

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC BÁCH KHOA**

**LÊ THỊ THỦY**

**PHÂN GIẢI ĐỒNG THAM CHIẾU ĐỐI TƯỢNG  
CHO PHÂN TÍCH CẢM XÚC**

Ngành: KHOA HỌC MÁY TÍNH

Mã số ngành: 62.48.01.01

**TÓM TẮT LUẬN ÁN TIẾN SĨ**

**TP. HỒ CHÍ MINH - NĂM 2020**

Công trình được hoàn thành tại **Trường Đại học Bách Khoa – ĐHQG-HCM**

Người hướng dẫn 1: GS. TS. Phan Thị Tươi

Phản biện độc lập 1:

Phản biện độc lập 2:

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án sẽ được bảo vệ trước Hội đồng đánh giá luận án họp tại

.....  
.....

vào lúc            giờ            ngày            tháng            năm

Có thể tìm hiểu luận án tại thư viện:

- Thư viện Trường Đại học Bách Khoa – ĐHQG-HCM
- Thư viện Đại học Quốc gia Tp.HCM
- Thư viện Khoa học Tổng hợp Tp.HCM

# CHƯƠNG 1 GIỚI THIỆU

## 1.1 Động cơ nghiên cứu

Phân giải đồng tham chiếu là một bài toán xuất hiện khá phổ biến trong lĩnh vực Xử lý ngôn ngữ tự nhiên. Trong mỗi văn bản, một đối tượng hay khía cạnh của một đối tượng có thể được nhắc đi nhắc lại nhiều lần bởi chính cụm từ ban đầu hoặc có thể là những cụm từ khác nhưng người đọc vẫn tham chiếu được đối tượng hay khía cạnh đó. Việc xác định sự liên kết hay 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 gọi là bài toán phân giải đồng tham chiếu.

Hiện nay, do sự đa dạng về mặt hàng cùng loại sản phẩm, nhu cầu của con người càng ngày khắt khe trong việc lựa chọn sản phẩm và với công nghệ Internet, mạng xã hội toàn cầu phát triển mạnh nên nguồn dữ liệu văn bản có nhiều ý kiến về các sản phẩm trên các trang web ngày một giàu và phong phú. Những đoạn văn bản có ý kiến đó gọi là văn bản có cảm xúc và việc quan tâm đến vấn đề phân tích văn bản cảm xúc cũng là một nhu cầu tất yếu của thực tiễn và khoa học.

Phân tích cảm xúc (Sentiment analysis) hay còn gọi là khai phá ý kiến (Opinion mining) là một lĩnh vực phân tích ý kiến, tình cảm, sự ước lượng, sự đánh giá, thái độ và cảm xúc của con người về các thực thể như sản phẩm, dịch vụ, tổ chức, cá nhân, vấn đề, sự kiện, chủ đề và các thuộc tính của chúng. Bài toán phân tích cảm xúc được chia thành ba mức: mức văn bản, mức câu, và mức khía cạnh. Tuy có nhiều phương pháp tiếp cận cho bài toán phân tích cảm xúc ở các mức, nhưng các tiếp cận này cũng chỉ mới dừng ở kết quả văn bản, câu hay khía cạnh và đối tượng được đề cập có tính phân cực gì về cảm xúc mà vẫn chưa quan tâm mối quan hệ giữa cảm xúc, khía cạnh và đối tượng. Vấn đề nghiên cứu còn để mở này chính là động cơ nghiên cứu của luận án.

## 1.2 Mục tiêu và phạm vi nghiên cứu

**Mục tiêu của luận án** là xây dựng mô hình phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc sử dụng các giải thuật xử lý ngôn ngữ tự nhiên, sự hỗ trợ của ontology cảm xúc và các phương pháp máy học, học sâu để giải quyết bài toán phân giải đồng tham chiếu nhiều đối tượng cho phân tích cảm xúc mức khía cạnh.

Việc xác định cảm xúc cho một khía cạnh của một đối tượng cụ thể cho thấy chi tiết hơn về chất lượng của một sản phẩm mà một khách hàng quan tâm. Trên cơ sở phân giải đồng tham chiếu trên văn bản có cảm xúc, tác giả luận án đề xuất (1) ***bài toán phân giải đồng tham chiếu đối tượng để tìm ra bộ ba gồm đối tượng, khía cạnh và cảm xúc.***

Tuy nhiên với văn bản có từ hai đối tượng trở lên thì việc tìm ra các khía cạnh, cảm xúc thuộc về đối tượng nào sẽ khó hơn rất nhiều. Xét ví dụ sau:

Ví dụ 1.1: “I gave my 6s plus to my wife and bought for myself an Apple iPhone7 two days ago. iPhone7 is beautiful. The battery is amazing. But what I really appreciate is the speaker producing good sound and its 128g storage.”

Trong ví dụ 1.1 xuất hiện hai đối tượng “6s plus” và “Apple iPhone7” (nói ngắn gọn là “iPhone7”). Ngoài ra có một số các khía cạnh như “battery”, “speaker”, “sound”, “storage” và các từ cảm xúc “beautiful”, “amazing”, “appreciate”, “good”. Vậy các từ chỉ khía cạnh và cảm xúc này sẽ đề cập đến đối tượng nào trong hai đối tượng trên. Với người đọc có thể dễ nhận thấy “battery” - “amazing”, “speaker” - “appreciate”, “sound” - “good” và “beautiful - khía cạnh ẩn” cùng nói về “Apple iPhone7”. Văn bản có từ hai đối tượng trở lên, việc xác định khía cạnh, cảm xúc cho đối tượng cụ thể trở nên phức tạp hơn.

Dựa trên đặc điểm của loại văn bản, yêu cầu của bài toán đặt ra và các tiếp cận hiện nay, luận án đề xuất (2) ***hướng giải quyết đồng tham chiếu đối tượng cho cảm xúc từ văn bản đơn giản chỉ có một đối tượng đến văn bản có từ hai đối tượng trở lên.***

Thực tế, các văn bản cảm xúc đề cập từ hai đối tượng trở lên khá nhiều vì khách hàng thường có sự so sánh các sản phẩm với nhau. Với các văn bản có đề cập nhiều hơn một đối tượng, để xác định khía cạnh và cảm xúc của đối tượng nào là vấn đề trọng tâm của bài toán phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc. Hiện nay, vấn đề này còn rất ít công trình khoa học được công bố.

### **1.3 Mô hình đề xuất**

Mô hình tổng quan của hệ thống phân giải đồng tham chiếu đối tượng mà luận án đề xuất được thể hiện qua Hình 1.1, với các mô đun sau.

Document: Một văn bản cảm xúc, bình luận về một sản phẩm theo miền.

PARSE: Mô đun phân tích cú pháp văn bản như gán nhãn, xác định mối quan hệ văn phạm phụ thuộc, v.v.

RECOAS: Nhận dạng các thành phần trong văn bản theo cấu trúc của ontology cảm xúc, như đối tượng, khía cạnh, cảm xúc và các mối quan hệ giữa các thành phần.

