Phân giải đồng tham chiếu dựa trên Ontology trong phân tích cảm xúc

Lê Thị Thủy, Phan Thị Tươi*, Quản Thành Thơ

***Tóm tắt*—Phân giải đồng tham chiếu thực thể và phân tích cảm xúc là hai bài toán độc lập khá phổ biến và được quan tâm rất nhiều trong cộng đồng xử lý ngôn ngữ tự nhiên. Tuy nhiên việc kết hợp cả hai bài toán thì vẫn chưa được quan tâm. Trong bài báo này, chúng tôi đề xuất một hướng ứng dụng cơ sở tri thức để giải quyết đồng tham chiếu đối tượng (thực thể) với khía cạnh có cảm xúc. Đồng thời, chúng tôi cũng xây dựng một mô hình cho bài toán phân giải đồng tham chiếu dựa trên Ontology trong phân tích cảm xúc của văn bản tiếng Anh. Cuối cùng bài báo đưa ra phương pháp đánh giá cho mô hình.**

***Từ khóa*—phân giải đồng tham chiếu, đối tượng và khía cạnh có cảm xúc, phân tích ý kiến, Ontology cảm xúc.**

1 GIỚI THIỆU

Việc xác định sự liên kết còn gọi là sự tham chiếu của các cụm từ cùng chỉ đến một đối tượng cụ thể trong xử lý ngôn ngữ tự nhiên (NLP) gọi là bài toán phân giải đồng tham chiếu (CR).

Hiện nay, với công nghệ Internet và nhu cầu mua sắm của con người càng cao thì những đoạn văn bản có nhiều ý kiến về các sản phẩm trên các trang web ngày một phong phú. Những đoạn văn bản có ý kiến đó còn gọi là văn bản có cảm xúc.

Ngày nhận bản thảo: 10-4 -2017, ngày chấp nhận đăng: 05 -10-2017.

Chúng tôi xin được cảm ơn công ty YouNet Media đã hỗ trợ tập dữ liệu văn bản cho phần thực nghiệm của bài báo..

Lê Thị Thủy, Phan Thị Tươi, Quản Thành Thơ - Khoa Khoa học và Kỹ thuật Máy tính, Trường Đại học Bách Khoa - ĐHQG-HCM. Số 268 Lý Thường Kiệt, Phường 14, Quận 10, Hồ Chí Minh

(E-mail: tuoi@cse.hcmut.edu.vn)

Ví dụ 1: "[1]I have just bought a Samsung Galaxy Note7. [2]I like it because it looks beautiful. [3]However, it is expensive. [4]It has a camera. [5]I took a photo and it is amazing."

Áp dụng CR, xác định được các chuỗi đồng tham chiếu: Core1(a Samsung Galaxy Note7[1], it[2,3], It[2,5], It[3,2], It[4,1]); Core2(Photo[5,4], It[5,6]). Áp dụng bài toán phân tích cảm xúc, xác định được các cặp cảm xúc: Sen1(It[2,5], beautiful[2]); Sen2(It[3,2], expensive[3]); Sen3(It[5,6], amazing[5]). Trong Sen1, "beautiful" là ý kiến tích cực của từ "It" trong câu 2, vị trí thứ 5. Trong Sen2, "expensive" là ý kiến tiêu cực của "It" trong câu 3, vị trí thứ 2. Trong Sen3, "amazing" là ý kiến tích cực của "It" trong câu 5, vị trí thứ 6. Kết hợp hai bài toán, nghĩa là kết hợp Core1 với Sen1 và Sen2, ta có ý kiến về "a Samsung Galaxy Note7" là "beautiful" và "expensive". Kết hợp Core2 và Sen3, xác định được "Photo" là "amazing".

