Tổng quan đề tài
Phương pháp đề xuất
Tổng quan quy trình
Tiền xử lý
Trích xuất cụm danh từ
Xây dựng dữ liệu học và kiểm tra
Tạo bộ phân loại và gom cụm
Thực nghiệm và đánh giá
Tổng kết

Luận văn tốt nghiệp

Phân giải đồng tham chiếu cho các đối tượng và thuộc tính trong khai khoáng ý kiến

Nguyễn Trọng Nghĩa 51202370
Nguyễn Đăng Trang 51203957

GVHD GS.TS Phan Thị Tươi
GVPB GS.TS Cao Hoàng Trụ

Đại học Bách Khoa TP. Hồ Chí Minh

Ngày 26 tháng 12 năm 2016

Luận văn tốt nghiệp

Tổng quan đề tài

Phương pháp đề xuất

Tổng quan quy trình

Tiền xử lý

Trích xuất cụm danh từ

Xây dựng dữ liệu học và kiểm tra

Tạo bộ phân loại và gom cụm

Thực nghiệm và đánh giá

Tổng kết

Giới thiệu đề tài

Phân giải đồng tham chiếu cho đối tượng và thuộc tính trong khai khoáng ý kiến hướng đến việc phân giải đồng tham chiếu cho các đối tượng và thuộc tính trong văn bản chứa ý kiến.

- Thương mại điện tử đang phát triển mạnh mẽ và người dùng có nhu cầu thể hiện ý kiến lên các sản phẩm trên mạng.
- Văn bản chứa ý kiến là các bài đánh giá (review), bài thảo luận (discussion),...
- Đối tượng và thuộc tính là các thực thể được thể hiện ý kiến, trong đó thuộc tính là bộ phận hoặc đặc tính của đối tượng.
- Nếu không có phân giải đồng tham chiếu cho các đối tượng và thuộc tính, ý kiến của người viết rất có thể sẽ được gán không đúng cho các thực thể.

Tổng quan đề tài (tt)

Phân giải đồng tham chiếu

Phân giải đồng tham chiếu là một bài toán kinh điển trong Xử lý ngôn ngữ tự nhiên. Mục tiêu của bài toán là tìm kiếm những từ, cụm từ hoặc ngữ cùng chỉ đến một khái niệm, thực thể trong thế giới thực.

Ví dụ

- *Beckham* will visit Vietnam tomorrow. *He* will attend a football event in Saigon.
- *The Samsung Galaxy S3* is one of the best phones I've ever used. I absolutely love the phone and *the photo quality* it has.

Tổng quan đề tài (tt)

