การสกัดคุณลักษณะสินค้าและความคิดเห็นในบทวิจารณ์สินค้า
โดยใช้แบบจำลองแมกซิมัมเอนโทรปีร่วมกับหลักความสัมพันธ์แบบพึ่งพา

EXTRACTING PRODUCT FEATURES AND OPINIONS IN PRODUCT REVIEWS USING MAXIMUM ENTROPY MODEL AND DEPENDENCY-BASED APPROACH

แสงกาญจน์ สมประเสริฐศรี
GAMGARN SOMPRASERTSRI

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2554

การสกัดคุณลักษณะสินค้าและความเห็นในการวิจารณ์สินค้า โดยใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา

EXTRACTING PRODUCT FEATURES AND OPINIONS IN PRODUCT REVIEWS USING MAXIMUM ENTROPY MODEL AND DEPENDENCY-BASED APPROACH

แกมกาญจน์ สมประเสริฐศรี

GAMGARN SOMPRASERTSRI

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2554

KMITL-2011-IT-D-001-001

EXTRACTING PRODUCT FEATURES AND OPINIONS IN PRODUCT REVIEWS USING MAXIMUM ENTROPY MODEL AND DEPENDENCY-BASED APPROACH

GAMGARN SOMPRASERTSRI

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2011

KMITL-2011-IT-D-001-001

หัวข้อวิทยานิพนธ์	การสกัดคุณลักษณะสินค้าและความเห็นในการวิจารณ์สินค้าโดยใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา
นักศึกษา	นางสาวแกมกาญจน์ สมประเสริฐศรี
รหัสประจำตัว	47066001
ปริญญา	ปรัชญาดุษฎีบัณฑิตสาขาวิชาเทคโนโลยีสารสนเทศ
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2554
อาจารย์ผู้ควบคุมวิทยานิพนธ์	ผศ.ดร.ภัทรชัย ลลิตโรจน์วงศ์

บทคัดย่อ

ปัจจุบันข้อมูลการวิจารณ์สินค้าถือเป็นข้อมูลที่มีประโยชน์ทั้งต่อผู้ที่สนใจซื้อสินค้า และผู้ผลิตสินค้า อย่างไรก็ตาม การวิเคราะห์ข้อมูลเหล่านี้ด้วยมนุษย์ต้องใช้เวลามากและสิ้นเปลืองสูง จึงจำเป็นต้องมีวิธีการสรุปความเห็นแบบอัตโนมัติขึ้น ปัญหาหลักในการสรุปความเห็น คือ การสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้านั้น งานวิจัยที่ผ่านมาใช้หลักการของการเกิดร่วมกันของคำ ซึ่งมีข้อจำกัดในการสกัดสำหรับประโยคที่มีโครงสร้างที่ซับซ้อน วิทยานิพนธ์นี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการที่สามารถสกัดคุณลักษณะสินค้าและความเห็นได้อย่างมีประสิทธิภาพ วิธีการที่นำเสนอนี้ใช้แบบจำลองแมกซิมั่มเอนโทรปีในการจัดประเภทให้เป็นความเห็นร่วมกับความสัมพันธ์แบบพึ่งพาระหว่างคุณลักษณะสินค้าและความเห็นซึ่งเป็นข้อมูลสำหรับการเรียนรู้ในการสกัดคุณลักษณะสินค้าและความเห็น ในการทดลองวัดประสิทธิภาพกับข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่และข้อมูลการวิจารณ์กล้องดิจิทัล ผลการทดลอง พบว่า ประสิทธิภาพของวิธีการที่นำเสนอในการสกัดคุณลักษณะสินค้าและความเห็นให้ค่าเอฟ เป็น 60.88% และ 66.19% ตามลำดับ และประสิทธิภาพในการสกัดประโยคที่แสดงความเห็นให้ค่าเอฟ เป็น 79.43% และ 71.62% ตามลำดับ เมื่อเปรียบเทียบกับวิธีการในงานวิจัยก่อนหน้านี้ซึ่งใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎ พบว่าวิธีการที่นำเสนอนี้มีประสิทธิภาพดีขึ้น อันจะทำให้สามารถนำไปประยุกต์ใช้ในงานสรุปความเห็นแบบอัตโนมัติที่มีประสิทธิภาพ และเป็นแนวทางในการพัฒนาวิธีการในการสกัดข้อสนเทศในงานด้านอื่นๆ ต่อไป