Với ví dụ 1, người đọc sẽ cảm nhận được năm câu của đoạn văn bản trên đều đề cập đến "Samsung Galaxy Note7" nhờ các từ "it" ở câu 2, câu 3 và câu 4, "Photo" trong câu 5 và "It" trong câu 5. Vậy vấn đề trong NLP đó là xác định được "Photo" là một khía cạnh của "Samsung Galaxy Note7", từ "beautiful" là ý kiến chỉ thuộc tính thiết kế và "expensive" chỉ thuộc tính giá thành của "Samsung Galaxy Note7".

Để thực hiện vấn đề này, tác giả đề xuất sử dụng cơ sở tri thức chuyên biệt giải quyết đồng tham chiếu giữa đối tượng với khía cạnh dựa theo công trình [1]. Tiếp theo đề xuất đồ thị đồng tham chiếu để tập hợp kết quả của hai bài toán cảm xúc và đồng tham chiếu, cuối cùng đưa ra các bộ đồng tham chiếu đối tượng với khía cạnh có cảm xúc.

Cấu trúc của bài báo như sau: phần 2 giới thiệu các nghiên cứu liên quan của bài toán CR và phân tích cảm xúc. Phần 3 đưa ra các đề xuất của bài báo: xây dựng Ontology cảm xúc về smartphone, mô hình CR dựa trên Ontology trong phân tích cảm xúc và đồ thị đồng tham chiếu. Phần 4 là kết quả thực nghiệm trên 100 văn bản có ý kiến về smartphone. Phần 5 đánh giá kết quả thực nghiệm của mô hình. Cuối cùng là kết luận và hướng phát triển của bài báo.

2 CÁC NGHIÊN CỨU LIÊN QUAN

Phân giải đồng tham chiếu

Vấn đề đồng tham chiếu được rất nhiều nhà nghiên cứu NLP quan tâm chủ yếu trên CR cụm danh từ, đại từ và thực thể có tên. Có rất nhiều cách tiếp cận để giải quyết vấn đề đồng tham chiếu, cụ thể:

- Phương pháp học máy có giám sát [2]; bán giám sát hoặc không giám sát [3];
- Phương pháp dựa trên đặc tính ngữ nghĩa của ngôn ngữ: từ vựng, cú pháp [4];
- Phương pháp dựa vào đồ thị [5];
- Sử dụng Knowledge Graph, Ontology [6];
- Các mô hình dựa theo luật [7].

Phân tích cảm xúc mức khía cạnh

Phân tích cảm xúc mức khía cạnh là xác định các ý kiến về thực thể ở từng đặc tính của nó. Bài toán được giải quyết theo nhiều hướng như mô hình hóa chủ đề [8]; Probabilistic Latent Semantic Analysis (PLSA) [9]; Latent Dirichlet Analysis (LDA) [10]. Ngoài ra, bài toán phân tích cảm xúc còn sử dụng Ontology chuyên biệt kết hợp với các luật ngôn ngữ để xử lý các từ cảm xúc [1]. Hiện nay đã ra đời nhiều công cụ phân tích cảm xúc như Trackur, SAS, Opentext, Statsoft, Clarabridge, TheySay, NetOwl, NICTA, Sentiment Analysis của Stanford, …

Bộ công cụ Stanford CoreNLP

Stanford CoreNLP là một bộ công cụ NLP khá lớn của [11], được sử dụng rộng rãi cả trong nghiên cứu NLP, trong thương mại và chính trị. Bộ công cụ này có kiến trúc đầy đủ các thành phần NLP, trong đó có hai tầng Coreference Resolution và Other Annotators (sentiment).

Tầng Coreference Resolution, [12] thực hiện CR cụm danh từ, đại từ và các thực thể có tên. Các tác giả kết hợp các hệ thống trên cơ sở luật, có giám sát và không giám sát. Mục tiêu của tầng Coreference Resolution là đơn giản, hướng đến độ chính xác từ cao nhất đến thấp nhất.

