c1$ tương ứng là các cây con tiềm năng sẽ được chấp nhận và quan hệ thu được từ cây con này là câu trả lời cho bài toán.

Tổng kết chương ba

Trong chương này, dựa trên phân tích các đặc trưng của dữ liệu Wikipedia tiếng Việt và cây phân tích cú pháp tiếng Việt, khóa luận đã đưa ra một phương án xây dựng tập dữ liệu học bán tự động và mô hình trích chọn quan hệ dựa trên phương pháp học có giám sát. Kết quả thực nghiệm ở chương sau cho thấy mô hình là hoàn toàn khả thi.

Chương 4. Thực nghiệm và đánh giá kết quả

4.1. Môi trường thực nghiệm

4.1.1. Cấu hình phần cứng

Bảng 4-1: Cấu hình phần cứng

Thành phần	Chỉ số
CPU	Intel Core 2 Duo 2.0Ghz
RAM	2GB
HDD	160GB
OS	Windows 7 Professional 32 bit

4.1.2. Công cụ phần mềm

Hệ thống sử dụng các công cụ sau:

Bảng 4-2: Danh sách các phần mềm sử dụng

STT	Tên phần mềm	Tác giả	Nguồn
1.	eclipse-SDK-3.4.0-win32		http://www.eclipse.org/downloads
2.	ColtechParser	Nguyễn Phương Thái	
3.	JvnTextpro	Nguyễn Cẩm Tú	
4.	weka-3-6-2		http://prdownloads.sourceforge.net/weka/weka-3-6-2.exe
5.	LibSVM	Chih-Chung Chang và Chih-Jen Lin	http://www.csie.ntu.edu.tw/~cjlin/libsvm/

4.2. Dữ liệu thực nghiệm

Dữ liệu thực nghiệm là hơn 4000 trang Wiki tiếng Việt được lấy từ [37]. Trong đó có 300 trang Wiki về các miền trường Đại học và cao đẳng trong cả nước.

4.3. Thực nghiệm

4.3.1. Mô tả cài đặt chương trình

Chương trình được tổ chức thành 4 gói:

- **RE.Crawler** : thực hiện các thu thập các trang Wiki theo miền hoặc theo từng trang cụ thể.
- **RE.Infobox** : trích chọn các bộ quan hệ dựa trên infobox của Wiki
- **RE.GrammarTree** : các thủ tục xử lý cây phân tích cú pháp và sinh vector đặc trưng
- **RE.Util** : Các thủ tục chuẩn hóa văn bản, xử lý xâu...

4.3.2. Xây dựng tập dữ liệu học dựa trên Wikipedia tiếng Việt

Đối với phương pháp học có giám sát, việc xây dựng tập dữ liệu học là đặc biệt quan trọng. Theo thống kê về các loại quan hệ được quan tâm nhất trong bài toán trích chọn quan hệ [21], khóa luận đã lựa chọn 3 quan hệ: “*năm thành lập*”, “*hiệu trưởng*” và “*ngày sinh*” để tiến hành thực nghiệm. Tập dữ liệu học cho mỗi quan hệ khoảng 350-400 câu. Quá trình xây dựng như sau:

a. Trích chọn infobox

Với mỗi trang Wiki, infobox của trang đó (nếu có) sẽ được trích chọn và tách ra thành các bộ quan hệ có dạng: $\langle E1 - R - E2 \rangle$, trong đó:

- E1: là thực thể trang Wiki đang xem xét
- R : quan hệ mà thực thể E1 có (chính là thành phần *thuộc tính* trong bảng infobox)
- E2: là thực thể có quan hệ R với E1 (là thành phần *giá trị* tương ứng với *thuộc tính* trong bảng infobox)

Ví dụ với trang Wiki “Đại học Quốc gia Hà Nội”, các bộ quan hệ trích chọn được là:

STT	Bộ quan hệ $\langle E1 - R - E2 \rangle$
-----	--

1.	<Đại học Quốc gia Hà Nội - Năm thành lập - 1906>
2.	< Đại học Quốc gia Hà Nội - Địa chỉ - 144 đường Xuân Thủy Quận Cầu Giấy, Hà Nội, Việt Nam>
3.	< Đại học Quốc gia Hà Nội - Website - www.vnu.edu.vn>
4.	< Đại học Quốc gia Hà Nội - Giám đốc - Mai Trọng Nhuận>
5.	< Đại học Quốc gia Hà Nội - Loại hình - Đại học quốc gia>
6.	<Đại_Hoc_Quoc_Gia_Ha_Noi - Điện thoại - +84-4-7547968>

Sau bước này thu được 864 bộ quan hệ.