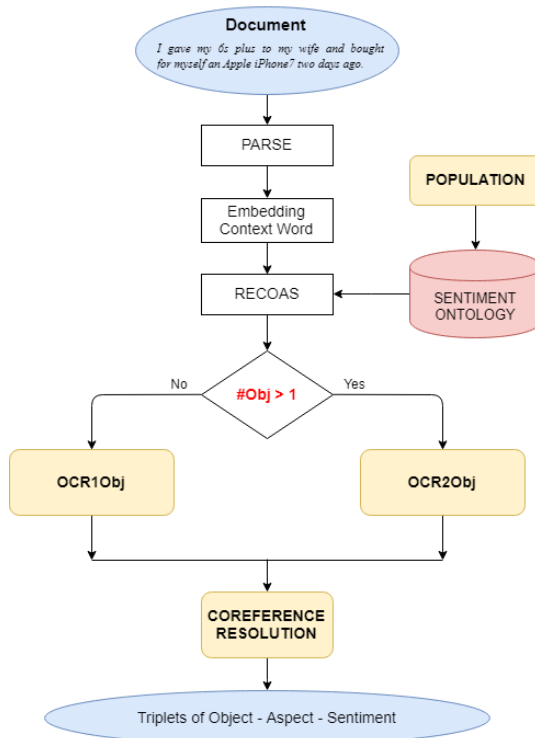
OCR1Obj: Mô đun phân giải đồng tham chiếu đối tượng trong văn bản chỉ đề cập đến một đối tượng.

OCR2Obj: Mô đun phân giải đồng tham chiếu đối tượng trong văn bản đề cập, có từ hai đối tượng trở lên.

EMBEDDING CONTEXT WORD (ECW): Mô đun chuyển đổi các từ thành các véc tơ dựa theo ngữ cảnh trái và phải của các từ của một câu trong một đoạn văn bản.

COREFERENCE RESOLUTION: Sử dụng đồ thị đồng tham chiếu truy xuất các bộ ba: đối tượng – khía cạnh – cảm xúc về các đối tượng được đề cập trong văn bản.

SENTIMENT ONTOLOGY: Cơ sở tri thức chứa các thể hiện của đối tượng, khía cạnh, cảm xúc về một loại sản phẩm và mối quan hệ giữa các thành phần.



Hình 1.1 Mô hình phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc

POPULATION: Mô đun làm giàu cơ sở tri thức ontology cảm xúc SENTIMENT ONTOLOGY.

TRIPLETS OF OBJECT – ASPECT – SENTIMENT: Kết quả của mô hình là các bộ ba: đối tượng – khía cạnh – cảm xúc về các đối tượng trong Document.

Để xây dựng mô hình phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc, luận án cần giải quyết một số bài toán sau.

**Bài toán 1:** Phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc có một đối tượng - OCR1Obj.

**Bài toán 2:** Làm giàu ontology cảm xúc hỗ trợ phân giải đồng tham chiếu đối tượng.

**Bài toán 3:** Xác định khía cạnh ẩn hỗ trợ phân giải đồng tham chiếu đối tượng.

**Bài toán 4:** Phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc có nhiều đối tượng – OCR2Obj.

### 1.3 Đóng góp chính của luận án

- **Đóng góp thứ nhất:** Xây dựng ontology cảm xúc SO và đồ thị đồng tham chiếu CRG hỗ trợ phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc. Đóng góp thứ nhất được công bố trong các công trình [2,4]
- **Đóng góp thứ hai:** Mô hình phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc cho văn bản có một đối tượng. Đóng góp thứ hai được công bố trong các công trình [2,4]
- **Đóng góp thứ ba:** Đề xuất mô hình làm giàu ontology cảm xúc. Đóng góp thứ ba được công bố trong công trình [3]
- **Đóng góp thứ tư:** Đề xuất mô hình xác định khía cạnh ẩn trong văn bản cảm xúc.
- **Đóng góp thứ năm:** Xây dựng mô hình phân giải đồng tham chiếu đối tượng cho văn bản cảm xúc có từ hai đối tượng trở lên. Đóng góp thứ năm được công bố trong các công trình [1].

### 1.4 Cấu trúc của luận án

Luận án được chia thành sáu chương và một phụ lục. Chương 1 trình bày mục tiêu, phạm vi nghiên cứu của luận án và những đóng góp chính của luận án; giới thiệu cấu trúc của luận án. Chương 2 trình bày các nghiên cứu liên quan như phân giải đồng tham chiếu, phân tích cảm xúc, phân giải đồng tham chiếu cho phân tích cảm xúc, xác định khía cạnh ẩn, xây dựng và làm giàu ontology trên miền chuyên biệt. Chương 3 trình bày mô hình, giải thuật phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc của văn bản có một đối tượng. Chương 4 trình bày mô hình làm giàu ontology cảm xúc hỗ trợ phân giải đồng tham chiếu. Chương 5 trình bày mô hình, giải thuật phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc cho văn bản có nhiều đối tượng. Chương 6 là

phần tổng kết trình bày những kết quả mà luận án đã đạt được cũng như những dự định nghiên cứu trong tương lai.

## **CHƯƠNG 2 CÁC NGHIÊN CỨU LIÊN QUAN**

### **2.1 Phân giải đồng tham chiếu đối tượng và khía cạnh cho phân tích cảm xúc**

Công trình Ding X và Liu B xác định các từ, cụm từ chỉ đối tượng (object) và thuộc tính (attribute) cùng tham chiếu đến một thực thể. Trong đó đối tượng là các thực thể có tên, như là tên các sản phẩm và dịch vụ. Các tác giả sử dụng tiếp cận học máy có giám sát với mô hình đề xuất huấn luyện trên 12 đặc trưng. Sau 10 năm, công trình trên đã được cải tiến bởi Chen J và các cộng sự với tiếp cận định hướng tri thức (knowledge-driven), cải tiến đề xuất khai phá tự động tri thức miền chuyên biệt từ dữ liệu văn bản thô và áp dụng tri thức theo miền chuyên biệt cho phân lớp đồng tham chiếu đối tượng, khía cạnh. Công trình Zhao Y đã xây dựng bộ phân lớp khía cạnh - đối tượng để đánh giá được mỗi cặp cần phân lớp và sử dụng Integer linear programming (ILP) để suy luận ra kết quả cuối cùng của bài toán.

### **2.2 Phân tích cảm xúc mức khía cạnh**

Phân tích cảm xúc mức khía cạnh (ABSA) theo Liu B là xác định các ý kiến về thực thể ở từng đặc tính của nó. Mỗi thực thể hay đối tượng trong một văn bản cảm xúc thường được đề cập đến các khía cạnh đi kèm với cảm xúc, quan điểm. Bài toán cần xác định từ cảm xúc nào tham chiếu đến khía cạnh gì trong văn bản. Bài toán phân tích ở mức này tập trung vào việc xác định các cặp khía cạnh - cảm xúc và đánh giá cảm xúc của khía cạnh.

### **2.3 Xây dựng ontology cảm xúc và làm giàu ontology**

Theo định nghĩa của Tung Thanh Nguyen, ontology cảm xúc là một cặp các tập khái niệm và tập các quan hệ. Ontology tổng quát bao gồm T-Box và A-Box. T-Box bao gồm các khái niệm khía cạnh, khái niệm cảm xúc và các mối quan hệ

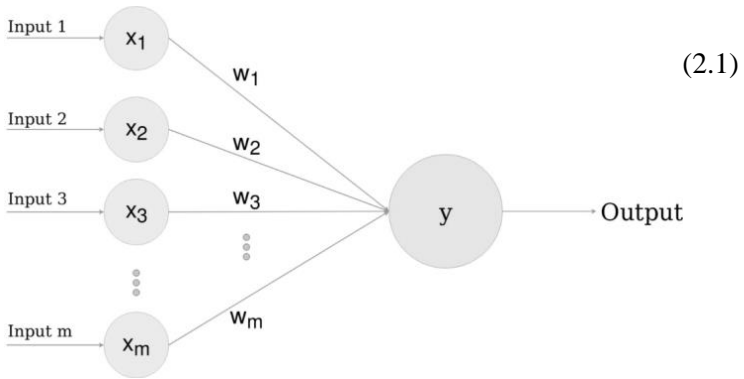


giữa các khái niệm; A-Box mô tả các trường hợp (instance) của các khái niệm. ontology cảm xúc được chia làm hai lớp khái niệm là khía cạnh và cảm xúc. Ontology cảm xúc được chia làm hai lớp khái niệm là khía cạnh và cảm xúc.

