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 21 tháng 12 năm 2016

Tổng quan đề tài

Giới thiệu

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.

- 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)

Mục tiêu

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

Giả định của đề tài là đối tượng và thuộc tính của đối tượng đã được tìm ra, bài toán của đề tài là đi tìm các chuỗi đồng tham chiếu của chúng. Trong đó chỉ xét các quan hệ đồng tham chiếu dạng *đồng nhất (identity)*.

Tổng quan đề tài (tt)

Ví dụ phân giải đồng tham chiếu

I've had **my 8100** for about two years. Dropped **it** on pavement, concrete, wood floors, etc. at least 30 times so **it** has a ton of scratches but **it**'s held up very well. My 2-year old has chewed on **it**, licked **it**, smacked **it**, etc. – still no problems. **Signal** has been solid, **intuitive usability** is very good. **My standard battery** was great for the first 20 months (several hours of **talk time**, a week of **standby**) and just now is needing a charge every other day. I'll definitely buy a Sanyo again.

Ví dụ phân giải đồng tham chiếu (tt)

- Trong review trên, đối tượng được nói đến là một chiếc điện thoại **my 8100**. Nó bao gồm các thuộc tính là **a ton of scratches**, **Signal**, **intuitive usability**, **My standard battery**, **talk time** và **standby**. Trên thực thể và các thuộc tính này, người dùng có thể hiện ý kiến cá nhân lên nó. Ví dụ: *Signal has been solid*.
- Các từ/cụm từ **my 8100** và **it** (được bôi đỏ) trong ví dụ trên đồng tham chiếu với nhau, nó cùng chỉ tới một thực thể là chiếc điện thoại **my 8100**.

Khái quát về đồng tham chiếu

Giới thiệu

Phân giải đồng tham chiếu là một bài toán kinh điển trong XLNNTN. Hai từ hoặc cụm từ được gọi là đồng tham chiếu với nhau nếu chúng cùng chỉ đến một thực thể. Mục tiêu của bài toán phân giải đồng tham chiếu là tìm kiếm tất cả các từ hoặc cụm từ đồng tham chiếu với nhau đó.

Các hướng tiếp cận

Có 3 hướng tiếp cận phổ biến đối với bài toán Phân giải đồng tham chiếu:

- Phương pháp học có giám sát
- Phương pháp học không giám sát
- Phương pháp xây dựng hệ thống các luật

Khái quát về đồng tham chiếu (tt)

Phương pháp học có giám sát

- Hai giải thuật được sử dụng phổ biến: Cây quyết định và SVM
- Mô hình xây dựng tập huấn luyện: Mô hình xét theo cặp, mô hình hướng thực thể, mô hình xếp hạng.

Phương pháp học không giám sát

Giải thuật thường được sử dụng là giải thuật gom cụm. Mỗi từ, cụm từ có một tập các thuộc tính đặc trưng. Giải thuật gom cụm phân chia tập các từ, cụm từ thành các cụm, trong đó mỗi cụm chứa các từ, cụm từ có cùng đặc điểm về thuộc tính.

Khái quát về đồng tham chiếu (tt)

Phương pháp xây dựng hệ thống các luật

Xây dựng tập luật gồm các luật về cú pháp, ngữ nghĩa xác định khi nào thì hai từ, cụm từ đồng tham chiếu đến nhau. Luật đứng đầu tiên có độ chính xác cao nhất, sau đó giảm dần với các luật đứng sau. Văn bản được đưa qua các luật, bắt đầu từ luật đầu tiên. Sau khi áp dụng tất cả các luật trên văn bản, ta thu được kết quả cuối cùng là tập các cụm đồng tham chiếu.

Khái quát về khai khoáng ý kiến

Giới thiệu

Khai khoáng ý kiến (Opinion Mining) hay **Phân tích cảm xúc (Sentiment Analysis)** là một lĩnh vực nghiên cứu của Xử lý ngôn ngữ tự nhiên (XLNNTN), hướng đến việc tìm ra ý kiến/cảm xúc của chủ thể (người viết, người nói,...) đối với một thực thể nào đó thể hiện qua ngôn ngữ.