Các bộ thể hiện quan hệ “năm thành lập”, “hiệu trưởng” và “ngày sinh” lần lượt được lấy ra. Thống kê kết quả được cho như bảng sau:

Quan hệ	Số lượng	Ví dụ bộ quan hệ <E1 – R – E2>
Hiệu trưởng	116	<Trường Đại học Văn Lang - Hiệu trưởng - TS. Nguyễn Dũng>
		<Học Viện Ngân Hàng Việt Nam - Hiệu trưởng - Tiến sĩ Tô Ngọc Hưng>
		<Trường Đại học Quốc Tế - Đại học Quốc Gia thành phố Hồ Chí Minh - Hiệu trưởng - Hồ Thanh Phong>
		<Trường Đại học Kiến Trúc Hà Nội - Hiệu trưởng - TS. Đỗ Đình Đức>
		<Trường Đại học Y Dược Cần Thơ - Hiệu trưởng - PGS. TS. Bác sĩ CK II Phạm Văn Linh>
		<Trường Đại học Bách Khoa Hà Nội - Hiệu trưởng - GS.TS. Nguyễn Trọng Giảng>
		<Trường Đại học Sư Phạm Hà Nội 2 - Hiệu trưởng - PGS.TS. Nguyễn Văn Mã>
		<Học Viện Kỹ Thuật Quân Sự - Hiệu trưởng - Giáo sư, TSKH Phạm Thế Long.>
		<Học Viện Y Dược Học Cổ Truyền Việt Nam - Hiệu trưởng - GS. TS.Trương Việt Bình>
		<Học Viện Ngoại Giao - Hiệu trưởng - PGS. TS. Dương Văn Quảng>
Năm thành lập	132	<Học Viện Ngân Hàng Việt Nam - Năm thành lập - 1998>
		<Trường Đại học Sư Phạm, Đại học Thái Nguyên - Năm thành lập - 25 tháng 12 năm 1987>
		<Trường Đại học Công nghiệp Hà Nội - Năm thành lập - 2005>
		<Trường Đại học Hà Hoa Tiên - Năm thành lập - 2007>
		<Trường Đại học Bà Rịa Vũng Tàu - Năm thành lập - 2006>
		<Học Viện Âm Nhạc Huế - Năm thành lập - 26 tháng 3 năm 2008>
		<Trường Đại Học Thành Tây - Năm thành lập - 10 tháng 10 năm 2007>
		<Trường Đại học Sư Phạm Đà Nẵng - Năm thành lập - 1975>

		<Khoa Quản trị Kinh doanh Đại học Quốc gia Hà Nội - Năm thành lập - 13 tháng 7 năm 1995>
		<Đại học Thái Nguyên - Năm thành lập - 1994>
		<Trường Đại học Điều Dưỡng Nam Định - Năm thành lập - 26 tháng 2 năm 2004>
Ngày sinh	160	<Nguyễn Tấn Dũng - ngày sinh - 17 tháng 11, 1949>
		<Nguyễn Văn Hiệu - ngày sinh - Ngày 21 tháng 07, 1938>
		<Phan Văn Khải - ngày sinh - 25 tháng 12, 1933>
		<Hồ Chí Minh - ngày sinh - 19 tháng 5, 1890>
		<Đinh Tiên Hoàn - ngày sinh - 924>
		<Nguyễn Đức Mạnh - ngày sinh - 11 tháng 9, 1940>
		<Gia Long - ngày sinh - 8 tháng 2 năm 1762>
		<Minh Mạng - ngày sinh - 25 tháng 5 năm 1791>
		<Nguyễn Du - ngày sinh - 3 tháng 1, 1766>
		<Trần Thái Tông - ngày sinh - 17 tháng 7, 1218>

b. Tìm kiếm trên Wiki

Để tìm các câu mô tả bộ quan hệ $\langle E1 - R - E2 \rangle$ vừa tìm được ở trên, ta tìm trong thực thể trang Wiki tương ứng. Các câu chứa cả ba thành phần của bộ quan hệ sẽ lấy ra và lưu vào trong cơ sở dữ liệu.

Quá trình này gồm 3 bước sau:

- Tạo truy vấn gửi tới modul *tìm kiếm* của Wiki. Từ khóa của truy vấn là quan hệ R và số lượng kết quả trả về. Wiki sẽ trả về một danh sách các trang Wiki có chứa từ khóa này.



Hình 15: Ví dụ về tìm kiếm trên Wikipedia

- Các trang trả về sẽ được thu thập, cho qua bước tiền xử lý (như ở mục tiếp theo)
- Các câu được trích ra có thể là một trong ba loại sau:
 - Loại 1: Câu chứa cả 3 thành phần của quan hệ
 - Loại 2: Câu chứa R và E1 hoặc R và E2
 - Loại 3: Câu chứa R

Các câu này sẽ được phân tích cú pháp, sinh cây quan hệ, sinh vector đặc trưng. Các vector đặc trưng có được từ câu loại 1 sẽ được gán nhãn tự động. Các vector đặc trưng có được từ câu loại 2 và 3 sẽ được gán nhãn bằng tay.

Tiền xử lý

Các trang sau khi được thu thập về sẽ được tiến hành tiền xử lý:

- Loại bỏ các thẻ html
- Tách câu
- Trích ra những câu chứa R
- Chuẩn hóa câu.

Việc loại bỏ các thẻ html, tách câu được thực hiện bởi bộ công cụ JvnTextPro[43], sau đó, những câu chứa R sẽ được lưu lại.