## 2.4 Các kiến thức nền tảng

### 2.4.1 Mạng nơ-ron nhân tạo – ANN

Với mạng ANN như Hình 2.1 còn gọi là mạng Perceptron một lớp, có nhiều đầu vào và một đầu ra. Đầu ra  $y$  được tính theo công thức (2.1), với  $x$  là đầu vào,  $w$  là trọng số.



Hình 2.1 Một nơ-ron nhân tạo - Perceptron

### 2.4.2 Mô hình ngôn ngữ - Tiền huấn luyện

BERT (Bidirectional Encoder Representations from Transformers) là một mô hình ngôn ngữ mới, được thiết kế để tiền huấn luyện việc biểu diễn từ, câu theo ngữ cảnh hai chiều trái và phải trên tất cả các tầng của mô hình. BERT là một kiến trúc đa tầng gồm nhiều lớp, là phương pháp học không giám sát và thực hiện hai nhiệm vụ chính: mô hình ngôn ngữ mặt nạ (Masked Language Model-MLM) và dự đoán câu tiếp theo (Next Sentence Prediction-NSP). Ngoài ra mô hình BERT còn sử dụng kỹ thuật Self-Attention, là phương pháp quan tâm đến các vị trí khác nhau trong một câu để đưa ra cách biểu diễn một chuỗi câu. Độ chính xác của mô hình đạt được 97% đến 98%.

## CHƯƠNG 3 PHÂN GIẢI ĐỒNG THAM CHIỀU CHO PHÂN TÍCH CẢM XÚC CÓ MỘT ĐỐI TƯỢNG

### 3.1.3 Tiền xử lý dữ liệu

Trong mô hình phân giải đồng tham chiều đối tượng, các văn bản cảm xúc đề cập đến các đối tượng cụ thể có tên riêng, ví dụ: “Samsung Galaxy A8”, “Apple iPhone7”. Để quá trình phân tách từ, cụm từ không làm mất ý nghĩa của tên các đối tượng, các cụm từ này sẽ được thay thế bởi một từ đại diện là OBJ. Hơn nữa việc thay thế này với mục đích là làm giảm số từ vựng của bộ phân tích cú pháp trong mô-đun tiền huấn luyện (ECW) sẽ được áp dụng trong các bài toán OCR1Obj và OCR2Obj.

## 3.1 Mô hình xác định khía cạnh ẩn - IAI

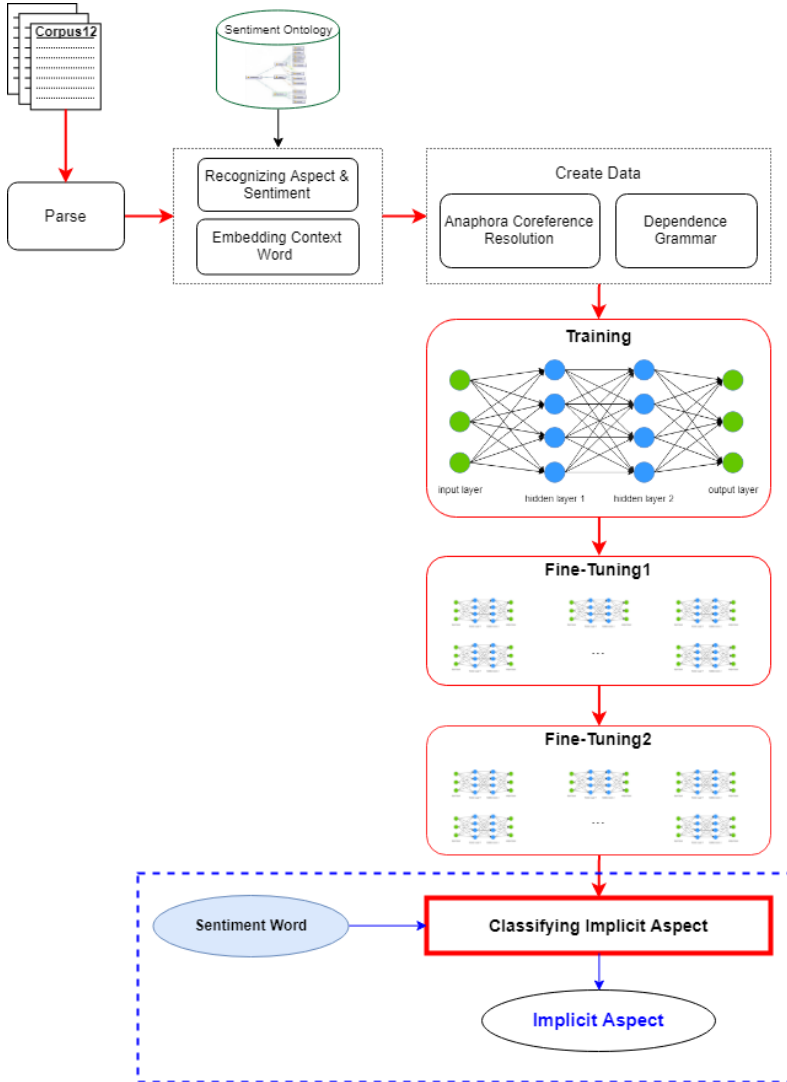
### 3.1.1 Mô hình IAI

Mô hình đề xuất xác định khía cạnh ẩn (Implicit Aspect Identification - IAI) được trình bày ở Hình 3.1. Trước tiên là Parse, mô-đun phân tích cú pháp văn bản, như gán nhãn từ loại (POS), phân tích cấu trúc cú pháp câu và các mối quan hệ trong câu. Tiếp theo là Recognizing Aspect & Sentiment (RAS), mô-đun gán nhãn ngữ nghĩa, nhận biết các từ chỉ khía cạnh và cảm xúc theo miền chuyên biệt thông qua cơ sở tri thức là ontology cảm xúc SO.

Embedding Context Word (ECW) là mô-đun chuyển đổi các từ thành các véc tơ dựa theo ngữ cảnh trái và phải của các từ của một câu trong một đoạn văn bản. Anaphora Coreference Resolution (ACR) là mô-đun phân giải đồng tham chiều đại từ, hỗ trợ xác định cặp cảm xúc - khía cạnh. Dependence Grammar (DG) là mô-đun hỗ trợ xác định cặp cảm xúc - khía cạnh cho một câu của văn bản bằng văn phạm phụ thuộc. ACR và DG cùng mục tiêu xác định các cặp cảm xúc - khía cạnh (đã được gán nhãn ngữ nghĩa bởi RAS và chuyển đổi thành các véc tơ số thực bởi ECW). Từ bốn mô-đun hỗ trợ tạo ra bộ dữ liệu huấn luyện cảm xúc - khía cạnh, mỗi từ cảm xúc đã được chuyển thành một véc tơ số thực có ngữ cảnh được gán nhãn khía cạnh tương ứng trong văn bản. Training là mô-đun huấn luyện xác định bộ trọng số  $W$ , áp dụng mạng nơ-ron nhân tạo nhiều

tăng nhiều đầu ra tương ứng với số lượng các khía cạnh trong ontology cảm xúc SO và các véc tơ cảm xúc làm đầu vào.

Sau khi mô đun Training học trên tập dữ liệu huấn luyện xác định được  $W$  và



Hình 3.1 Kiến trúc mô hình xác định khía cạnh ẩn IAI

để nâng cao độ chính xác cho việc xác định khía cạnh ẩn, mô đun Fine-Tuning1 sử dụng W huấn luyện tiếp trên từng bộ dữ liệu tương ứng với mỗi từ cảm xúc.