Luận văn tốt nghiệp

Tổng quan đề tài

Phương pháp đề xuất

Tổng quan quy trình

Tiền xử lý

Trích xuất cụm danh từ

Xây dựng dữ liệu học và kiểm tra

Tạo bộ phân loại và gom cụm

Thực nghiệm và đánh giá

Tổng kết

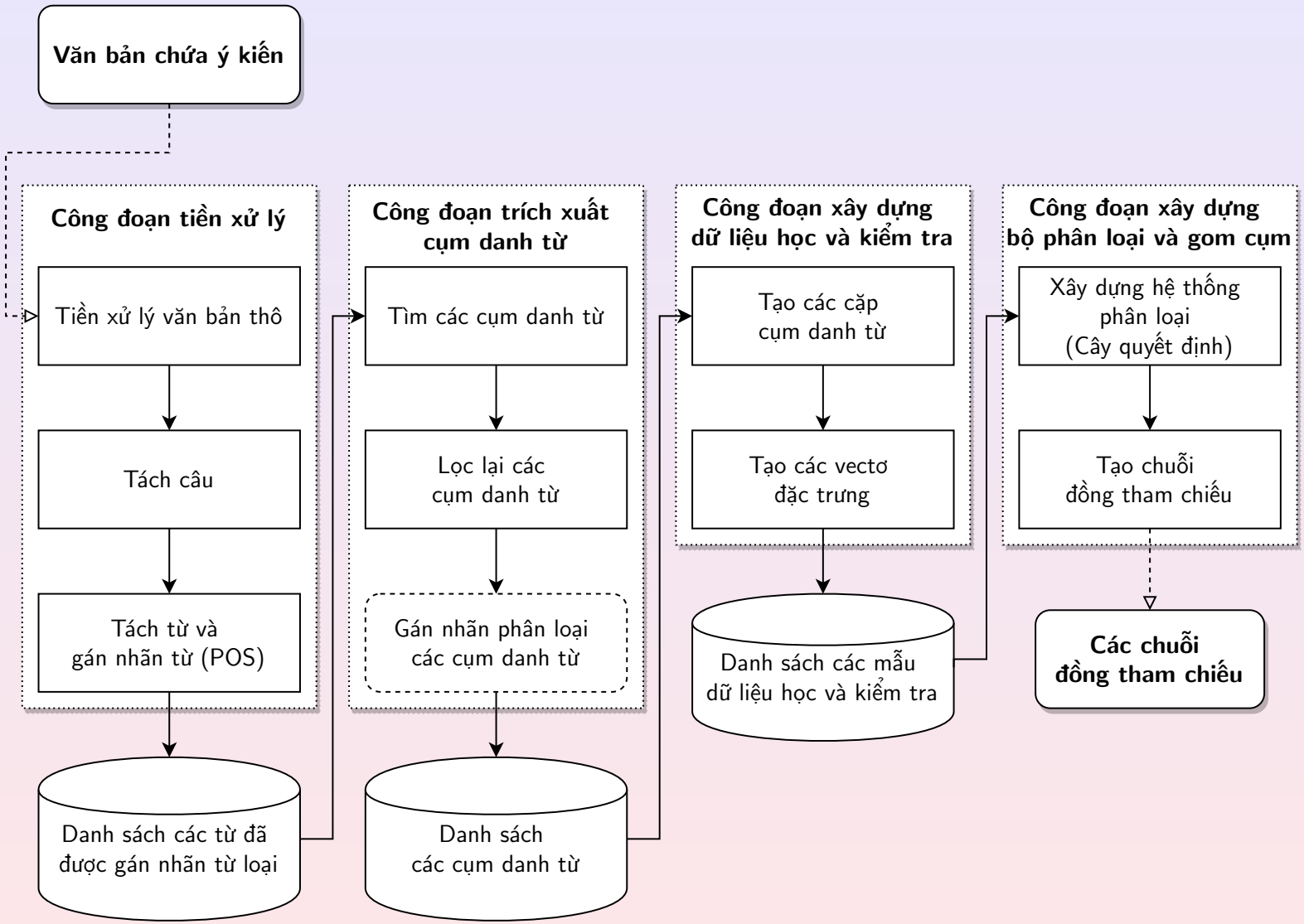
Mục tiêu đề tài

Tìm ra các từ/cụm từ trong văn bản chứa ý kiến cùng chỉ về một đối tượng hoặc thuộc tính nào đó, tức là tìm các chuỗi đồng tham chiếu của đối tượng và thuộc tính.

Phạm vi đề tài

- Giả định rằng các đối tượng và thuộc tính đã được tìm ra
- Chỉ xét các quan hệ đồng tham chiếu dạng *đồng nhất (identity)*