Có một số ký tự đặc biệt mà bộ phân tích cú pháp không xử lý cần được loại bỏ hoặc thay thế bằng kí hiệu tương đương. Các ký hiệu mở ngoặc “(”, đóng ngoặc “)” này thường được sử dụng mang ý nghĩa chú thích nên để không làm mất đi ý nghĩa, các cặp đóng mở ngoặc sẽ được thay thế bởi dấu gạch gang “-” tương ứng. Ví dụ: câu “**Trường Đại học Bách khoa Hà Nội** ([tiếng Anh](#): *Hanoi University of Technology*, viết tắt là *HUT*) là trường [đại học kỹ thuật](#) đa ngành, được thành lập tại [Hà Nội](#) ngày [15 tháng 10](#) năm [1956](#).” sẽ được chuẩn hóa thành “**Trường Đại học Bách khoa Hà Nội - tiếng Anh**: *Hanoi University of Technology*, viết tắt là *HUT* - là trường [đại học kỹ thuật](#) đa ngành, được thành lập tại [Hà Nội](#) ngày [15 tháng 10](#) năm [1956](#).”

4.3.3. Sinh vector đặc trưng

a. Phân tích cú pháp

- Tách từ: sử dụng bộ tách từ JvnTextpro[43] của Nguyễn Cẩm Tú.

- Đưa câu về dạng chuẩn đầu vào vào bộ phân tích cú pháp.
- Phân tích cú pháp sử dụng bộ phân tích cú pháp coltechparser của Nguyễn Phương Thái và cộng sự [38]

Nhân xét:

- Kết quả thực nghiệm cho thấy kết quả phân tích cú pháp sẽ phụ thuộc rất lớn vào việc tách từ.
- Phân tích cú pháp các câu sau khi đã tách từ sẽ cho cây phân tích cú pháp tốt hơn.

b. Trích chọn cây con biểu diễn quan hệ R và sinh vector đặc trưng

Sử dụng thuật toán như đã trình bày ở mục 3.3.4.2 ta sẽ sinh được các cây con có khả năng biểu diễn quan hệ $\langle E1 - R - E2 \rangle$ (gọi tắt là *cây con*)

Các thuộc tính của vector đặc trưng $v = (v_1, v_2, v_3, v_4, v_5, v_6, v_7)$ thể hiện khả năng mà cây con đó biểu diễn quan hệ R, cụ thể được xác định như sau trong quá trình thực nghiệm:

- Cụm nhãn trung tâm: Khả năng cây con thể hiện quan hệ R đang tìm (chứ không phải là quan hệ R' nào khác). Giá trị càng cao thì khả năng càng lớn. Nếu $Node_R$ là nút trên *cây con* biểu diễn R, gọi:
 - $num1$ là số nút lá của $Node_R$
 - $num2$ là số nút lá của $Node_R$ có giá trị trùng với từ khóa thể hiện R

Khi đó: v_1 được tính theo công thức

$$v_1 = \begin{cases} 0 & \text{node lá của } Node_R \text{ có chứa từ như “không”} \\ \frac{num2}{num1} & \text{trong trường hợp còn lại} \end{cases}$$

- Cụm nhãn thể hiện E1, E2: Khả năng các nút biểu diễn thực sự là thực thể. Giá trị càng cao thì khả năng càng lớn. Nếu $Node_{Ei}$ là nút trên *cây con* biểu diễn E_i , gọi:
 - $num1$ là số nút lá của $Node_{Ei}$
 - $num2$ là số nút lá của $Node_R$ biểu diễn thực thể E_i (đã xác định trước như theo giả thiết bài toán)

Khi đó: v_2, v_3 được tính theo công thức

$$v = \frac{num\ 2}{num\ 1}$$

- Đường dẫn tới nhãn E1, E2:
 - v_4 : số nút đi từ nút biểu diễn E1 sang nút biểu diễn R
 - v_6 : số nút đi từ nút biểu diễn E2 sang nút biểu diễn R
 - $v_5 = \frac{\sum w_t}{v_4}$ với w_t là trọng số của các nút trên đường đi từ nút biểu diễn E1 sang nút biểu diễn R với chú ý rằng $v_5=0$ nếu $v_4=0$
 - $v_7 = \frac{\sum w_t}{v_6}$ với w_t là trọng số của các nút trên đường đi từ nút biểu diễn E2 sang nút biểu diễn R với chú ý rằng $v_7=0$ nếu $v_6=0$
 - w_t được tính theo như mô tả trong mục 3.3.4.2
- Trong quá trình thực nghiệm áp dụng, trọng số của nút lá được gán bằng một mang ý nghĩa, các từ được sử dụng đều được xem là tương đương nhau.
 Cây con ở hình 14 có vector đặc trưng $v = (0.5; 1.0; 1.0; 3.0; 0.0; 2.0; 0)$