Tầng gán nhãn cảm xúc (Other Annotators - sentiment), [13] sử dụng ngân hàng cây có nhãn cảm xúc (Stanford Sentiment Treebank) và Recursive Neural Tensor Network - RNTN để phân lớp các câu từ rất tiêu cực đến rất tích cực thể hiện bằng các ký hiệu: --, -, 0, +, ++.

3 PHÂN GIẢI ĐỒNG THAM CHIẾU ĐỐI TƯỢNG VỚI KHÍA CẠNH, CẢM XÚC (OBJECT ASPECT SENTIMENT COREFERENCE - OBASCORE).

Các khái niệm của một số thuật ngữ sau được sử dụng trong bài báo này.

Đối tượng (Object) là một khái niệm chỉ đến một thực thể hay tên riêng của một vật cụ thể.

Khía cạnh (Aspect) là một khái niệm đề cập đến một thành phần (component) hay một thuộc tính (tính chất - attribute) của một đối tượng.

Cảm xúc (Sentiment) là những khái niệm gồm những từ mang suy nghĩ chủ quan, ý kiến về một khía cạnh của đối tượng.

Sau khi nghiên cứu công cụ Stanford CoreNLP, tác giả nhận thấy bộ công cụ này chưa giải quyết được hai vấn đề. Thứ nhất, chưa xác định được cảm xúc của khía cạnh ẩn; Thứ hai, chưa xác định các khía cạnh thuộc về đối tượng nào trong văn bản. Để khắc phục những hạn chế nêu trên, bài báo đề xuất phương pháp ứng dụng Ontology cảm xúc hỗ trợ CR trên bộ công cụ Stanford CoreNLP.

Ontology cảm xúc hỗ trợ phân giải đồng tham chiếu

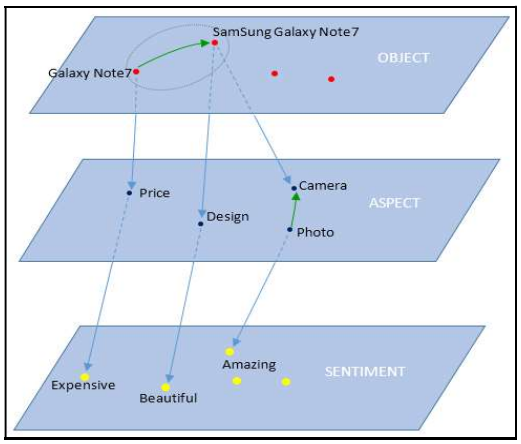
Tác giả xây dựng Ontology cảm xúc cho smartphone trên cơ sở áp dụng và phát triển công

trình [1], bằng hai tập (C, R).

Trong đó C = (C^O, C^A, C^S) là tập các tập các khái niệm. $\mathbf{C^O}$ là tập khái niệm của đối tượng (Object), ví dụ: Samsung, IPhone, Oppo, …; $\mathbf{C^A}$ là tập khái niệm của khía cạnh (Aspect). Khía cạnh có hai lớp con Component (camera, battery, …) và Attribute (design, price, …); $\mathbf{C^S}$ là tập khái niệm cảm xúc (Sentiment) có các trị thuộc các lớp tích cực (Positive), tiêu cực (Negative) và trung hòa (Neutral).

R = (R^T, R^N, R^S) là tập các tập quan hệ giữa các class với nhau, giữa các cá thể trong cùng class hoặc khác class. R^T là tập các quan hệ có phân cấp cha con (subconcept-of); R^N là tập các quan hệ không phâp cấp cha con (hasAttribute, hasComponent,…); R^S là tập các quan hệ cảm xúc (isPositive, isNegative, isNeutral).

Cá thể là thực thể hay đối tượng cụ thể: của đối tượng: Galaxy J3, Oppo A37, …; của khía cạnh: price, design, camera, …; của cảm xúc: cheap, expensive, beautiful,…



Hình 1. Kiến trúc Ontology cảm xúc

Kiến trục Ontology cảm xúc được minh họa ở hình 1, có ba lớp: Object, Aspect và Sentiment. Đường mũi tên giữa các cá thể trong cùng một lớp hoặc giữa các lớp khác nhau thể hiện sự đồng tham chiếu giữa đối tượng - khía cạnh, khía cạnh - cảm xúc.