Tổng quan quy trình

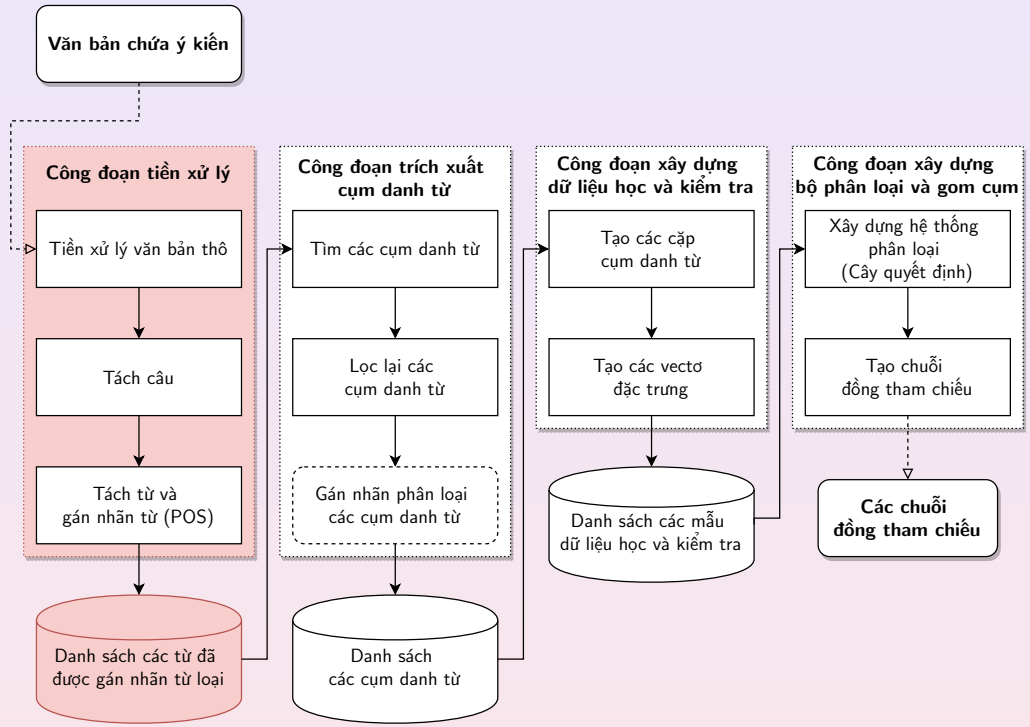


Hình: Tổng quan quy trình phân giải đồng tham chiếu

Tiền xử lý

Tiền xử lý văn bản thô

Sửa lỗi chính tả và một số lỗi nhỏ khác do cách viết không chuẩn mực của người dùng.



Tách câu, tách từ và gán nhãn từ loại

- Công cụ Stanford [1]
- Penn Treebank POS Tag [2]