Kết quả của Fine-Tuning1 là các mô hình, được ký hiệu W1s, dùng để phân lớp cho mô đun tiếp theo, Fine-Tuning2. Tương tự như Fine-Tuning1, mô đun Fine-Tuning2 sẽ tiếp tục huấn luyện từ W1s trên các tập dữ liệu cảm xúc tham chiếu đến khía cạnh mang tính chất thuộc tính (không phải là thiết bị, thành phần của đối tượng) để thu hẹp phạm vi khía cạnh ẩn cần xác định và cho kết quả W2s, nâng cao độ chính xác cho mô đun cuối cùng, Classifying Implicit Aspect.

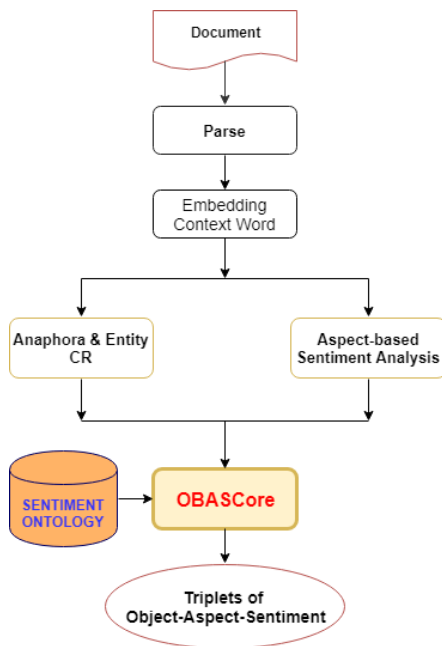
Classifying Implicit Aspect là mô đun cuối cùng, xác định khía cạnh ẩn cho các từ cảm xúc dựa vào các W2s của Fine-Tuning2. Với đầu vào là một từ cảm xúc được biểu diễn bởi một véc tơ cảm xúc, bộ phân lớp sử dụng mô hình tương ứng với cảm xúc để xác định khía cạnh ẩn.

### **3.1.2 Thực nghiệm và đánh giá**

Mô-đun Training có 69905 mẫu dữ liệu được trích xuất từ Corpus12 với 389103 văn bản cảm xúc về điện thoại thông minh (smartphone), trong đó: dữ liệu huấn luyện có 64374 mẫu và dữ liệu thử nghiệm có 5598 mẫu và độ chính xác 0.82. Mô hình IAI được đánh giá trên 110 văn bản cảm xúc về smartphone lần lượt qua từng mô đun: Training (IAI1), Fine-Tuning1 (IAI2) và Fine-Tuning2 (IAI3). Kết quả độ chính xác đạt được của IAI3 (0.88) lớn hơn IAI2 và IAI1. Ngoài ra IAI được so sánh với công trình của Rana TA (0.71) và cũng cho thấy kết quả tốt hơn rất nhiều.

## **3.2 Mô hình phân giải đồng tham chiếu cho phân tích cảm xúc có một đối tượng**

Mô hình phân giải đồng tham chiếu đối tượng cho các văn bản cảm xúc có một đối tượng là xác định đồng tham chiếu bộ ba: đối tượng - khía cạnh - cảm xúc được luận án đề xuất và trình bày ở Hình 3.2.



Hình 3.2 Mô hình phân giải đồng tham chiếu cho phân tích cảm xúc có một đối tượng.

### 3.2.1 *Parse – Phân tích cú pháp*

Thực hiện gán nhãn từ loại, nhãn quan hệ ngữ pháp, tạo cây cú pháp v.v.

### 3.2.2 *Anaphora & Entity CR - Phân giải đồng tham chiếu đại từ, thực thể*

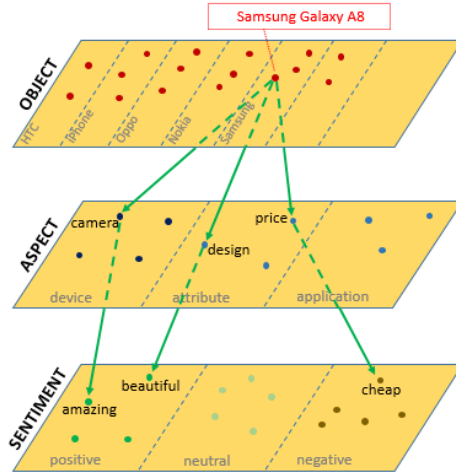
Từ đặc điểm này của văn bản, luận án chia thành hai trường hợp riêng biệt xác định đồng tham chiếu trong văn bản cho mô-đun Anaphora & Entity CR, đó là: (1) xác định các đại từ đề cập đến đối tượng, thực thể ban đầu được gọi là phân giải đồng tham chiếu đại từ; (2) xác định các cụm danh từ cùng đề cập đến đối tượng, thực thể ban đầu gọi là phân giải đồng tham chiếu thực thể có tên.

### 3.2.3 *Aspect-based sentiment analysis - Phân tích cảm xúc mức khía cạnh*

Xác định mối quan hệ giữa đối tượng - cảm xúc, khía cạnh - cảm xúc trong một văn bản.

### 3.2.4 *Ontology cảm xúc*

Kiến trúc ontology cảm xúc được minh họa trong hình 3.3.



Hình 3.3 Kiến trúc Ontology cảm xúc.

Ontology cảm xúc gồm ba lớp: lớp OBJECT bao gồm tên các đối tượng theo miền, được ký hiệu **Obj**; Lớp thứ hai ASPECT, bao gồm các khía cạnh của đối tượng, được ký hiệu **Asp**; Lớp thứ ba SENTIMENT, là các ý kiến, cảm xúc hay nhận xét về các khía cạnh của đối tượng và được ký hiệu **Sen**. Ngoài ba lớp khái niệm, trong ontology cảm xúc còn có mối quan hệ giữa các lớp với các thành phần (instance). Ontology cảm xúc với kiến trúc ở Hình 3.3 có thể: nhận diện các từ thuộc lớp đối tượng, khía cạnh hoặc cảm xúc; xác định khía cạnh (aspect) ảnh hưởng cho cảm xúc; xác định khía cạnh (aspect) với đối tượng (object); ontology có thể giải quyết đồng tham chiếu thực thể cụm danh từ; xác định cảm xúc là tích cực hay tiêu cực.

### 3.2.5 *Đồ thị đồng tham chiếu (CoReference Graph – CRG)*

Đồ thị đồng tham chiếu (CRG), là công cụ cuối cùng để sinh ra các bộ ba đối tượng, khía cạnh và cảm xúc. CRG là đồ thị có hướng và có trọng số với mỗi đỉnh là các từ, cụm từ thuộc bộ đồng tham chiếu đại từ, thực thể có tên hoặc bộ phân tích cảm xúc mức khía cạnh. Trọng số của mỗi cạnh là khoảng cách giữa

các lớp mà mỗi từ, cụm từ thuộc về lớp đó. Luận án áp dụng các thuật toán tìm kiếm, duyệt cây để xác định các bộ ba đối tượng - khía cạnh - cảm xúc trên một miền chuyên biệt.