Sau khi có được tập đồng tham chiếu thực thể, tập cảm xúc và Ontology cảm xúc hỗ trợ phân giải đồng tham chiếu, tác giả kết hợp ba thành phần này trong đồ thị đồng tham chiếu.

Đồ thị đồng tham chiếu (CoReference Graph - CRG)

Đồ thị CRG là một đồ thị có hướng được biểu diễn bằng cặp (V, E):

- V là tập các đỉnh chứa các cá thể của lớp đối tượng, lớp khía cạnh, lớp cảm xúc và các cụm danh từ, các đại từ biểu diễn đối tượng hay khía cạnh.

- E là tập các cung nối các đỉnh, có hướng thể hiện ba mối quan hệ đồng tham chiếu: tham chiếu thực thể (Core), tham chiếu cảm xúc (Sent) và tham chiếu khía cạnh (Asp).

- Trọng số của đồ thị thể hiện khoảng cách giữa các đỉnh, với các đỉnh thuộc lớp Object và Aspect hoặc giữa lớp Aspect và Sentiment có trọng số bằng 1, giữa lớp Object và Sentiment có trọng số bằng 2, giữa các đỉnh cùng một lớp (đồng tham chiếu thực thể) có trọng số 0.

Các tính chất của đồ thị:

- Các đỉnh không trùng nhau, khác nhau về từ (tiếng Anh), vị trí trong câu và vị trí câu.

- Đồ thị CRG có thể có từ hai đồ thị con trở lên.

- Nếu quan hệ Sent(v1,v2) có v1 là các cụm danh từ hoặc đại từ đại diện cho các cá thể của đối tượng, khía cạnh, và v2 chỉ có thể là các cá thể của cảm xúc.

- Nếu quan hệ Asp(v1,v2) có v1 là các cụm danh từ, đại từ đại diện cho các cá thể của đối tượng thì v2 chỉ có thể là các cá thể của khía cạnh.

- Đồ thị CRG sẽ có các đỉnh treo là các cá thể của cảm xúc hoặc các đại từ.

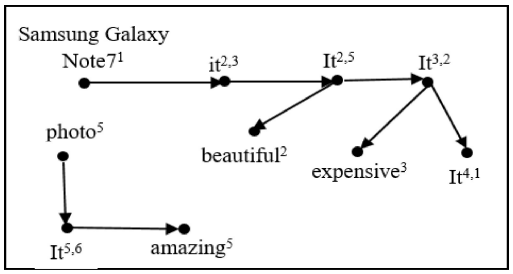
Xét lại ví dụ 1: "[1]I have just bought a Samsung Galaxy Note7. [2]I like it because it looks beautiful.

[3]However, it is expensive. [4]It has a camera. [5]I took a photo and it is amazing." Sử dụng bộ công cụ Stanford CoreNLP, ta có được kết quả đồng tham chiếu và gán nhãn cảm xúc như sau: Core1(a Samsung Galaxy Note7[1], it[2,3], It[2,5], It[3,2], It[4,1]); Core2(Photo[5,4], It[5,6]); Sen1(It[2,5], beautiful[2]); Sen2(It[3,2], expensive[3]); Sen3(It[5,6], amazing[5]).