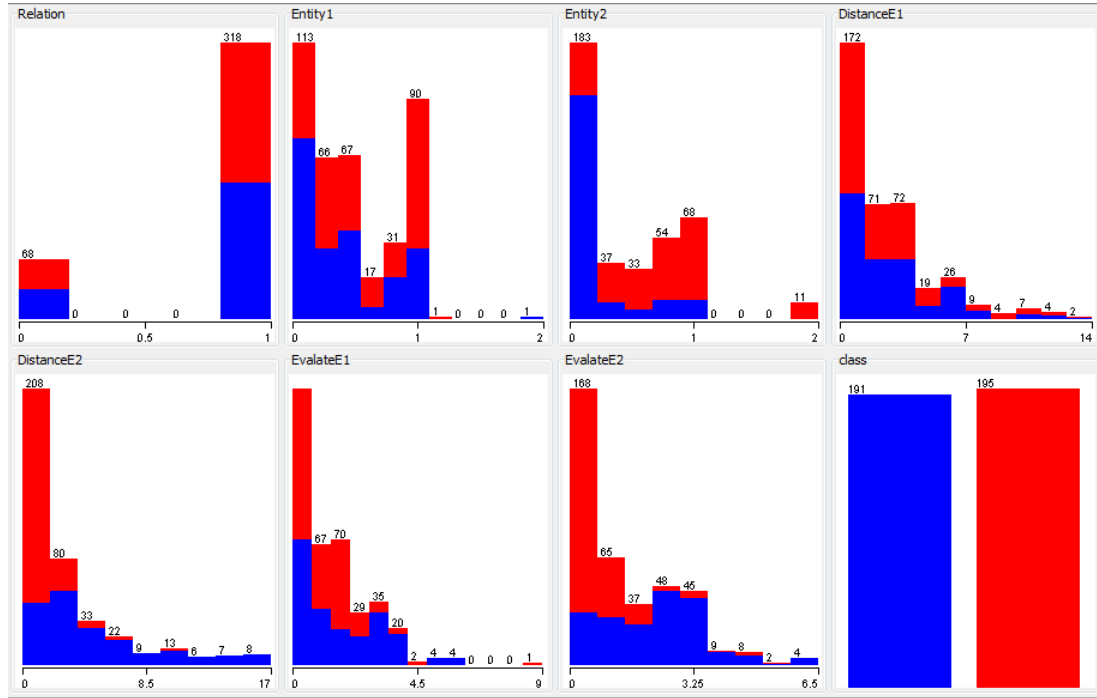
Nhận xét:

- Thực nghiệm cho thấy, giá trị của v_4, v_5, v_6, v_7 càng nhỏ thì cây con thu được càng có khả năng thể hiện đúng bộ quan hệ $\langle E - R - E \rangle$. Điều này cũng phù hợp với thực tế là khi các thành phần trên cây phân tích cú pháp càng gần nhau, thì mức độ quan hệ giữa chúng sẽ càng cao hơn.
- Điều này cũng chứng tỏ rằng, các công thức đưa ra tính vector đặc trưng là hợp lý.
- Tuy nhiên, vẫn còn một số nhập nhằng khi xác định trường hợp cụm nhãn trung tâm chứa từ khóa biểu diễn R nhưng lại chứa thêm các từ “không”.

4.3.4. Bộ phân lớp SVM

Sử dụng phần mềm Weka[26] và LibSVM[44] để tiến hành huấn luyện mô hình và kiểm thử.

Một ví dụ thông kê về dữ liệu học trong trường hợp quan hệ “năm thành lập” của mô hình được cho trên hình vẽ:



Hình 16 : Bảng thống kê dữ liệu học của quan hệ “ngày sinh”

4.4. Đánh giá

4.4.1. Đánh giá hệ thống

Hệ thống được đánh giá chất lượng thông qua ba độ đo: độ chính xác (precision), độ hồi tưởng (recall) và độ đo F (F-messure). Ba độ đo này được tính toán theo các công thức sau:

$$pre_{C_i} = \frac{correctC_i}{correctC_i + incorrectC_i}$$

$$rec_{C_i} = \frac{correctC_i}{correctC_i + incorrectC_0}$$

$$rec_{C_0} = \frac{correctC_0}{correctC_0 + incorrectC_i}$$

$$F_{C_i} = \frac{2 * pre_{C_i} * rec_{C_i}}{pre_{C_i} + rec_{C_i}}$$

Ý nghĩa của các giá trị $correctC_i$, $incorrectC_i$ được định nghĩa như bảng 4-3.

4.4.2. Phương pháp đánh giá

Hệ thống thử nghiệm theo phương pháp đánh giá chéo. Theo phương pháp này, dữ liệu thực nghiệm được chia thành 10 phần bằng nhau, lần lượt lấy 9 phần để huấn luyện và 1 phần còn lại để kiểm tra, kết quả sau 10 lần thực nghiệm được ghi lại và đánh giá tổng thể.