### 3.2.6 Mô đun OBASCore

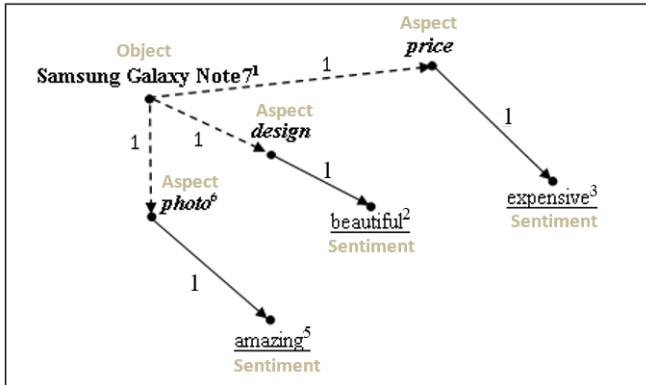
Hình 3.4 trình bày chi tiết giải thuật của OBASCore. Kết quả thực hiện các bước của giải thuật hình 3.4 được minh họa ở hình 3.5. Đồ thị CRG xác định đối tượng Samsung Galaxy Note7 có ba cặp khía cạnh - cảm xúc là *photo - amazing*, *design - beautiful*, *price - expensive*.

### 3.2.7 Thực nghiệm và đánh giá

Kết quả thực nghiệm mô hình phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc có một đối tượng trên 320 văn bản cảm xúc về smartphone đạt độ chính xác 76% (sử dụng SO xác định khía cạnh ẩn) và 78% (có kết hợp mô hình IAD). Với sự hỗ trợ của SO và CRG giải quyết bài toán OCR1Obj đạt kết quả khá tốt. Tuy nhiên bài toán còn phụ thuộc vào dữ liệu của SO, chính vì vậy việc làm giàu cho SO là cần thiết.

<p><u>Input:</u> Tập đồng tham chiếu C, tập cảm xúc S, Ontology cảm xúc O</p> <p><u>Output:</u> Tập các đối tượng - khía cạnh - cảm xúc ObjAspSen</p> <p><u>Algorithm:</u></p> <ol style="list-style-type: none"><li>1. Khởi tạo CRG</li><li>2. Add(<math>V^C, E^C, CRG</math>)</li><li>3. AddNoDuplicate(<math>V^S, E^S, CRG</math>)</li><li>4. ClassifyVertices(<math>CRG, O</math>)</li><li>5. UpdateWeight(<math>CRG, O</math>)</li><li>6. AddEdges(<math>V^C, O, 0, CRG</math>)</li><li>7. Xét các đỉnh đầu v không là đỉnh cuối:     sum = SumOfWeights(v, outerVertices)     <b>If</b> (sum=2) addVertexAspect(outerVertices, O, CRG)     <b>Else If</b> (sum=1)         removeVertex(intermediateVertex, CRG)     <b>Else</b> removeVertex(outerVertices, CRG)</li><li>8. AddEdges(V.Object, V.Aspects, 1, CRG)</li><li>9. <b>For</b> each edge (v1, v2) of E: if getWeight(v1, v2)=0</li></ol>
---

Hình 3.4 Giải thuật của mô đun OBASCore



Hình 3.5 CRG của ví dụ 1.1 được thực hiện bởi giải thuật hình 3.4

### 3.3 Kết luận

Ứng dụng ontology cảm xúc SO, đồ thị đồng tham chiếu CRG do luận án đề xuất cho bài toán phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc có một đối đã xác định được đối tượng, khía cạnh kèm theo cảm xúc. Trong văn bản. Ngoài ra, với sự hỗ trợ của mô hình xác định khía cạnh ẩn IAI trong bài toán OCR1Obj cũng thu được kết quả tốt hơn so với sử dụng SO đơn thuần.

## CHƯƠNG 4 LÀM GIÀU ONTOLOGY CẢM XÚC HỖ TRỢ PHÂN GIẢI ĐỒNG THAM CHIẾU CHO PHÂN TÍCH CẢM XÚC

### 4.1 Giới thiệu

Làm giàu dữ liệu ontology (Ontology Population) là nhiệm vụ cập nhật thêm dữ liệu hay gọi là làm giàu tri thức cho các lớp khái niệm trong ontology. Quá trình làm giàu dữ liệu cho ontology không làm thay đổi cấu trúc của ontology, ví dụ như hệ thống phân cấp khái niệm và các mối quan hệ không phân loại (non-taxonomic) được giữ nguyên. Sự thay đổi đó là tập các hiện thực hóa (instances) của khái niệm và mối quan hệ trong phạm vi miền.

### 4.2 Giới thiệu một số mô hình làm giàu dữ liệu ontology cảm xúc

Hiện nay có rất nhiều công trình nghiên cứu làm giàu dữ liệu cho ontology với các phương pháp tiếp cận khác nhau. Phương pháp học máy có giám sát và



không giám sát, sử dụng mô hình phân lớp để xác định các thể hiện (D. Celjuska và các cộng sự, M. Jiang và các cộng sự, v.v). Phương pháp học sâu sử dụng các thuật toán CNN, RNN để xác định các nguồn dữ liệu phụ thuộc miền (D. Zeng và các cộng sự, A. Ayadi và các cộng sự, C. Liu và các cộng sự). Ngoài ra, có các phương pháp dựa theo luật, thống kê và kết hợp các phương pháp để áp dụng trên các miền chuyên biệt khác nhau.

### **4.3 Mô hình làm giàu ontology cảm xúc ESO (Enriched Sentiment Ontology)**

Mô hình làm giàu SO, được thể hiện ở hình 4.1, với tiếp cận là sự kết hợp phương pháp dựa theo luật, độ tương tự ngữ nghĩa giữa các từ, nhận diện tính phân cực của cảm xúc dựa trên các bộ từ điển khá phổ dụng như WordNet và Opinion Lexicon trong lĩnh vực xử lý ngôn ngữ tự nhiên.

*Document* là đầu vào của mô hình ESO có dạng văn bản không cấu trúc đề cập đến một đối tượng và có ý kiến, cảm xúc về các khía cạnh của đối tượng đó.

*Parse* thực hiện phân tích cú pháp. Luận án sử dụng bộ công cụ parse của Stanford để thực hiện việc gán nhãn từ loại và phân tích cú pháp theo văn phạm phụ thuộc (DG). Các nhãn từ loại sau phân tích cú pháp là danh từ riêng (NNP), danh từ chung (NN), động từ (VERB), tính từ (ADJ) và trạng từ (ADV). Đây là năm loại nhãn từ được luận án lọc ra để phân lớp ngữ nghĩa.

*Annotating of semantics* được dùng để gán nhãn ngữ nghĩa theo SO cho các từ với nhãn từ loại NNP, NN, VERB, ADJ và ADV. Những từ còn lại không thuộc vào các lớp của SO (không tồn tại) sẽ được gán nhãn UNKN.

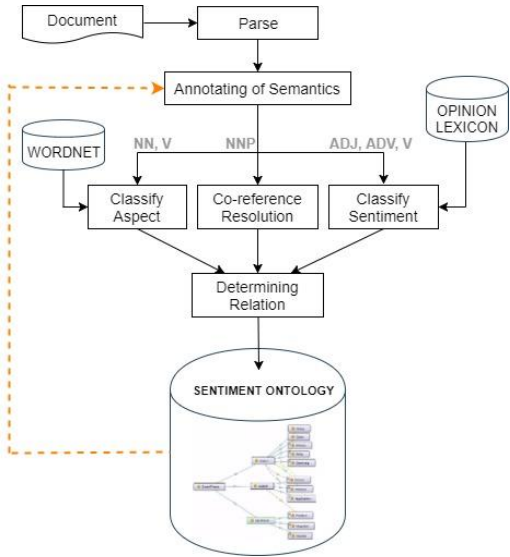
*Coreference resolution* sử dụng DG như “compound” xác định các danh từ riêng đồng tham chiếu với nhau. “Compound” là quan hệ văn phạm giữa các danh từ hoặc danh từ riêng. Ngoài ra để phân lớp cho các Obj, tác giả sử dụng phương pháp so trùng giữa các Object với các lớp con.