Từ kết quả của Stanford, ta có CRG = (V, E) được minh họa ở hình 2, trong đó:

V = {Samsung Galaxy Note7[1], it[2,3], It[2,5], It[3,2], It[4,1], Photo[5], beautiful[2], expensive[3], amazing[5]}

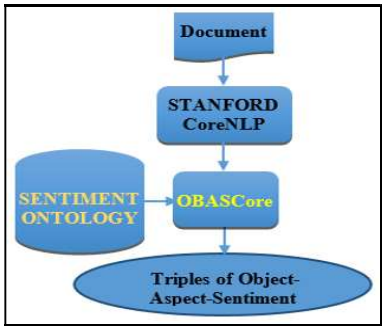
E = {Core(a Samsung Galaxy Note7[1], it[2,3]); Core(it[2,3], It[2,5]); Core(It[2,5], It[3,2]); Core(It[3,2], It[4,1]); Core2(photo[5], It[5,6]) Sen1(It[2,5], beautiful[2]); Sen2(It[3,2], expensive[3]); Sen3(It[5,6], amazing[5])}



Hình 2. Đồ thị CRG của ví dụ 1 từ kết quả của bộ Stanford CoreNLP

Mô hình phân giải đồng tham chiếu OBASCore

Mô hình CR đối tượng - khía cạnh - cảm xúc được đề xuất trình bày ở hình 3.



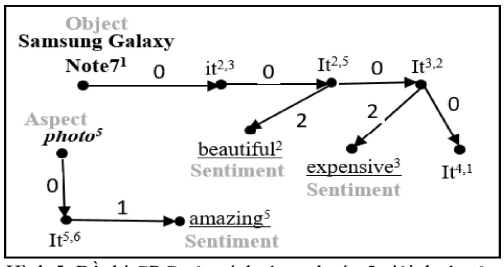
Hình 3. Mô hình phân giải đồng tham chiếu OBASCore

Mô hình có các mô đun: *Document* là văn bản có ý kiến về smartphone; *Stanford CoreNLP* là bộ công cụ CR; *Ontology* là một cơ sở tri thức có cảm xúc về smartphone; *OBASCore* là mô đun do tác giả đề xuất, xử lý kết quả xuất ra của *Stanford CoreNLP*. *OBASCore* sử dụng Ontology để xác định các tập đồng tham chiếu giữa đối tượng với khía cạnh có cảm xúc (Triples of Object-Aspect-Sentiment). Giải thuật mô tả chức năng của mô đun OBASCore được trình bày ở hình 4.

1. Khởi tạo CRG
2. Thêm đỉnh và cung từ tập C
3. Thêm đỉnh, cung từ tập S không trùng đỉnh
4. Phân loại các đỉnh theo các class của Ontology
5. Cập nhật trọng số cho các cung theo định nghĩa CRG
6. Thêm cạnh giữa các đỉnh đồng tham chiếu trong cùng class dựa trên Ontology
7. Xét các đỉnh đầu v không là đỉnh cuối:
   Tính tổng trọng số từ v đến các đỉnh treo.
   Nếu tổng trọng số bằng 2, thêm đỉnh là khía cạnh trong Ontology tương ứng với đỉnh treo.
   Nếu tổng trọng số bằng 1, loại bỏ đỉnh trung gian.
   Ngược lại, loại bỏ đỉnh treo.
8. Thêm cung giữa đối tượng với các khía cạnh và gán trọng số bằng 1
9. Xét các đỉnh thuộc tập V: nếu cung (v1,v2) có trọng số bằng 0 thì loại bỏ v2.

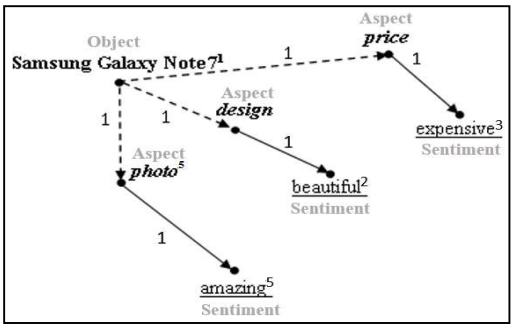
Hình 4. Giải thuật của mô đun OBASCore

Áp dụng thuật toán cho ví dụ 1 đến bước 3 và 4, ta có đồ thị như hình 2. Sau bước 4 và 5, đồ thị hình 2 được gán giá trị trọng số cho các cung và các đỉnh được phân loại theo lớp tương ứng trong Ontology. Kết quả minh họa ở hình 5, với đỉnh thuộc lớp đối tượng là **Samsung Galaxy Note7**; đỉnh thuộc lớp khía cạnh là ***photo***; đỉnh thuộc lớp cảm xúc là beautiful, expensive, amazing; các đỉnh còn lại không thuộc lớp nào trong Ontology.