Bảng 4-3 : Các giá trị đánh giá hệ thống phân lớp

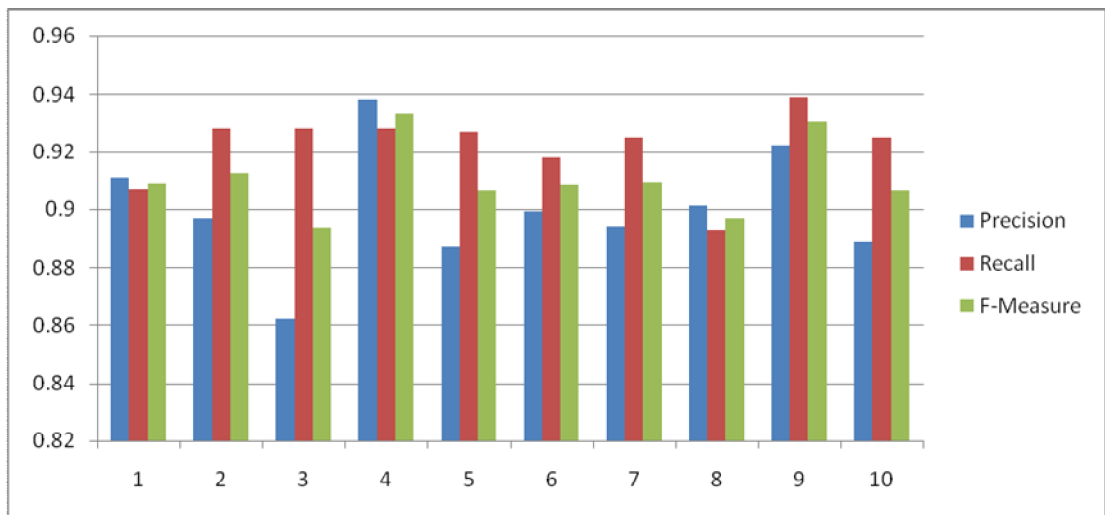
	C_0	C_1
C_0	$correctC_0$	$incorrectC_0$
C_1	$incorrectC_1$	$correctC_1$

Với:

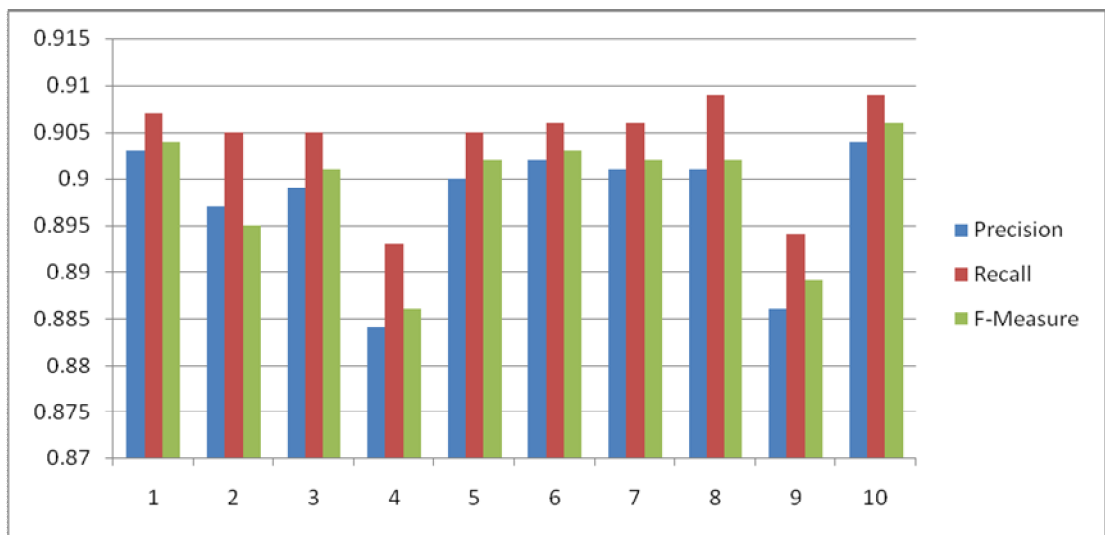
Giá trị	Ý nghĩa
$correctC_0$	Số kết quả được phân lớp vào C_0 là đúng
$incorrectC_0$	Số kết quả được phân lớp vào lớp C_0 là sai
$incorrectC_1$	Số kết quả được phân lớp vào lớp C_1 là sai
$correctC_1$	Số kết quả được phân lớp vào lớp C_1 là đúng

4.4.3. Kết quả kiểm thử

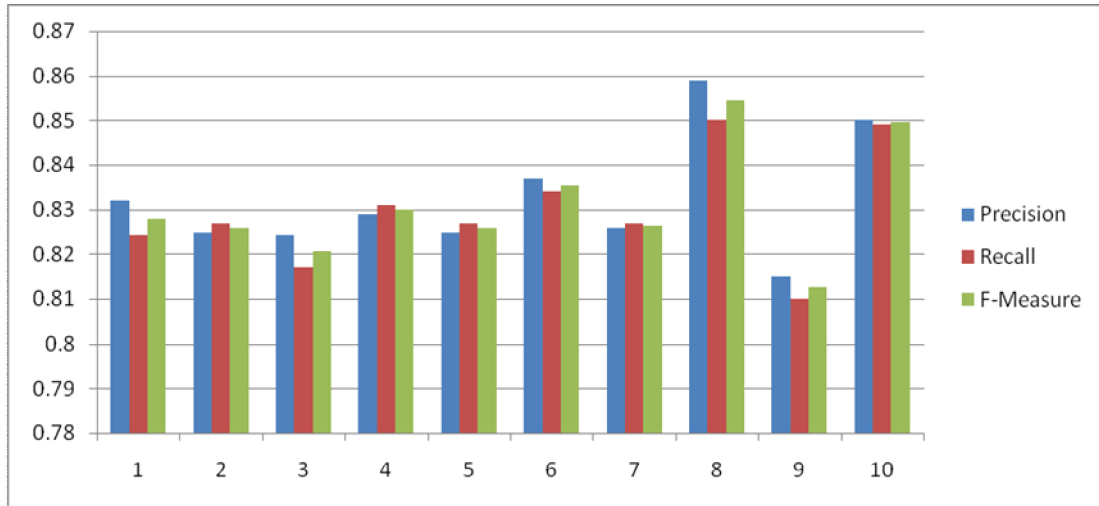
Kết quả kiểm thử của 3 quan hệ “năm thành lập”, “hiệu trưởng” và “ngày sinh” cho kết quả như sau:



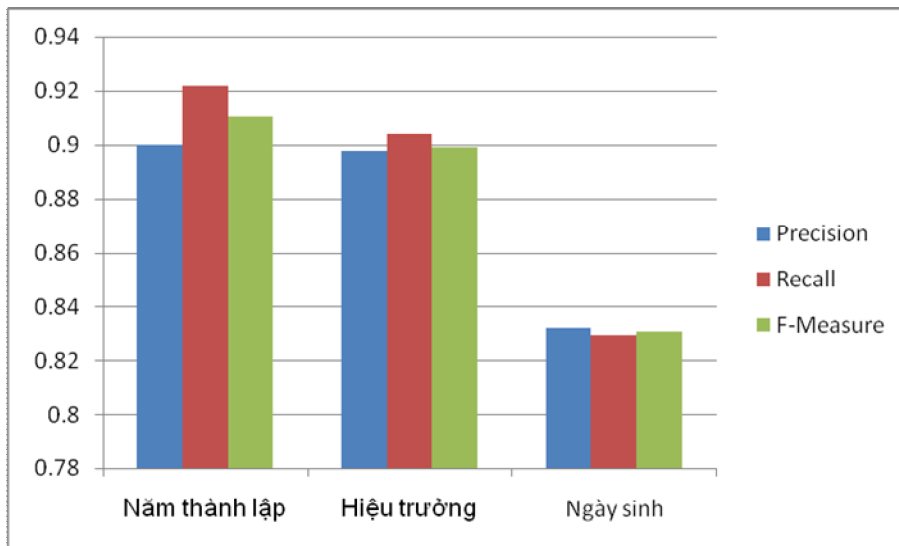
Hình 17: Kết quả kiểm thử đối với quan hệ “năm thành lập”



Hình 18: Kết quả kiểm thử đối với quan hệ “hiệu trưởng”



Hình 19: Kết quả kiểm thử đối với quan hệ “ngày sinh”



Hình 20: So sánh kết quả trung bình của ba quan hệ

4.5. Nhận xét

Bước đầu thực nghiệm hệ thống trích chọn quan hệ dựa trên cây phân tích cú pháp cho kết quả tương đối khả quan. Độ đo F_1 trung bình cho từng quan hệ thử nghiệm “năm thành lập”, “hiệu trưởng”, “ngày sinh” lần lượt là 91,06% , 89,9% và 83,08%. Tuy vẫn còn nhiều trường hợp nhập nhằng nhưng tôi tin rằng một khi đã xây dựng được tập dữ liệu huấn luyện đủ lớn, thu thập được các nguồn tra cứu dồi dào hơn và kết hợp thêm các đặc trưng khác, cũng như đưa ra được trọng số các nút riêng theo từng quan hệ, hệ thống còn có thể đạt được độ chính xác cao hơn nữa trong tương lai.

Kết luận

Từ việc nghiên cứu bài toán trích chọn quan hệ, khóa luận đã đưa ra mô hình trích chọn quan hệ thực thể dựa trên cây phân tích cú pháp trên miền dữ liệu Wikipedia tiếng Việt. Qua những kết quả thực nghiệm đạt được cho thấy mô hình là khả thi và có thể áp dụng được.

Về mặt nội dung, khóa luận đã đạt được những kết quả sau:

- Giới thiệu bài toán trích chọn quan hệ và các khái niệm liên quan.
- Tìm hiểu và phân tích các phương pháp trích chọn quan hệ điển hình, trong đó tập trung vào các phương pháp có sử dụng cây phân tích cú pháp.
- Dựa vào đặc trưng của Wikipedia tiếng Việt, đưa ra được mô hình xây dựng tập dữ liệu học bán tự động
- Áp dụng mô hình học có giám sát SVM để xây dựng mô hình trích chọn quan hệ dựa vào cây phân tích cú pháp trên miền dữ liệu của Wikipedia tiếng Việt đạt kết quả khả quan.

Bên cạnh những, do hạn chế về mặt thời gian và kiến thức khóa luận vẫn còn hạn chế sau:

- Khóa luận chưa xây dựng được giao diện người dùng và kết quả thực nghiệm ở một số trường hợp chưa đạt độ chính xác như mong muốn

Về định hướng nghiên cứu, việc giải quyết bài toán theo tiếp cận có giám sát là bước khởi đầu tốt. Trong thời gian tới, khóa luận sẽ được phát triển theo các hướng sau:

- Một là, hoàn thiện bước xây dựng tập dữ liệu học sao cho có thể thực hiện được trên nhiều quan hệ tiến tới xây dựng bộ phân lớp đa lớp.
- Hai là, thử nghiệm mô hình học không giám sát trên vector đặc trưng đã xây dựng được.
- Ba là, tích hợp modul này vào hệ thống xây dựng tự động ontology cho tiếng Việt trên miền ứng dụng các trường đại học Việt Nam nhằm phục vụ việc tìm kiếm hướng thực thể.