[1] http://stanfordnlp.github.io/CoreNLP

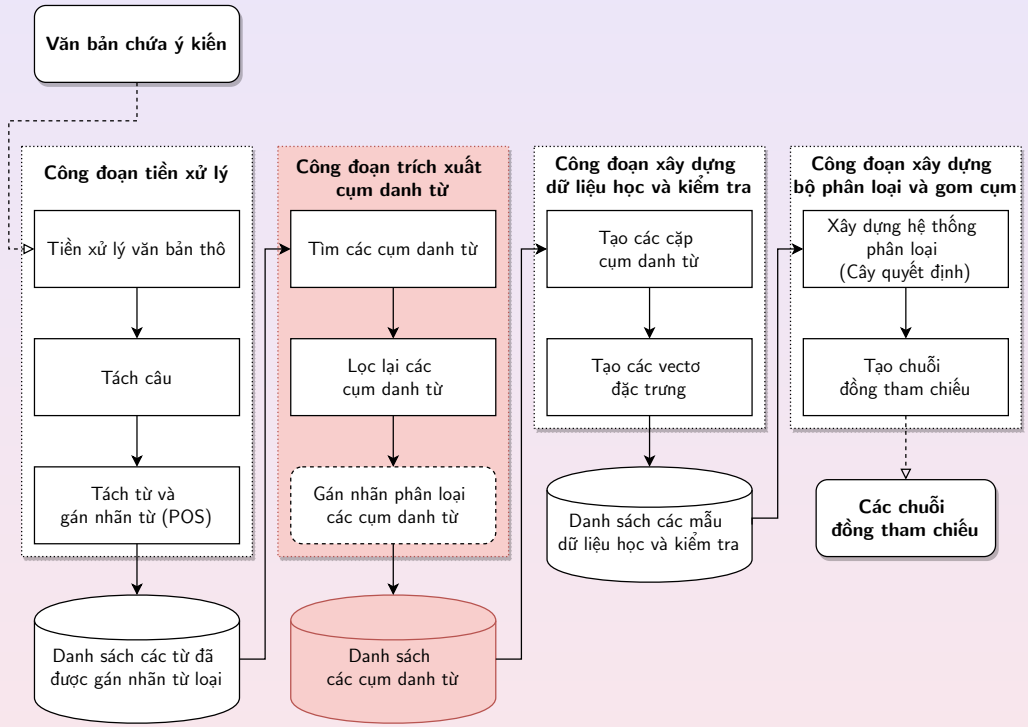
Trích xuất cụm danh từ

Tìm các cụm danh từ

Công cụ CRFChunker [3].

Lọc lại các cụm danh từ

Một số cụm danh từ sẽ không được đưa vào bộ phân loại vì chúng không thể chỉ về đối tượng và thuộc tính.



Gán nhãn cụm danh từ

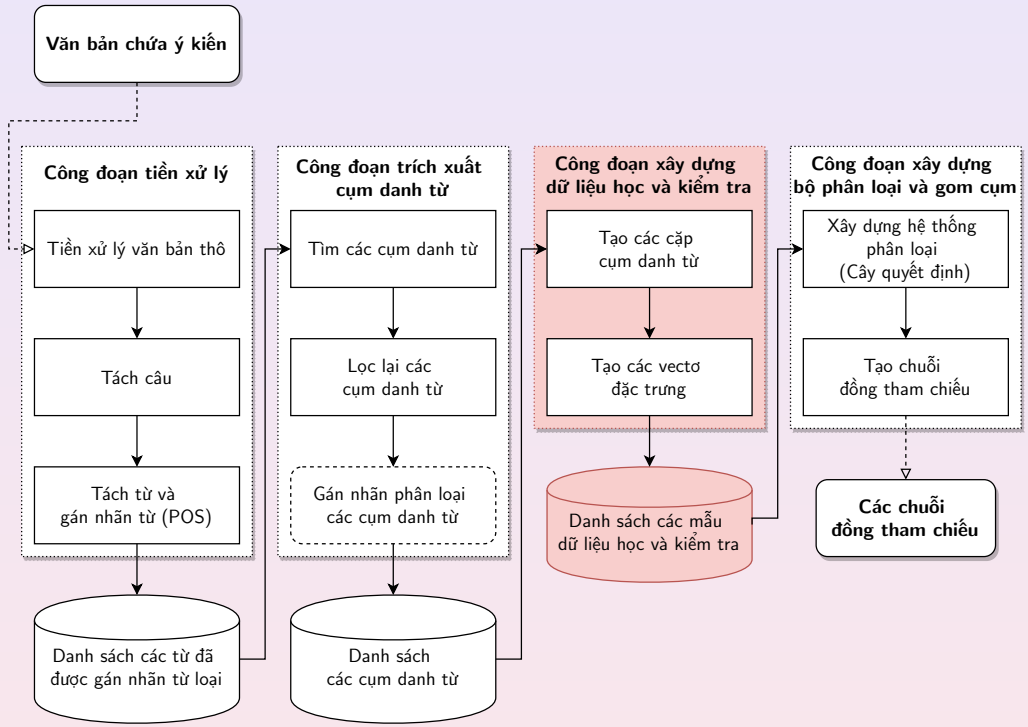
Ví dụ: Đoạn văn bản đã được gán nhãn: <0,-1,0 The Note 3> is a lot lighter than <1,-1,0 my HTC EVO>. <2,0,2 It>'s very fast and has <3,0,1 so many features> that <4,-1,0 an IPhone5> can't touch.

[3] http://crfchunker.sourceforge.net

Xây dựng dữ liệu học và kiểm tra

Tạo cặp các cụm danh từ

Tạo tất cả các cặp gồm hai cụm danh từ trong văn bản, trong đó ít nhất một trong hai cụm danh từ phải là đối tượng hoặc thuộc tính.