Thesis	Extracting Product Features and Opinions in Product Reviews Using Maximum Entropy Model and Dependency-Based Approach
Student	Miss Gamgarn Somprasertsri
Student ID	47066001
Degree	Doctor of Philosophy in Information Technology
Programme	Information Technology
Year	2011
Thesis Advisor	Asst.Prof.Dr. Pattarachai Lalitrojwong

ABSTRACT

Recently, product reviews is considered as a significant informative resource which is useful for both potential customers and product manufacturers. However, an analysis of those data with human often takes much time and be expensive. Therefore it is more efficient to automatically summarize various reviews. The high-level problem of opinion summarization addresses how to extract product features and its opinions. Many previous works usually depend on the co-occurrence of words. The co-occurrence based approaches are not sufficient to extract sentences which are grammatical complex. The thesis objective is to develop an approach for extracting product features and opinions that can solve these problems effectively. This approach is done by applying the maximum entropy model to classify if a pair of words is a product feature and its opinion. The classification learning process also uses the information obtained from the analysis of dependency between a product feature and its opinion. On cellular phone reviews and digital camera reviews, the experimental results show that the macro-averaged F-measure of the proposed approach for extracting product features and pinions are 60.88% and 66.19% respectively. The macro-averaged F-measure of the proposed approach for extracting opinion sentences are 79.43% and 71.62% respectively. The evaluation shows that the proposed approach provide more effectiveness than both using nearby adjectives and using extraction rules. This proposed methodology can be applied in automatic opinion summarization and used as a way improving information extraction methods in other applications.

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จได้ด้วยความกรุณาจากอาจารย์ที่ปรึกษา ผศ.ดร.ภัทรชัย ลลิตโรจน์วงศ์ ที่ให้ความช่วยเหลือ ให้คำชี้แนะช่วยแก้ปัญหา ตลอดจนให้ความรู้และประสบการณ์ที่ดีแก่ข้าพเจ้า

ขอขอบพระคุณ ศ.ดร.ชิดชนก เหลือสินทรัพย์ รศ.ดร.บุญธีร์ เครือตราชู รศ.ดร.อาริต ธรรมโน รศ.ดร.วรพจน์ กรีสุระเดช และผศ.ดร.พรฤดี เนติโสภากุล กรรมการสอบหัวข้อและวิทยานิพนธ์ที่ได้กรุณาให้คำแนะนำตลอดจนข้อชี้แนะ ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ยิ่งขึ้น

ขอขอบคุณมหาวิทยาลัยมหาสารคามที่ได้ให้โอกาสและให้การสนับสนุนการศึกษาครั้งนี้ และขอขอบคุณเพื่อนๆและน้องๆ ในห้องปฏิบัติการทุกคน

สุดท้ายต้องขอขอบคุณคุณแม่หมุด สมประเสริฐศรี พี่ๆ น้องๆ และหลานๆ ของข้าพเจ้า รวมถึงคุณอาภรณ์ คำก้อน ที่คอยช่วยเหลือและเป็นกำลังใจที่ดีตลอดมาจนในที่สุดทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

สำหรับคุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดามารดา ซึ่งเป็นที่รักและเคารพยิ่ง ตลอดจนครูอาจารย์ที่เคารพทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้และถ่ายทอดประสบการณ์ที่ดีให้แก่ข้าพเจ้า

แกมกาญจน์ สมประเสริฐศรี

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

สารบัญตาราง (ต่อ)

สารบัญรูป

สารบัญรูป (ต่อ)