Hình 5. Đồ thị CRG của ví dụ 1 sau bước 5 giải thuật của mô đun OBASCore

Thực hiện tiếp bước 6, 7, 8 và 9 của giải thuật OBASCore ở hình 5, ta có đồ thị cuối cùng của ví dụ 1 như hình 6. Khi đó đồ thị CRG xuất hiện hai đỉnh ***price*** và ***design*** được xác định dựa vào hai từ cảm xúc beautiful và expensive thông qua Ontology cảm xúc. Cung nét đứt thể hiện đồng tham chiếu giữa đối tượng - khía cạnh. Cung nét liền thể hiện tham chiếu khía cạnh - cảm xúc.



Hình 6. CRG của ví dụ 1 được thực hiện bởi giải thuật của mô đun OBASCore

Đồ thị CRG xác định Samsung Galaxy Note7 có ba cặp khía cạnh - cảm xúc: photo - amazing, design - beautiful, price - expensive.

4 KẾT QUẢ THỰC NGHIỆM.

Giải thuật OBASCore chạy trên tập dữ liệu với 100 đoạn văn bản có cảm xúc về các smartphone được cung cấp bởi công ty YouNet Media (http://www.younetmedia.com/) chuyên về phân tích trực tuyến. Kết quả thu được trình bày ở bảng 1. Tập dữ liệu này được chia thành ba dạng sau:

Dạng 1: Có khía cạnh

a) "I have just bought a Samsung Galaxy Note7. Its design is beautiful. The price is expensive."

b) "Galaxy Note 5 is a perfect phone. I like it. Note 5 has a 2GB RAM. It is strong and powerful."

Dạng 2: Không rõ khía cạnh

a) "I have just bought a Samsung Galaxy Note7. It is expensive."

b) "I have just bought a Samsung Galaxy Note7. It is amazing."

Dạng 3: Có đối tượng và không có từ đồng tham chiếu trực tiếp

a) "The Samsung Galaxy S5 is very beautiful. The price is not cheap."

b) "I bought my Galaxy S5 from store yesterday. I loved the screen. It is so attractive."

Bảng 1. Kết quả thực nghiệm cho 100 đoạn văn bản có cảm xúc về smartphone

TT	Các dạng câu	Số câu	Kết quả	
			Đúng	Sai
1	Có khía cạnh	45	33	12
2	Không rõ khía cạnh	45	39	6
3	Có đối tượng và không có từ đồng tham chiếu trực tiếp	10	5	5

Với kết quả ở bảng 1, dạng 1 có lỗi vì văn bản ở dạng 1b có động từ sở hữu "has/có" (*Note 5 has a 2GB RAM)* khi đó CR không rút trích được "2GB RAM" nên "It" ở câu 5 (dạng 1b) không tham chiếu đến "2GB RAM" dẫn đến kết quả phân tích cảm xúc cho khía cạnh không chính xác.

Dạng 2 xảy ra trường hợp phân tích cảm xúc không đúng vì một từ cảm xúc có thể đề cập đến nhiều khía cạnh. Ví dụ ở dạng 2b, từ "amazing" có thể chỉ đến khía cạnh là một tấm hình "photo" hay khía cạnh là chất lượng "quality" của "*Samsung Galaxy Note7"*. Đây là sự nhập nhằng nghĩa của từ cảm xúc.