*WordNet* (G.A. Miller) là một cơ sở tri thức phân cấp trong tiếng Anh được sử dụng như một cơ sở dữ liệu từ vựng chứa 155.287 từ và 117.659 tập đồng nghĩa

cho các từ loại danh từ, động từ, tính từ và danh từ. Tất cả các tập đồng nghĩa có quan hệ ngữ nghĩa.

Classify Aspect là mô-đun phân lớp khía cạnh, sử dụng độ đo tương tự ngữ nghĩa giữa các từ (Semantic Similarity between Words - SSW) và WordNet. Công thức tính SSW được áp dụng theo công thức (4.1).

$$s(w_1, w_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \tag{4.1}$$



Hình 4.1 Mô hình đề xuất ESO làm giàu ontology cảm xúc

Giá trị của  $\alpha$  và  $\beta$  phụ thuộc vào cơ sở tri thức được sử dụng. Trong công thức (4.1),  $l$  là độ dài đường đi ngắn nhất giữa hai từ  $w_1$  và  $w_2$ ,  $h$  là độ sâu của lớp đầu tiên trong mạng ngữ nghĩa phân cấp là từ gần nghĩa chung nhất của  $w_1$  và  $w_2$  (subsumer).

Classify Sentiment, phân lớp cảm xúc sử dụng bộ từ vựng cảm xúc Opinion Lexical (H. Minqing và L. Bing) để phân lớp các từ loại tính từ, trạng từ và động từ vào ba lớp positive, negative và neutral.

Determining Relation, xác định mối quan hệ giữa các thành phần Obj-Asp và Aso-Sen. Luận án sử dụng các quan hệ văn phạm phụ thuộc có tác động lên các từ loại NN, NNP, VERB, ADJ và ADV để áp dụng trên kết quả phân lớp cảm xúc, phân lớp khía cạnh và phân giải đồng tham chiếu.

#### **4.4 Thực nghiệm và đánh giá**

Mô hình ESO được thực nghiệm trên tập dữ liệu thu thập từ YouNetMedia.com. Với 614 câu có ý kiến nhận xét về các smartphone, luận án thu được các từ, cụm từ cho lớp Obj, Asp và Sen tương ứng là 106, 250 và 167 từ. Đối với lớp Obj, độ chính xác đạt 89%, lớp Asp đạt 72% và lớp Sen đạt 83%. Kết quả này phản ánh đúng thực tế vì khía cạnh thường là danh từ hoặc động từ xuất hiện với tần suất không nhỏ. ESO còn phụ thuộc vào WordNet và Opinion Lexical, đó là hạn chế của mô hình.

#### **4.5 Kết luận**

Với đặc thù riêng về cấu trúc của SO hỗ trợ bài toán phân giải đồng tham chiếu cho phân tích cảm xúc nên mỗi lớp khái niệm sẽ có những tiếp cận khác nhau để làm giàu các thể hiện cho các lớp. Cụ thể, lớp đối tượng sử dụng văn phạm phụ thuộc, lớp khía cạnh sử dụng cơ sở tri thức WordNet và lớp cảm xúc dựa trên bộ từ vựng cảm xúc.

### **CHƯƠNG 5 PHÂN GIẢI ĐỒNG THAM CHIẾU ĐỐI TƯỢNG CHO PHÂN TÍCH CẢM XÚC CÓ NHIỀU ĐỐI TƯỢNG**

#### **5.1 Đặt vấn đề**

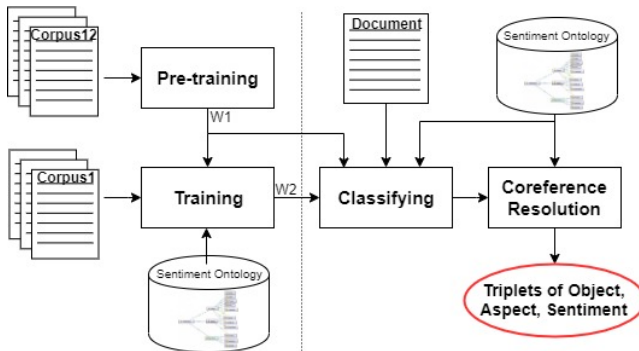
Trong văn bản có nhiều quan điểm, cảm xúc và đề cập đến nhiều khía cạnh của các đối tượng khác nhau sẽ làm cho bài toán phân giải đồng tham chiếu cho cảm xúc càng trở nên phức tạp. Với bài toán phân giải đồng tham chiếu đối tượng xác định được bộ ba đối tượng - khía cạnh - cảm xúc trong văn bản có nhiều đối tượng, luận án đã kết hợp cơ sở tri thức với học máy, học sâu, văn phạm phụ thuộc và đồ thị đồng tham chiếu để giải quyết.

## 5.2 Mô hình phân giải đồng tham chiếu cho phân tích cảm xúc có nhiều đối tượng – CROAS

Mô hình CROAS gồm hai giai đoạn huấn luyện và thử nghiệm, được trình bày như Hình 5.1.

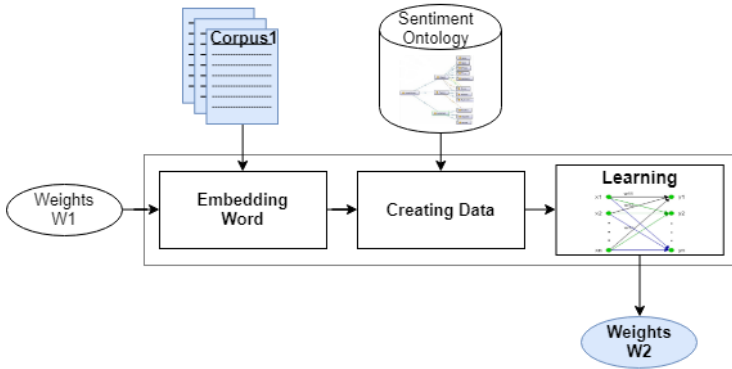
### 5.2.1 Giai đoạn huấn luyện

Giai đoạn đầu tiên này sử dụng kho ngữ liệu các văn bản Corpus12 để tiền huấn luyện (Pre-training) và Corpus1, SO để huấn luyện (Training). Kho ngữ liệu Corpus12 bao gồm các văn bản chỉ đề cập đến một đối tượng (Corpus1) và các văn bản đề cập từ hai đối tượng trở lên (Corpus2). Từ kho ngữ liệu ban đầu Corpus12, giai đoạn này sử dụng mô hình BERT sinh ra bộ trọng số  $W1$  biểu diễn véc-tơ có ngữ cảnh. Mô đun huấn luyện học trên Corpus1 để xác định bộ



Hình 5.1 Kiến trúc của mô hình CROAS

trọng số  $W2$  được dùng để phân lớp đối tượng cho các từ khóa cạnh và cảm xúc mà chúng tham chiếu đến. Hình 5.2 trình bày cấu trúc mô đun huấn luyện, bao gồm ba mô đun con: nhúng từ (Embedding Word –EW), tạo dữ liệu huấn luyện (Creating Data) và huấn luyện (Learning). Trong đó mô-đun Creating Data lọc các cặp dữ liệu *Asp-Obj* và *Sen-Obj* dựa trên SO và văn phạm phụ thuộc. Tiếp theo mô-đun Learning được huấn luyện theo giải thuật 5.1.



Hình 5.2 Mô đun Training của mô hình CROAS

Các biến sau được sử dụng trong giải thuật 5.1.

- $E$  là hàm mất mát giữa  $y$  và  $\hat{y}$
- $dW2$  là giá trị cần cập nhật cho  $W2$
- $db$  là giá trị cần cập nhật cho độ lệch  $b$
- $W2$  được cập nhật mới thông qua hệ số học  $\eta$  và  $dW2$ .