Tạo các vectơ đặc trưng

(<f1,f2,f3,. . . ,fn>, y)

- <f1,f2,f3,. . . ,fn>: tập đặc trưng
- Y: giá trị phân loại

Xây dựng dữ liệu học và kiểm tra (tt)

Nhóm đặc trưng	Đặc trưng	Giải thích
Nhóm liên quan khai khoáng ý kiến	Tính nhất quán về ý kiến	Bằng 1 nếu NP1 và NP2 có cùng thiên hướng ý kiến. Bằng 0 nếu chúng khác về thiên hướng ý kiến. Nếu không xác định được thiên hướng ý kiến cho NP1 hoặc NP2 thì bằng 2.
	Sự kết hợp giữa thực thể và từ chỉ ý kiến	Bằng 0,1,2,3,4,10 tùy vào xếp hạng PMI
Nhóm liên quan ngữ pháp	NP1 là đại từ	Bằng true nếu NP1 là đại từ, ngược lại bằng false
	NP2 là đại từ quan hệ	Bằng 1 nếu NP2 là đại từ quan hệ, ngược lại bằng false
	NP2 là đại từ	Bằng true nếu NP2 là đại từ, ngược lại bằng false
	Tính thống nhất về số	Bằng true nếu NP1 và NP2 thống nhất về số (trong ngữ pháp), ngược lại bằng false
	Từ hạn định	Bằng true nếu NP2 bắt đầu bằng "the", ngược lại bằng false
	Từ chỉ trỏ	Bằng true nếu NP2 bắt đầu bằng "this", "that", "these", those", ngược lại bằng false
	Cả NP1 và NP2 đều là tên riêng	Bằng true nếu NP1 và NP2 đều là tên riêng, ngược lại bằng false

Bảng: Các đặc trưng được sử dụng trong hệ thống

Xây dựng dữ liệu học và kiểm tra (tt)

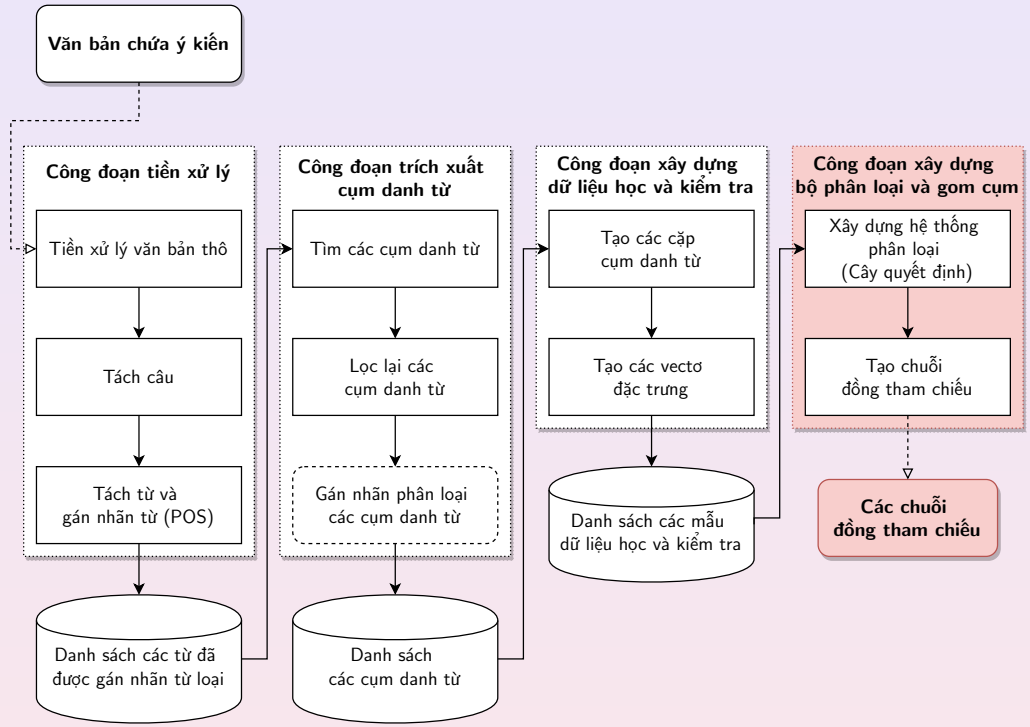
Nhóm đặc trưng	Đặc trưng	Giải thích
Nhóm liên quan từ vựng	Giống nhau hoàn toàn	Bằng 1 nếu NP1 và NP2 giống nhau hoàn toàn về mặt từ vựng
	Danh từ chính giống nhau	Bằng true nếu NP1 và NP2 có danh từ chính giống nhau, ngược lại bằng false
	Chuỗi con của nhau	Bằng true nếu NP1 và NP2 là chuỗi con (cha) của nhau, ngược lại bằng false
Khác	Khoảng cách	Khoảng cách về câu chứa NP1 và NP2. Nếu chúng nằm cùng một câu thì bằng 0
	Từ khóa "is" nằm ở giữa	Bằng true nếu có "is" không đi kèm với chỉ định so sánh nào nằm ở giữa NP1 và NP2, ngược lại thì bằng false
	Từ khóa "has" nằm ở giữa	Bằng true nếu có "has" nằm ở giữa NP1 và NP2, ngược lại thì bằng false