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เว็บไซต์โดยทั่วไปในปัจจุบันมักจะให้บุคคลทั่วไปได้แสดงความเห็นผ่านเว็บไซต์ ซึ่งอาจจะเป็นความเห็นที่มีต่อสินค้า ความเห็นที่มีต่อบริการ หรือความเห็นที่มีต่อหน่วยงานนั้นๆ โดยเฉพาะเว็บไซต์ขายสินค้าที่ลูกค้าสามารถแสดงความเห็นต่อสินค้าได้ ตัวอย่างเว็บไซต์ที่มีการแสดงความเห็นต่อสินค้าของลูกค้า ได้แก่ เว็บไซต์ cnet.com, epinions.com, shopping.com และ amazon.com การแสดงความเห็นต่อสินค้านี้เป็นการวิจารณ์สินค้าซึ่งจะมีทั้งวิจารณ์สินค้าในด้านบวกและด้านลบ ข้อมูลการวิจารณ์สินค้าดังกล่าวเป็นข้อมูลที่มีประโยชน์ทั้งต่อผู้สนใจจะซื้อสินค้าและผู้ขายหรือผู้ผลิตสินค้า ผู้สนใจจะซื้อสินค้าสามารถใช้เป็นข้อมูลในการตัดสินใจซื้อสินค้าและผู้ผลิตสินค้าจะได้ทราบถึงความพึงพอใจหรือไม่พึงพอใจของลูกค้าที่มีต่อสินค้า สามารถนำข้อมูลไปพัฒนาสินค้าต่อไป แต่เนื่องจากข้อมูลการวิจารณ์สินค้ามีจำนวนมากและเพิ่มขึ้นอย่างรวดเร็ว จึงเป็นการยากที่ผู้สนใจซื้อสินค้าจะอ่านข้อมูลการวิจารณ์สินค้าต่างๆ ได้ทั้งหมด และเป็นปัญหาต่อผู้ขายหรือผู้ผลิตสินค้าในการวิเคราะห์ข้อมูลเหล่านั้นด้วย นอกจากนี้แล้ว ลักษณะของการวิจารณ์สินค้ายังเป็นลักษณะของการเขียนบรรยายเป็นประโยค ไม่ใช่ลักษณะของการถามตอบ ซึ่งจะทำให้ผู้เขียนสามารถถ่ายทอดความเห็นได้อย่างอิสระ ไม่ถูกจำกัดด้วยคำถามเหมือนคำตอบแบบปลายปิด แต่การเขียนบรรยายเป็นประโยคแบบอิสระนี้ ทำให้การวิเคราะห์ข้อมูลด้วยมนุษย์ต้องใช้เวลามากและมีความสิ้นเปลืองสูง จากปัญหาดังกล่าวข้างต้น จึงจำเป็นต้องมีการจัดการข้อมูลการวิจารณ์สินค้าเหล่านี้ เช่น การสรุปความเห็น เพื่อให้ได้ข้อมูลในลักษณะที่สามารถนำไปใช้งานได้อย่างมีประสิทธิภาพ ทำให้มีงานวิจัยทางด้านการสรุปความเห็นแบบอัตโนมัติจากข้อมูลการวิจารณ์สินค้าเกิดขึ้นในปัจจุบันอาทิเช่น [1][2] [3] [4] [5][6][7][8]

การสรุปความเห็นจากข้อมูลการวิจารณ์สินค้า จะพิจารณาจากคุณลักษณะของสินค้าและด้านของความเห็นที่แสดงต่อคุณลักษณะของสินค้านั้นๆ แล้วแสดงผลลัพธ์การจำแนกความเห็นตามคุณลักษณะสินค้าและด้านความเห็น การสรุปความเห็นในลักษณะนี้ จะทำให้สามารถทราบถึงความพึงพอใจหรือไม่พึงพอใจของลูกค้าที่มีต่อคุณลักษณะสินค้าต่างๆ โดยทั่วไปแล้วการสรุปความเห็นสามารถแบ่งงานออกได้เป็น 3 งาน [4] ดังนี้

1. สกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้านั้นๆ
2. พิจารณาด้านของความเห็นว่าเป็นด้านใด เช่น ด้านบวกหรือด้านลบ
3. แสดงผลลัพธ์ของการสรุปโดยแยกตามคุณลักษณะสินค้าและด้านของความเห็น

เมื่อพิจารณางานทั้ง 3 ของการสรุปความเห็นแล้ว พบว่างานหลักที่สำคัญของการสรุปความเห็น คือ การสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้านั้นๆ เพราะการจะสรุปความเห็นตามคุณลักษณะสินค้าได้ถูกต้องนั้น จะขึ้นอยู่กับการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้านั้นในประโยคให้ถูกต้อง แต่ถ้าการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้านั้นผิด ก็จะทำให้ผลลัพธ์ของการสรุปความเห็นนั้นไม่ถูกต้องด้วย

ตัวอย่างการสกัดคุณลักษณะสินค้า และความเห็นที่มีต่อคุณลักษณะสินค้าจากการวิจารณ์กล้องดิจิทัล "*This camera is very easy to use. The viewing screen is easy to see and very clear. The pictures are clear and good color. To compare other digital cameras we have used, this one is definitely superior and we would highly recommend.*" จากตัวอย่างข้อมูลการวิจารณ์กล้องดิจิทัล เราสามารถสกัดวลีความเห็น ได้ดังนี้ "*very easy to use*" "*viewing screen is easy to see and very clear*" และ "*pictures are clear and good color*" จะเห็นได้ว่าวลีเหล่านี้แสดงถึงความเห็นของลูกค้าที่มีต่อคุณลักษณะสินค้าของกล้องดิจิทัล คือ การใช้งาน ("*use*") จอภาพ ("*viewing screen*") และภาพ ("*picture*") ซึ่งลูกค้ามีความเห็นด้านบวก โดยพิจารณาได้จากคำแสดงความเห็น "*easy*" "*clear*" และ "*good*" ที่มีต่อคุณลักษณะสินค้าดังกล่าว

การสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้าของงานวิจัยที่ผ่านมา จะแยกงานออกเป็นงานย่อย 2 ส่วน คือ การสกัดคุณลักษณะสินค้าและการสกัดความเห็นที่มีต่อคุณลักษณะสินค้านั้น โดยกระบวนการจะเริ่มจากการสกัดคุณลักษณะสินค้าก่อน แล้วจึงนำคุณลักษณะสินค้าที่สกัดได้นั้นไปค้นหาหรือสกัดคำที่แสดงความเห็นที่มีต่อคุณลักษณะสินค้านั้นในประโยค ซึ่งวิธีการสกัดคุณลักษณะสินค้าของงานวิจัยที่ผ่านมาส่วนใหญ่จะอาศัยหลักการทางสถิติโดยพิจารณาเฉพาะคำที่คาดว่าจะเป็นคุณลักษณะสินค้า ได้แก่ การใช้ Association Rule Mining [2] การใช้ Point-Wise Mutual Information [4] และการใช้ความน่าจะเป็น [8] ส่วนการสกัดความเห็นที่มีต่อคุณลักษณะสินค้านั้นจะยึดหลักการของการเกิดร่วมกันระหว่างคำที่แสดงถึงคุณลักษณะสินค้าและคำที่แสดงความเห็น ได้แก่ วิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันโดยพิจารณาจากการเกิดขึ้นร่วมกันของคำคุณศัพท์ที่ทำหน้าที่ขยายคำนามที่แสดงถึงคุณลักษณะสินค้า [2] และวิธีการที่ใช้กฎซึ่งสร้างจากโครงสร้างทางไวยากรณ์ของรูปแบบที่เกิดขึ้นร่วมกันระหว่างคุณลักษณะสินค้าและความเห็นในประโยค [4]