PHỤ LỤC

Bảng 5-1: Bảng các nhãn được sử dụng trong cây phân tích cú pháp

Kí hiệu nhãn	Phân loại	Ví dụ	Kí hiệu nhãn	Phân loại	Ví dụ
No - Danh từ riêng	No	Bùi Thúc Anh, Hà Nội...	A – Tính từ	Ai – Tính từ chỉ tính chất	Trong vắt, mênh mông
N - Danh từ	Ns – danh từ đơn thể	quần, áo, bạn...		An – Tính từ định lượng	Cao (hai mét), rộng (vài sải tây)..
	Nc – danh từ tổng thể	quần áo, binh lính, bạn bè...	P – Đại từ	Pp – Đại từ xưng hô	
	Na – Danh từ trừu tượng	giai điệu		Pd – Đại từ chỉ định	Đây, đó, kia...
	Nu – danh từ đơn vị đo lượng	lít rượu, nắm muối, mẫu đất, phút suy nghĩ...		Pn – Đại từ chỉ số lượng	Bấy, bấy nhiều, tất cả
V - Động từ	Vt – ngoại động từ	ăn bánh, xây nhà...	R – Phó từ	Pi – Đại từ ngghi vấn	Ai, gì, đâu, bao giờ, bao nhiều...
	Vi – nội động từ	ngủ, nói, làm việc		Rd - Phó từ chỉ hướng	Vào (nhà), xuống (cầu tháng), (sản xuất) ra

	Ve – động từ tồn tại	Còn, mất, hết...		Rt – Phó từ chỉ thời gian	
	Va – Động từ tiếp thụ	Bị, phải, được...	C – Giới từ		Do, của, với, hay, nếu
	Vv – Động từ tình thái	Muốn, dám, quả quyết	M – Trợ từ		Chinish, chợt, ngay, tất nhiên, à, ừ, hả, hử
	Vg – động từ tổng hợp	mua bán, đánh đập...	E – Cảm từ		Ái chà, ôi chao, dạ, vâng
Vz – Động từ “là”			Nl – Loại từ		Cái, con, cây, người, tắm...
NP – cụm danh từ		Tất cả những chiếc kẹo	Nq – Số từ		Một, hai, ba, dăm,...
VP – cụm động từ		Đang ăn cơm, yêu cô ấy, bán cho họ	Y – từ viết tắt		CHXH, TTCK, CNTT
AP – Cụm tính từ		Xinh quá, mỏng cùi, giỏi về thể thao	X – Từ không xác định		
RP – Cụm phó từ		Vẫn chưa	SBAR – mệnh đề phụ		Quyền sách mà anh mượn; khỏe vì chơi thể

					thao đều đặn
PP – Cụm giới từ		vào Sài Gòn	S – Câu trần thuật		Tôi đi học bằng xe đạp
QP – cụm từ chỉ số lượng		Năm trăm, hơn 200	SQ – Câu ngghi vấn		Ai đang ở trong nhà?
SE – Câu cảm thán		Ái chà,...	SC – Câu cầu khiến		Không được làm ồn, đi đi em...

TÀI LIỆU THAM KHẢO

Tiếng Việt

- [1] Hà Quang Thụy, Phan Xuân Hiếu, Đoàn Sơn, Nguyễn Trí Thành, Nguyễn Thu Trang, Nguyễn Cẩm Tú (2009). Giáo trình Khai phá dữ liệu Web. *NXBGDVN*, 10-2009.
- [2] Nguyễn Thị Minh Huyền, Phan Xuân Hiếu, Nguyễn Lê Minh, Lê Thanh Hương (2009). Báo cáo kết quả sản phẩm các công cụ xử lý ngôn ngữ tự nhiên tiếng Việt. *Đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt"*
- [3] Nguyễn Hồng Côn (2008). Cấu trúc cú pháp câu tiếng Việt: chủ - vị hay đề - thuyết. *Hội nghị khoa học về Việt Nam học*.

Tiếng Anh

- [4] Abdulrahman Almuhareb (2006). Attributes in lexical acquisition. *PhD Thesis*. University of Essex.
- [5] Adrian Iftene, Alexandra Balahur-Dobrescu (2008). Named Entity Relation Mining using Wikipedia. *The Sixth International Language Resources and Evaluation LREC08* (2008), European Language Resources Association (ELRA), Pages: 2–9517408
- [6] Anne-Marie Vercoustre, Jovan Pehcevski, James A. Thom (2007). Using Wikipedia Categories and Links in Entity Ranking. *INEX 2007*: 321-335.
- [7] Brin, S. (1998). Extracting patterns and relations from the world wide web. *WebDB 1998*: 172-183.
- [8] Bunescu R. C., and Mooney R. J. (2005). A shortest path dependency kernel for relation extraction. *HLT/EMNLP 2005*: 724–731.
- [9] Chinchor, N. and Marsh, E. (1998). Information extraction task definition (version 5.1). *The 7th Message Understanding Conference*. http://acl.ldc.upenn.edu/muc7/ie_task.html.
- [10] Corina Roxana Girju (2002). Text mining for semantic relations. *PhD. Thesis*. The University of Texas at Dallas, 2002.
- [11] Coyle, B., and Sproat, R. (2001). WordsEye: An automatic text-to-scene conversion system. *The Siggraph Conference*, Los Angeles, USA.