Các mô hình khai khoáng ý kiến

- Cấp độ văn bản (Document level)
- Cấp độ câu (Sentence level)
- Cấp độ thực thể - thuộc tính (Entity-Aspect level)

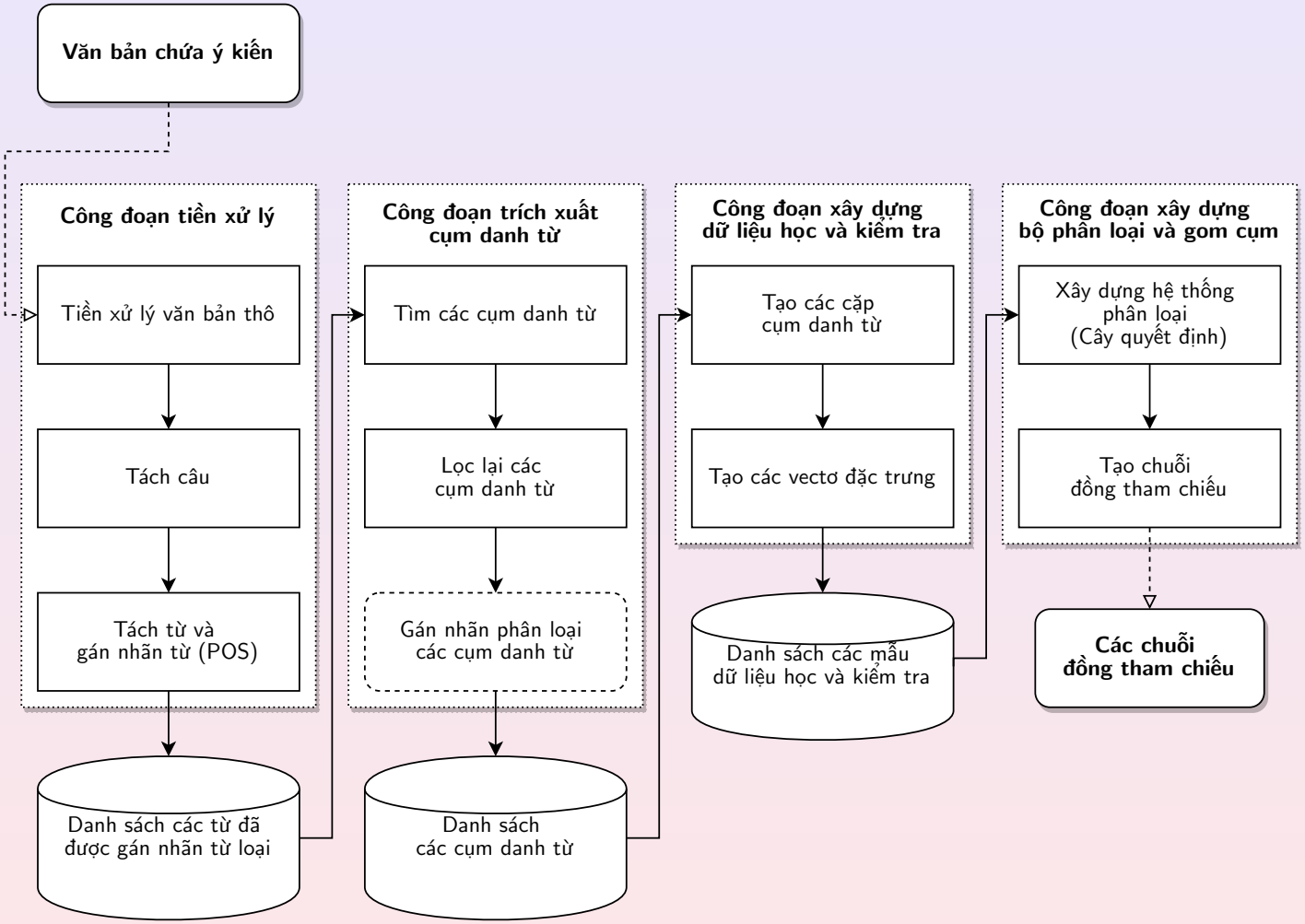
Khái quát về khai khoáng ý kiến (tt)

Khai khoáng ý kiến cấp độ thực thể - thuộc tính

- Ở cấp độ thực thể - thuộc tính, nhiệm vụ là tìm tất cả các ý kiến (Opinion) chứa trong văn bản: Một ý kiến bao gồm 5 thành phần (Theo Bing Liu trong cuốn Sentiment Analysis and Opinion Mining [1])
  (e,a,s,h,t)
  e: thực thể **a**: thuộc tính của thực thể **s**: ý kiến trên thực thể đó **h**: chủ thể thể hiện ý kiến **t**: thời điểm đưa ra ý kiến
- Ví dụ: Một ý kiến được rút trích từ *ví dụ 1* {my 8100, my battery, tích cực (great), người viết review, thời gian viết review}

[1] https://www.cs.uic.edu/ liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