Dạng 3 có thể xảy ra trường hợp "*bị sót"* đối tượng vì không có từ đồng tham chiếu trực tiếp, đối tượng được nhắc lại bằng cách sử dụng khía cạnh của nó ở câu tiếp theo và CR không xác định được. Nếu đối tượng được đề cập trực tiếp có cảm

xúc, thì việc phân tích cảm xúc đã giúp đồ thị CRG không "bỏ sót" đối tượng (dạng 3a).

Ngoài lỗi xuất hiện ở ba dạng văn bản trên thì Ontology cũng có thể là nguyên nhân chủ quan gây ra lỗi. Khi Ontology không đầy đủ tri thức thì việc tìm kiếm và suy luận dẫn đến kết quả không chính xác.

5 ĐÁNH GIÁ KẾT QUẢ THỰC NGHIỆM.

Mô hình trình bày ở hình 3 có kết quả thực nghiệm phụ thuộc vào kết quả đồng tham chiếu thực thể, kết quả phân tích cảm xúc và đồ thị đồng tham chiếu kết hợp Ontology. Vì vậy tác giả đề xuất phương pháp đánh giá mô hình như sau: Tính độ truy hồi R và độ chính xác P cho từng kết quả theo số cặp đồng tham chiếu, số cặp cảm xúc và số bộ đối tượng - khía cạnh - cảm xúc của một văn bản. Sau đó tính trung bình cộng trên tập dữ liệu có n văn bản. Áp dụng phương pháp này cho tập dữ liệu có 100 câu như bảng 1, kết quả đánh giá của mô hình 3 thu được như bảng 2.

Bảng 2. Kết quả đánh giá thực nghiệm của mô hình OBASCore với 100 văn bản

Độ đo(1)	**Coreference** (2)	**Sentiment** (3)	**Ontology + CRG** (4)
R	0,83	0,88	**0,79**
P	0,85	0,89	**0,76**

Trong bảng 2, cột 4 là kết quả đánh giá cuối cùng của mô hình. Với kết quả như bảng 2, phương pháp CR dựa trên Ontology trong phân tích cảm xúc cho dạng câu đơn và câu ghép đạt kết quả tương đối cao.

Hiệu quả của Ontology kết hợp CRG sẽ giảm so với coreference và sentiment nhưng không đáng kể. Nếu một trong hai đầu vào không chính xác thì đầu ra của OBASCore cũng sẽ sai và ngay cả khi đầu vào đúng thì kết quả của OBASCore cũng có thể sai do Ontology có thể thiếu tri thức. Tuy nhiên, so với đánh giá của thuật toán gốc Coreference (78,93%) và Sentiment (80,7%) thì kết quả của bài báo cao hơn, cụ thể như bảng 2.

6 KẾT LUẬN

Ứng dụng Ontology cảm xúc cho bài toán phân tích cảm xúc mức khía cạnh kết hợp với CR xác định được đối tượng cụ thể có khía cạnh và cảm xúc của người viết trong một văn bản.

Tuy nhiên với sự kết hợp này còn một số hạn chế như bỏ sót đối tượng khi không có đồng tham chiếu; phân giải đồng tham chiếu không chính xác khi có những giới từ phủ định; rút trích cụm danh từ chưa đầy đủ. Những hạn chế này sẽ được tác giả tiếp tục nghiên cứu và giải quyết. Ngoài ra để nâng cao hiệu quả đồng tham chiếu giữa đối tượng - khía cạnh, khía cạnh - cảm xúc cần phải phát triển, mở rộng Ontology cảm xúc.

TÀI LIỆU THAM KHẢO

[1] Tung Thanh Nguyen, Tho Thanh Quan, Tuoi Thi Phan, "Sentiment search: an emerging trend on social media monitoring systems", *ASLIB Journal of Information Management, Vol. 66 Iss: 5, ISSN: 2050-3806, SCI-E,* 2014.

[2] Kevin Clark and Christopher D. Manning. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. *Association for Computational Linguistics (ACL),* 2016.