Bảng: Các đặc trưng được sử dụng trong hệ thống (tt)

Tạo bộ phân loại và gom cụm

Áp dụng giải thuật học máy

Giải thuật cây quyết định J48 (trên Weka) được sử dụng để phân loại cho các cặp ứng viên.



Gom cụm

Được thực hiện trên cơ sở tính chất bắc cầu của quan hệ đồng tham chiếu. Giả sử bộ phân loại cho kết quả các cặp cụm danh từ (A,B) và (B,C) đồng tham chiếu với nhau. Khi đó ta gom được (A,B,C) thành một cụm đồng tham chiếu.

Luận văn tốt nghiệp

Tổng quan đề tài

Phương pháp đề xuất

Tổng quan quy trình

Tiền xử lý

Trích xuất cụm danh từ

Xây dựng dữ liệu học và kiểm tra

Tạo bộ phân loại và gom cụm

Thực nghiệm và đánh giá

Tổng kết

Dữ liệu thực nghiệm

Dữ liệu được thu thập từ các bài đánh giá (review) trên *amazon.com* và các bài thảo luận (discussion) từ *http://www.howardforums.com*. Đây là các bài đánh giá và thảo luận về điện thoại. Tập dữ liệu gồm 157 bài, với mỗi bài có trung bình 7-8 câu.

> The fact that the GS5 is from Samsung makes it the Z2's biggest competitor. The Z2's strongest points are the side-mounted camera button. The Z2 has slightly louder speakers, slightly better battery life than the GS5. It has a fingerprint reader for easier unlocking, a better looking screen and a removable battery. Between the two phones, I'd pick the GS5 for its brighter display and for the ease of use the fingerprint reader brings. That said, watch out because the GS5 doesn't use on-screen menu buttons so handling it can be tricky unless you stick it in a case.

Hình: Ví dụ về một bài đánh giá (review) lấy từ amazon.com

Thực nghiệm và đánh giá (tt)

Tổng quan đề tài
Phương pháp đề xuất
Tổng quan quy trình
Tiền xử lý
Trích xuất cụm danh từ
Xây dựng dữ liệu học và kiểm tra
Tạo bộ phân loại và gom cụm
Thực nghiệm và đánh giá
Tổng kết

Các độ đo

Sử dụng 3 độ đo phổ biến để đánh giá hệ thống:

- P: độ đúng đắn (precision)
- R: độ phủ (recall)
- F: độ F (F-measure)

Các phương pháp

Sử dụng 3 hệ đo khác nhau để tính các độ đo trên, gồm:

- Hệ đo MUC
- Hệ đo B-CUBED
- Hệ đo CEAF-Φ_4

Luận văn tốt nghiệp

Tổng quan đề tài

Phương pháp đề xuất

Tổng quan quy trình

Tiền xử lý

Trích xuất cụm danh từ

Xây dựng dữ liệu học và kiểm tra

Tạo bộ phân loại và gom cụm

Thực nghiệm và đánh giá

Tổng kết

Các hệ thống được đánh giá

- Hệ thống cơ sở: Nhóm liên quan ngữ pháp + Nhóm liên quan từ vựng + Nhóm thuộc tính khác.
- Hệ thống SC: Hệ thống cơ sở + *Tính nhất quán về ý kiến (SC)*.
- Hệ thống EOA: Hệ thống cơ sở + *Sự kết hợp giữa thực thể và từ chỉ ý kiến (EOA)*.
- Hệ thống đầy đủ: Tất cả đặc trưng

Thực nghiệm và đánh giá (tt)

Luận văn tốt nghiệp

Tổng quan đề tài

Phương pháp đề xuất

Tổng quan quy trình

Tiền xử lý

Trích xuất cụm danh từ

Xây dựng dữ liệu học và kiểm tra

Tạo bộ phân loại và gom cụm

Thực nghiệm và đánh giá

Tổng kết

	Hệ đo MUC			Hệ đo B3			Hệ đo CEAF-Φ_4		
	P	R	F	P	R	F	P	R	F
Hệ thống cơ sở	0.757	0.534	0.621	0.794	0.524	0.626	0.621	0.557	0.586
Hệ thống SC	0.742	0.630	0.680	0.735	0.608	0.661	0.666	0.593	0.627
Hệ thống EOA	0.735	0.558	0.632	0.766	0.542	0.632	0.616	0.568	0.591
Hệ thống đầy đủ	0.730	0.632	0.676	0.724	0.610	0.658	0.661	0.594	0.626

Bảng: Kết quả thực nghiệm

Luận văn tốt nghiệp

Tổng quan đề tài

Phương pháp đề xuất

Tổng quan quy trình

Tiền xử lý

Trích xuất cụm danh từ

Xây dựng dữ liệu học và kiểm tra

Tạo bộ phân loại và gom cụm

Thực nghiệm và đánh giá

Tổng kết

Nhận xét

- Đặc trưng liên quan đến Khai khoáng ý kiến đã ảnh hưởng tích cực đến kết quả
- Đặc trưng *Tính nhất quán về ý kiến (SC)* đã có ảnh hưởng tương đối lớn đến hệ thống
- Đặc trưng *Sự kết hợp giữa thực thể và từ chỉ ý kiến (EOA)* ít có ảnh hưởng tích cực đến hệ thống
- Việc kết hợp hai đặc trưng mới *Tính nhất quán về ý kiến (SC)* và *Sự kết hợp giữa thực thể và từ chỉ ý kiến (EOA)* vẫn còn gặp vấn đề

Tổng kết

Kết quả đạt được

- Từ những kiến thức về Phân giải đồng tham chiếu và Khai khoáng ý kiến, đưa ra được phương pháp giải quyết bài toán.
- Hiện thực hệ thống dựa trên phương pháp đề xuất, cho kết quả đầu ra khả quan.

Khó khăn, hạn chế

- Dữ liệu tự thu thập có nhiều câu phức tạp, gây ảnh hưởng đến các đặc trưng liên quan Khai khoáng ý kiến.
- Do giới hạn về thời gian nên chưa kịp thử nghiệm cho tiếng Việt.

Tổng kết (tt)

Hướng phát triển

- Tìm thêm các đặc trưng mới liên quan đến Khai khoáng ý kiến để tăng hiệu suất hệ thống.
- Cải thiện đặc trưng Sự kết hợp giữa thực thể và từ chỉ ý kiến.
- Thử nghiệm cho tiếng Việt.