- [12] Daniel Sleator and Davy Temperly (1993). Parsing English with a Link Grammar. *Third International Workshop on Parsing Technologies*. <http://www.cs.cmu.edu/afs/cs.cmu.edu/project/link/pub/www/papers/ps/LG-IWPT93.pdf>.
- [13] Dat P. T. Nguyen, Yutaka Matsuo, Mitsuru Ishizuka (2007). Relation Extraction from Wikipedia Using Subtree Mining. *AAAI 2007*: 1414-1420
- [14] Eugene Agichtein, Luis Gravano (2000). Snowball: Extracting Relations from Large Plain-Text Collections. *ACM DL 2000*: 85-94.
- [15] Fabian M. Suchanek, Georgiana Ifrim, Gerhard Weikum (2006). LEILA: Learning to Extract Information by Linguistic Analysis. *COLING/ACL 2006 (Workshop On Ontology Learning And Population)*.
- [16] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum (2008). YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3): 203-217.
- [17] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Seaghdha, Sebastian Pado, Marco Pennacchiotti, Lorenza Romano and Stan Szpakowicz (2009). Multi-Way Classification of Semantic Relations Between Pairs of Nominals. *The NAACL-HLT-09 Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-09)*, Boulder, USA, May 2009.
- [18] Jinxiu Chen, Donghong Ji, Chew Lim Tan, Zhengyu Niu (2005). Unsupervised Feature Selection for Relation Extraction. *The 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, <http://www.aclweb.org/anthology/I/I05/I05-2045.pdf>
- [19] Jonathan Yu, James A. Thom and Audrey Tam (2007). Ontology evaluation using Wikipedia categories for browsing. *CIKM 2007*: 223-232.
- [20] Kai-Hsiang Yang, Chun-Yu Chen, Hahn-Ming Lee, and Jan-Ming Ho (2008). EFS: Expert Finding System Based on Wikipedia Link Pattern Analysis. *The 2008 IEEE International Conference on Systems, Man and Cybernetics (SMC 2008)*: 631-635.
- [21] Kambhatla N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *ACL 2004*.

- [22] Kim S., Lewis P., Martinez K. and Goodall S. (2004). Question Answering Towards Automatic Augmentations of Ontology Instances. *The Semantic Web: Research and Applications: First European Semantic Web Symposium, ESWS*: 152-166.
- [23] L.Denoyer and P.Gallinari (2006). The Wikipedia XML corpus. *SIGIRForum*, **40**(1): 64–69.
- [24] Larry Sanger (2005). The Early History of Nupedia and Wikipedia: A Memoir. *Open Sources 2.0*, ed. DiBona, Cooper, and Stone. O'Reilly, 2005 (Pre-published in slashdot.org, Apr. 2005).
- [25] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007). Open information extraction from the Web. *IJCAI 2007*: 2670-2676.
- [26] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, **11**(1):10-18.
- [27] Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G. Payan, Kunbin Qu, Ming Li (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, **20**(18):3604-3612.
- [28] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates (2004). Web-Scale Information Extraction in KnowItAll. [WWW 2004](#): 100-110.
- [29] I. Fahmi (2009). Automatic term and relation extraction for medical question answering system, *PhD Thesis*, University of Groningen, Netherlands
- [30] Sanghee Kim, Paul H. Lewis, Kirk Martinez (2004). The Impact of Enriched Linguistic Annotation on the Performance of Extracting Relation Triples. *CICLing 2004*: 547-558.
- [31] Valpola, H. (2000). Bayesian Ensemble Learning for Nonlinear Factor Analysis. *PhD Thesis*, Helsinki University of Technology.
- [32] Zhou GuoDong, Zhang Min. Extracting relation information from text documents by exploring various types of knowledge. *Information Processing and Management 43* (2007): 969–982.
- [33] <http://en.wikipedia.org/wiki/Help:Infobox>
- [34] http://en.wikipedia.org/wiki/Subject_Verb_Object

- [35] http://en.wikipedia.org/wiki/Dependency_graph
- [36] <http://inex.is.informatik.uni-duisburg.de/>
- [37] <http://static.wikipedia.org/downloads/2008-06/vi/>
- [38] <http://vlsp.vietlp.org:8080/demo/?page=home>
- [39] <http://wordnet.princeton.edu/>
- [40] <http://www.abisource.com/projects/link-grammar/>
- [41] <http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html> . Information about the sixth Message Understanding Conference.
- [42] http://www.db.dk/bh/Lifeboat_KO/CONCEPTS/semantic_relations.htm
- [43] Nguyen Cam Tu (2008). “JVnTextpro: A Java-based Vietnamese Text Processing Toolkit”
- [44] <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>