### ***Giải thuật 5.1*** Huấn luyện trong mô-đun Training

**Input:** Example( $x, y$ )

**Output:** Weight  $W2$

- 1: InitialRandom( $W2$ )
- 2: **for**  $i = 0$  **to** *steps* **do**
- 3:      $\hat{y} = \text{FeedForward}(W2, x)$
- 4:      $E = \text{CalcLoss}(\hat{y}, y)$
- 5:      $dW2 = \text{Calc}_dW2(E, x)$
- 6:      $db = \text{CalcBias}(E)$
- 7:      $W2 = \text{UpdateWeight}(\eta, dW2)$
- 8:      $b = \text{UpdateBias}(\eta, db)$
- 9: **end for**

Quá trình xác định  $W2$  hoàn thành sẽ kết thúc giai đoạn huấn luyện.

### **5.2.2** *Giai đoạn thử nghiệm*

Trong mô-đun Classifying (giải thuật 5.2), BERT được sử dụng để chuyển đổi các từ trong Document sang các véc-tơ số thực có ngữ cảnh, ký hiệu  $V_s$ . Khi

đó, SO lọc ra các véc-tơ của các từ chỉ đối tượng, khía cạnh và cảm xúc theo phạm vi miền dữ liệu (chuyên biệt).

### ***Giải thuật 5.2*** Mô-đun Classifying

---

**Input:** Document  $D$ ,  $SO$ ,  $W1$ ,  $W2$

**Output:** Pairs of *Object-Aspect*, *Object-Sentiment*  $P$

1:  $V_s = \text{EmbeddingWord}(D, W1)$

2:  $x, y = \text{FilteringVector}(V_s, SO)$

3:  $\hat{y} = \text{FeedForward}(x, W2)$

4: **for**  $i = 0$  **to**  $\text{size}(\hat{y})$  **do**

5:      $j = \text{MinDistance}(\hat{y}_i, y)$

6:      $P.\text{addPair}(x_i, y_j)$

7: **end for**

---

Giải thuật 5.2 mô tả các bước thực thi của Classifying với đầu vào  $x$  là các véc-tơ chỉ khía cạnh và cảm xúc, đầu ra  $y$  là các véc-tơ chỉ đối tượng trong Document. Để xác định các khía cạnh và cảm xúc thuộc về đối tượng nào, thuật toán sử dụng khoảng cách Euclidean để tính khoảng cách giữa giá trị dự đoán  $\hat{y}$  với véc-tơ  $y$ . Khi đó các khía cạnh và cảm xúc sẽ được phân lớp cho đối tượng có khoảng cách Euclidean nhỏ nhất. Đầu ra của giải thuật 5.2 là các cặp  $P$ : đối tượng- khía cạnh và đối tượng-cảm xúc. Các cặp  $P$  được đưa vào đồ thị CRG trong mô-đun Coreference Resolution (hình 5.5) và thu được các bộ ba đối tượng, khía cạnh và cảm xúc. Quá trình thực hiện của mô-đun được trình bày trong giải thuật 5.3 (mô-đun Coreference Resolution).

Kết quả từ giải thuật 5.2 sẽ xây dựng được đồ thị CRG ban đầu với các đỉnh từ  $P$  bằng phương thức AddVertices. Khi đó các cạnh lần lượt được tạo giữa các đỉnh nhờ mối quan hệ văn phạm phụ thuộc bằng phương thức AddEdges. Các quan hệ của văn phạm phụ thuộc giúp nâng cao độ chính xác của mô hình CROAS. Bước tiếp theo, trọng số của đồ thị được cập nhật dựa vào SO.

Trọng số của cạnh là 0 nếu nó được tạo bởi hai đỉnh là đồng tham chiếu thực thể (ví dụ: Apple iPhone7 và iPhone7). Trọng số là 1 nếu hai đỉnh thuộc cặp đối tượng, khía cạnh (iPhone7, battery) hoặc khía cạnh, cảm xúc (battery, amazing). Trọng số bằng 2 nếu hai đỉnh thuộc cặp đối tượng, cảm xúc

(iPhone7, beautiful).amazing). Trọng số bằng 2 nếu hai đỉnh thuộc cặp đối

### **Giải thuật 5.3** Mô-đun Coreference Resolution

**Input:** Pairs of Object-Aspect, Object-Sentiment  $P$ ;

Pairs of dependency grammar  $DG$ ; Sentiment ontology  $SO$

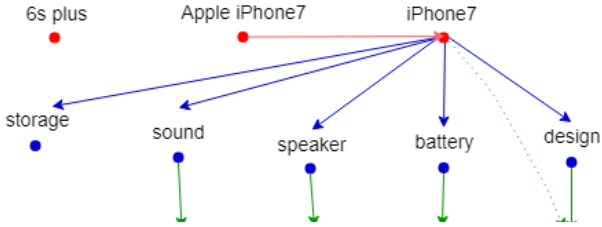
**Output:** Triplets of *Object, Aspect, and Sentiment*

```
1: Initializing CRG( $v1, v2$ )
2: AddVertices( $P, CRG$ )
3: AddEdges( $DG, CRG$ )
4:  $DG.v2 = \text{true}$ 
5: if  $P.v2 \neq \text{true}$  then
6:     AddEdges( $P, CRG$ )
7: end if
8: UpdateWeight( $CRG, SO$ )
9: if  $\text{weight}(v.Obj, vi) = 2$  then
10:     $asp = SO.Asp(vi)$ 
11:    addEdge( $v.Obj, asp$ )
12:    addEdge( $asp, vi$ )
13: end if
```

Để có được đầu ra cuối cùng là các bộ ba, các cạnh có trọng số bằng 2 sẽ được thay bởi hai cạnh có chung một đỉnh và đỉnh này là một khía cạnh ảnh hưởng với một đỉnh là cảm xúc trong các cạnh này. Đỉnh chung này được xác định bằng hai phương pháp hoặc sử dụng  $SO$  hoặc sử dụng IAI để tìm khía cạnh ảnh.

Xét lại ví dụ 1.1: “I gave my 6s plus to my wife and bought for myself an Apple iPhone7 two days ago. iPhone7 is beautiful. The battery is amazing. But what I really appreciate is the speaker producing good sound and its 128g storage.”

Kết quả thử nghiệm trên ví dụ 1.1 được kết quả cuối cùng thể hiện trên đồ thị CRG như Hình 5.3. Từ Hình 5.3, ta có bốn bộ ba: (1) iPhone7, sound, good; (2) iPhone7, speaker, appreciate; (3) iPhone7, battery, amazing; và (4) iPhone7, design, beautiful. Đây là kết quả cuối cùng của bài toán phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc có nhiều đối tượng.



Hình 5.3 CRG của ví dụ 1.1.

### 5.2.3 Thực nghiệm và đánh giá

**Tập dữ liệu và cài đặt tham số:** Miền dữ liệu của Corpus12, Corpus1 và Document là các văn bản cảm xúc về điện thoại thông minh được thu thập từ trang web Amazon (<https://www.amazon.com/>, <http://snap.stanford.edu/data/>) và <http://www.younetmedia.com>).

Corpus12 với 389103 văn bản, tương đương với 3.203.008 mẫu dữ liệu được huấn luyện trên mô hình BERT<sub>LARGE</sub> (24 tầng, số nút ẩn bằng 1024 và 16 self-attention heads) với 300,000 và 500,000 bước. Corpus1 với 74,3MB (9456 mẫu). Để đánh giá hiệu quả của mô hình, luận án sử dụng Document có 1000 văn bản cảm xúc.