[3] Ng, Vincent. "Unsupervised models for coreference resolution". *In Proceedings of EMNLP, pp. 640–649, Honolulu, HI, 2008.*

[4] Aria Haghighi and Dan Klein. "Simple Coreference Resolution with Rich Syntactic and Semantic Features". *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.* 2009

[5] Cristina Nicolae and Gabriel Nicolae, BESTCUT: A Graph Algorithm for Coreference Resolution. *EMNLP 2006.*

[6] Prokofyev, R., Tonon, A., Luggen, M., Vouilloz, L., Difallah, D. E., & Cudré-Mauroux, P., Sanaphor: Ontology-based coreference resolution. In *International Semantic Web Conference* (458-473). 2015.

[7] Lappin, Shalom and Herbert Leass. "An algorithm for pronominal anaphora resolution". *Computational Linguistics, 20(4): 535–561.* 1994.

[8] Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen, "Opinion Word Expansion and Target: Extraction through Double Propagation", *Computational Linguistics Vol.37, No.1,* Pages 9-27. 2011.

[9] Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. Topic sentiment mixture: modeling facets and opinions in

weblogs. In *Proceedings of the 16th international conference on World Wide Web* (pp. 171-180). *ACM. 2007, May*.

[10] Zhao, W.X., Jiang, J., Yan, H., & Li, X. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (56-65) 2010.

[11] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S. and McClosky, D.,"The Stanford CoreNLP Natural Language Processing Toolkit". *In ACL (System Demonstrations) (pp. 55-60)* 2014.

[12] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu and Dan Jurafsky, "Deterministic coreference resolution based on entity-centric, precision-ranked rules". *Computational Linguistics 39(4),* 2013.

[13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. "Recursive deep models for semantic compositionality over a sentiment tree-bank". *InEMNLP 2013, pages 1631–1642.*

Lê Thị Thủy là nghiên cứu sinh của Trường Đại học Bách Khoa, ĐHQG - HCM. Hiện nay, Lê Thị Thủy là giảng viên trường Đại Học Công nghiệp Tp.HCM, 12 Nguyễn Văn Bảo, Q. Gò Vấp, HCM. Email: lethithuyit@iuh.edu.vn

Phan Thị Tươi là Giáo sư Tiến sĩ công tác tại khoa Khoa học và Kỹ thuật Máy tính, Trường Đại học Bách Khoa, ĐHQG-HCM. Phan Thị Tươi nhận bằng Tiến sĩ Khoa học Máy tính từ Đại học Charles, Cộng hòa Séc năm 1985. Các hướng nghiên cứu bao gồm Trình biên dịch, Truy vấn thông tin và Xử lý ngôn ngữ tự nhiên. Phan Thị Tươi là nghiên cứu viên chính của các dự án trọng điểm cấp Quốc gia và đã xuất bản nhiều bài báo trên các tạp chí và hội nghị uy tín Quốc gia và Quốc tế.

Quản Thành Thơ hiện là giảng viên của Trường Đại học Bách Khoa, ĐHQG-HCM, 268 Lý Thường Kiệt, Q.10, Tp. Hồ Chí Minh.

Coreference resolution Ontology-based in sentiment analysis

Le Thi Thuy, Phan Thi Tuoi*, Quan Thanh Tho
Ho Chi Minh City University of Technology, VNU-HCM
Corresponding author: tuoi@cse.hcmut.edu.vn

Receive: 10-4-2017, Accepted: 05-10-2017

***Abstract*—Entity co-reference resolution and sentiment analysis are independent problems and popular research topics in the community of natural language processing. However, the combination of those two problems has not been getting much attention. Thus, this paper susgests to apply knowledge base to solve co-reference between object and aspect with sentiment. In addition, the paper also proposes the model of Ontology-based co-reference resolution in sentiment analysis for English text. Finally, we also discuss evaluation methods applied for our model and the results obtained.**

***Index Terms*—co-reference resolution; object and aspect with sentiment; sentiment analysis; sentiment Ontology.**