Tổng quan quy trình



Hình: Tổng quan quy trình phân giải đồng tham chiếu

Tiền xử lý

Tiền xử lý văn bản thô

Chuẩn hóa dữ liệu thô để việc sử dụng công cụ Stanford xử lý có hiệu quả cao.
Ví dụ:
Tiền xử lý cho văn bản sau: *This phone is an excellent phone.It has a very good color screen, it is well disigned, the camera pretty good for a phone*

- Câu sau phải cách dấu chấm câu trước bằng ít nhất một dấu khoảng cách (Chỗ khoanh tròn) => Thêm dấu cách sau dấu chấm cuối câu trước.
- *disigned* sai chính tả => Sửa lỗi chính tả, thành *designed*.
- Dùng *the camera pretty good* là sai cú pháp văn phạm (*be pretty good*) => Chỉnh sửa về từ ngữ cho đúng văn phạm.

Tiền xử lý (tt)

Tách câu, tách từ và gán nhãn từ loại

Việc tách câu, tách từ đóng vai trò quan trọng trong các bước tiếp theo. Những công việc này thực hiện được là nhờ vào gói công cụ Stanford [2]. Các nhãn từ loại (part-of-speech) cũng được trích ra nhờ vào gói công cụ này, chúng được sử dụng theo như Penn Treebank POS Tag [3].

[2] http://stanfordnlp.github.io/CoreNLP
[3] http://web.mit.edu/6.863/www/PennTreebankTags.html

Trích xuất cụm danh từ

Tìm các cụm danh từ

Dựa vào kết quả gán nhãn từ loại và công cụ CRFChunker [4], tìm được các cụm danh từ trong các câu. Các cụm danh từ có khả năng nói tới đối tượng và thuộc tính, mục tiêu hướng đến của phép phân giải đồng tham chiếu.
Ví dụ: <This phone> is <an excellent phone>. <It> has <a very good color screen>, <it> is well designed, <the camera> is pretty good for <a phone>.

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. Đó là các cụm danh từ chỉ người, thời gian, số lượng,...

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

Trích xuất cụm danh từ (tt)

Gán nhãn cụm danh từ

- Mục đích của việc gán nhãn cụm danh từ là nhằm xác định cụm danh từ nào chỉ đến đối tượng, thuộc tính đồng thời biểu diễn sự tham chiếu giữa chúng. Qua đó việc tạo tập dữ liệu cho việc học và test một cách dễ dàng.
- Mỗi cụm danh từ được gán nhãn có dạng <ID,REF,TYPE NP>. Trong đó:
  - ID: chỉ vị trí của cụm danh từ trong câu.
  - REF: ID của cụm danh từ mà cụm danh từ hiện tại đang tham chiếu đến.
  - TYPE NP: là giá trị dùng để phân loại cụm danh từ NP. Tập giá trị: {0,1,2,3}.

Trích xuất cụm danh từ (tt)

Ví dụ 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. I love <5,-1,3 the camera > and <6,5,2 it> takes <7,-1,3 great pictures>.

Xây dựng dữ liệu học và kiểm tra

Cách tạo tập dữ liệu học và kiểm tra

Từ tập dữ liệu, ta tìm được một danh sách L gồm các cụm danh từ.

- Tập huấn luyện (Instances) bao gồm mọi cặp (NP1, NP2) được tạo ra từ danh sách L và thỏa mãn điều kiện NP1 hoặc NP2 là cụm danh từ chỉ đối tượng/thuộc tính.
- Do đề tài chỉ thực hiện với mục đích phân giải đồng tham chiếu các đối tượng/thuộc tính nên không xét trường hợp các cặp (NP1,NP2) mà cả NP1 và NP2 đều không chỉ đối tượng/thuộc tính.

Xây dựng dữ liệu học và kiểm tra (tt)

Ví dụ xây dựng dữ liệu học và kiểm tra

Xem xét đoạn văn bản đã được đánh dấu dưới đây: I have owned <0,-1,0 the Nokia 6101> since <1,0,1 late October 2005>. I absolutely love <2,0,0 it>! I wouldn't trade <3,2,0 it> for <4,0,1 the world>.

```
4,6,false,false,true,false,1,true,false,false,2,false,false,false,false,10,false,false
0,6,false,false,true,false,3,true,false,false,2,false,false,false,false,10,false,false
4,5,false,false,true,false,1,true,false,false,2,false,false,false,false,10,false,false
0,5,false,false,true,false,3,true,false,false,2,false,false,false,false,10,false,true
3,4,true,false,false,false,0,true,false,true,2,false,false,false,false,1,false,false
2,4,true,false,false,false,1,true,false,false,2,false,false,false,false,1,false,false
1,4,false,false,false,false,2,true,false,false,2,false,false,false,false,4,false,false
0,4,false,false,false,false,2,true,false,false,2,false,false,false,false,0,false,false
0,3,false,true,false,false,2,true,false,false,2,false,false,false,false,10,false,true
0,2,false,true,false,false,1,true,false,false,2,false,false,false,false,10,false,true
0,1,false,false,false,false,0,true,false,false,2,true,false,false,false,10,false,false
```

Hình: Các mẫu dữ liệu học và kiểm tra.