**Phần cứng và phần mềm:** Mô hình CROAS được phát triển trên môi trường Colab.Google với GPU và 8-core TPU. Mô đun Pre-training và Training sử dụng các siêu tham số: batch size, số bước huấn luyện, hệ số học và số bước trên một epoch được trình bày ở Bảng 5.1.

Bảng 5.1 Các thông số cài đặt của Pre-training và Training

	Pre-training		Training
Số bước lặp	300,000	500,000	500,000
Kích thước Batch	32	32	16
Hệ số học	5e-5	5e-5	1.0
Số bước/epoch	100,094	100,094	591
Thời gian thực thi	3.9 ngày (với TPU)	6.5 ngày (với TPU)	5.2 ngày (với GPU)



**Kết quả đánh giá:** Bảng 5.2 trình bày kết quả độ chính xác và độ truy hồi của mô hình CROAS. Trong đó CROAS\* và CROAS\*\* tương ứng với CROAS có sử dụng SO và IAI để xác định khía cạnh ẩn.

Bảng 5.2 Kết quả đánh giá CROAS trên 1000 văn bản bình luận về smartphone

Số bước huấn luyện của Pre-training	Mô hình	Độ chính xác (%)	Độ truy hồi (%)
300 000	CROAS*	77.00	75.04
	CROAS**	81.74	79.66
500 000	CROAS*	88.00	85.76
	CROAS**	<b>90.64</b>	88.33

Bảng 5.2 cho thấy, khi tăng số bước huấn luyện lên 500000 thì CROAS\*\* có độ chính xác đạt 90.64% và độ truy hồi là 88.33% cũng cao hơn CROAS\* và so với 300000 bước huấn luyện. Bảng 5.3 so sánh kết quả thực nghiệm của CROAS với hai công trình Baseline dựa trên các cặp đồng tham chiếu đối tượng, khía cạnh. Kết quả OA1 (với 300000 bước huấn luyện) và OA2 (500000 bước) tốt hơn so với Baseline1 (X. Ding, B. Liu) và Baseline2 (Y. Zhao và các cộng sự).

Bảng 5.3 So sánh kết quả thực nghiệm của CROAS và hai công trình Baseline

Phương pháp	Độ chính xác (%)	Độ truy hồi (%)
Baseline1	69.80	74.23
Baseline2	77.96	76.16
OA1	79.00	77.02
OA2	89.10	86.87

### 5.3 Kết luận

Mô hình CROAS kết hợp máy học trên các véc-tơ có ngữ cảnh với SO, CRG và văn phạm phụ thuộc đạt hiệu quả tốt. Ngoài ra kết hợp IAI trong CROAS, kết quả cao hơn so với SO. Tuy nhiên, CROAS chưa hiệu quả trong các văn bản có chứa chuỗi so sánh (more ... than ...).

## CHƯƠNG 6 KẾT LUẬN

Kết quả nghiên cứu phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc đã giải quyết được bốn bài toán con của luận án: xác định khía cạnh ẩn, làm giàu ontology cảm xúc, phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc có một đối tượng và nhiều đối tượng.

Về mặt thực tiễn, luận án có thể ứng dụng thực tế vì gần với nhu cầu thực tiễn của người dùng. Về ý nghĩa khoa học, luận án đã khai thác được các giải thuật của xử lý ngôn ngữ tự nhiên; khai thác tiếp cận học máy, học sâu trong xử lý ngôn ngữ tự nhiên; khai thác được ngữ cảnh của từ ngữ trong câu, văn bản; và đã có đề xuất các phương pháp luận cho bốn bài toán cụ thể với những đóng góp chính sau.

### 6.1 Các đóng góp của luận án

- **Đóng góp thứ nhất:** Xây dựng ontology cảm xúc và đồ thị đồng tham chiếu CRG hỗ trợ phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc.
- **Đóng góp thứ hai:** Mô hình phân giải đồng tham chiếu đối tượng cho phân tích cảm xúc cho văn bản có một đối tượng.
- **Đóng góp thứ ba:** Đề xuất mô hình làm giàu ontology cảm xúc.
- **Đóng góp thứ tư:** Xây dựng mô hình xác định khía cạnh ẩn.
- **Đóng góp thứ năm:** Xây dựng mô hình phân giải đồng tham chiếu đối tượng cho văn bản cảm xúc có nhiều đối tượng.

### 6.2 Hướng phát triển

Mặc dù đạt được kết quả khá tốt và được đánh giá bằng thực nghiệm, luận án vẫn còn một số vấn đề nhỏ cần quan tâm và phát triển trong tương lai.

Vấn đề 1: Phát triển và làm giàu ontology cảm xúc bằng phương pháp tự động là nhu cầu cần thiết.

Vấn đề 2: Phát triển mô hình CROAS để khắc phục vấn đề phụ thuộc bước tiền huấn luyện.

Vấn đề 3: Phát triển mô hình của luận án trên văn bản cảm xúc tiếng Việt sẽ có ý nghĩa khoa học và thực tiễn cao trong lĩnh vực NLP ở Việt Nam.

## DANH MỤC CÔNG TRÌNH ĐÃ CÔNG BỐ

### Tạp chí quốc tế

1. T. Le Thi, T. Phan Thi, and T. Quan Thanh. Machine learning using context vectors for object coreference resolution. Computing (2021). <https://doi.org/10.1007/s00607-021-00902-4>.

### Tạp chí trong nước

1. T. Le Thi, T. Phan Thi, and T. Quan Thanh. (2017). Coreference resolution Ontology-based in sentiment analysis. Science and Technology Development Journal, 20(K9), 23-30. <https://doi.org/10.32508/stdj.v20iK9.1673>.

### Kỹ yếu hội nghị quốc tế

1. T. Le Thi, T. Phan Thi, and T. Quan Thanh, "Instance-Based Enrichment of Sentiment Ontology," 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), Danang, Vietnam, 2019, pp. 1-6.

2. T. Le Thi, T. Quan Thanh, and T. Phan Thi. 2017. Ontology-Based Entity Coreference Resolution For Sentiment Analysis. In Proceedings of the Eighth International Symposium on Information and Communication Technology (SoICT 2017). ACM, New York, NY, USA, 50-56. DOI: <https://doi.org/10.1145/3155133.3155168>

3. T. Le Thi, H. Vo Thanh, T. Mai Duc, T. Quan Thanh, T. Phan Thi. "An Ontology-based Coreference Resolution Approach for Aspect-level Sentiment Analysis". RIVF 2016, HaNoi, Vietnam.

4. T. Le Thi, H. Vo Thanh, T. Mai Duc, T. Quan Thanh, T. Phan Thi. (2016) Sentiment Analysis Using Anaphoric Coreference Resolution and Ontology Inference. In: Sombattheera C., Stolzenburg F., Lin F., Nayak A. (eds) Multi-disciplinary Trends in Artificial Intelligence. MIWAI 2016. Lecture Notes in Computer Science, vol 10053. Springer, Cham.

### Đề tài nghiên cứu khoa học

1. PGS. TS Quản Thành Thơ, Phân giải đồng tham chiếu cho tiếng Việt trong quá trình phân tích cảm xúc hướng đến khía cạnh. ĐHQG-HCM. 2016.

2. GS. TS Phan Thị Tươi, Lê Thị Thủy, Phân giải đồng tham chiếu đối tượng, khía cạnh trên cơ sở cảm xúc trong các bài nhận xét tiếng Anh. ĐHBK-HCM. 2018.
3. PGS. TS Quản Thành Thơ, Kết hợp rút trích nét đối tượng và học máy để phân tích cảm xúc trên khía cạnh cho văn bản tiếng Việt. ĐHQG-HCM.2018.