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 phương pháp chọn lại mẫu

Khắc phục tình trạng tập dữ liệu chênh lệch bằng cách dùng phương pháp chọn lại mẫu (resampling methods). Các thông số áp dụng (Dùng Resample trong WEKA):

- Z: thể hiện kích thước tập dữ liệu mới so với tập dữ liệu cũ. Chúng tôi chọn $Z = 100$, kích thước tập dữ liệu học mới bằng với tập dữ liệu học cũ.
- B: thể hiện mức độ chênh lệch dữ liệu trong tập dữ liệu mới. Chúng tôi lựa chọn $B = 0.1$.

Tạo bộ phân loại và gom cụm (tt)

Áp dụng giải thuật học máy

Áp dụng mô hình học có giám sát. Giải thuật cây quyết định J48 (trên Weka) được sử dụng để phân loại cho các cặp ứng viên.

Gom cụm

Dựa vào kết quả phân loại các cặp ứng viên, thực hiện gom cụm đồng tham chiếu. Việc gom cụm này 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.

Thực nghiệm và đánh giá

Chương trình

Hệ thống được hiện thực bằng ngôn ngữ Java (JDK1.8), dùng tới các thư viện của công cụ Stanford, Weka và code của công cụ CRFChunker. Chương trình được tổ chức thành các gói (package) sau:

- element
- gui
- process
- test
- util
- weka

Thực nghiệm và đánh giá (tt)

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)

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

Thực nghiệm và đánh giá (tt)

Các hệ thống được đánh giá

Các hệ thống được đánh giá bao gồm:

- Hệ thống cơ sở: Hệ thống sử dụng tập thuộc tính là các đặc trưng thuộc Nhóm liên quan ngữ pháp, Nhóm liên quan từ vựng và Nhóm thuộc tính khác.
- Hệ thống SC: Hệ thống cơ sở sử dụng thêm đặc trưng *Tính nhất quán về ý kiến (SC)*.
- Hệ thống EOA: Hệ thống cơ sở sử dụng thêm đặc trưng *Sự kết hợp giữa thực thể và từ chỉ ý kiến (EOA)*.
- Hệ thống đầy đủ: Hệ thống sử dụng tập thuộc tính là tất cả các đặc trưng đã nêu

Thực nghiệm và đánh giá (tt)

	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

Thực nghiệm và đánh giá (tt)

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

Thực nghiệm và đánh giá (tt)

Phân tích

Nguyên nhân đặc trưng *Sự kết hợp giữa thực thể và từ chỉ ý kiến (EOA)* chưa ảnh hưởng nhiều đến hệ thống:

- Số lượng từ chỉ ý kiến trong tập dữ liệu vẫn chưa nhiều.
- Các từ chỉ ý kiến trong tập dữ liệu vẫn chưa có tính đặc trưng cao.
- Tập dữ liệu học của đặc trưng EOA bị nhiễu.
- Theo giả thuyết cặp nào có giá trị đặc trưng EOA càng nhỏ thì càng có khả năng đồng tham chiếu. Tuy nhiên trên thực tế không hoàn toàn như vậy.

Thực nghiệm và đánh giá (tt)

Phân tích (tt)

Nguyên nhân đặc trưng *Tính nhất quán về ý kiến (SC)* chưa ảnh hưởng nhiều đến hệ thống:

- Giải thuật tìm thiên hướng ý kiến gắn với đối tượng/thuộc tính được nêu ra trong câu vẫn còn đơn giản và chỉ cho kết quả đúng đối với các câu tương đối đơn giản.

Một số lý do làm cho hệ thống vẫn chưa đạt kết quả cao:

- Hai đặc trưng liên quan đến Khai khoáng ý kiến vẫn chưa có nhiều ảnh hưởng đến hệ thống.
- Có những trường hợp một đối tượng được diễn tả bởi nhiều tên gọi khác nhau mà không đặc trưng nào trong hệ thống phát hiện ra được.
- Đặc trưng *Danh từ chính giống nhau* ảnh hưởng nhiều đến độ đúng đắn của hệ thống.

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 áp dụng mô hình 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 mô hình cho tiếng Việt.