วิธีการสกัดความเห็นดังกล่าวให้ผลดีกับประโยคที่มีโครงสร้างทางไวยากรณ์แบบง่าย ซึ่งประกอบด้วยประธาน กริยา และกรรม เช่น ประโยค "*The pictures are beautiful.*" แต่อย่างไรก็ตาม การสกัดความเห็นที่พิจารณาจากการเกิดขึ้นร่วมกันของคำที่ใกล้กัน โดยวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎ ยังมีข้อจำกัดในการสกัดความเห็นสำหรับประโยคแบบยาวที่มีโครงสร้างทางไวยากรณ์ที่ซับซ้อนขึ้น ดังตัวอย่างประโยคข้างล่างต่อไปนี้

(1) *It has **movie mode** that works good for a digital camera.*

(2) *It is great having the **LCD display**.*

(3) *I bought my **canon g3** about a month ago and i have to say i am very satisfied.*

(4) *The nice thing is that it uses the **SD memory card**.*

จากประโยคตัวอย่างข้างบน 4 ประโยค คำที่ขีดเส้นใต้จะหมายถึงคุณลักษณะสินค้า ส่วนคำที่เป็นตัวเข้มจะหมายถึงคำแสดงความเห็น ซึ่งเห็นได้ว่า คุณลักษณะสินค้าและคำแสดงความเห็นในประโยคทั้ง 4 ประโยค จะไม่อยู่ในโครงสร้างที่สามารถใช้วิธีการที่ผ่านมาซึ่งพิจารณาจากการเกิดขึ้นร่วมกันของคำที่ใกล้กัน ทั้งวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎได้

จากปัญหาดังกล่าว งานวิจัยนี้จึงนำเสนอวิธีการใหม่ในการสกัดคุณลักษณะสินค้าและความเห็น โดยใช้หลักการผสมระหว่างการใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพาในการสกัดคุณลักษณะสินค้าและความเห็น เนื่องจากข้อสนเทศที่ได้จากการวิเคราะห์ความสัมพันธ์แบบพึ่งพาเป็นข้อสนเทศจากบริบทในระยะไกล ไม่ถูกจำกัดขอบเขตของคำเหมือนวิธีการที่ผ่านมาช่วยให้สามารถสกัดคุณลักษณะสินค้าและความเห็นที่ไม่อยู่ในรูปแบบของการเกิดร่วมกันในระยะใกล้ได้

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการในการสกัดคุณลักษณะสินค้าและความเห็นที่มีประสิทธิภาพดียิ่งขึ้น ทั้งในแง่ของความถูกต้องในการสกัดและในแง่ของจำนวนประโยคที่สามารถสกัดได้ โดยประยุกต์ใช้หลักการผสมระหว่างการใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับการวิเคราะห์ความสัมพันธ์แบบพึ่งพา

1.3 ทฤษฎีหรือแนวความคิดที่ใช้ในการวิจัย

การสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้านั้น แนวทางส่วนใหญ่จะเริ่มจากการสกัดคุณลักษณะสินค้าก่อนแล้วจึงสกัดความเห็นที่มีต่อคุณลักษณะสินค้า การสกัดความเห็นที่มีต่อคุณลักษณะสินค้าด้วยวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน และวิธีการที่ใช้กฎ ยังมีข้อจำกัดดังที่กล่าวมาแล้ว งานวิจัยนี้จึงได้นำทฤษฎีด้านการเรียนรู้ด้วยเครื่อง และทฤษฎีด้านการประมวลผลภาษาธรรมชาติมาประยุกต์ใช้ร่วมกัน โดยใช้การวิเคราะห์ความสัมพันธ์แบบพึ่งพาร่วมกับแบบจำลองแมกซิมั่มเอนโทรปี (Maximum Entropy Model) ซึ่งเป็นแบบจำลองที่สร้างในลักษณะของการแจกแจงแบบสม่ำเสมอ (Uniform Distribution) โดยจะจำลองทุกสิ่งตามข้อเท็จจริงทั้งหมดที่มีอยู่ แต่ไม่ตั้งข้อสันนิษฐานใดๆเกี่ยวกับข้อเท็จจริงที่ไม่รู้ คุณลักษณะเด่นของแบบจำลองนี้ คือ เป็นแบบจำลองที่สามารถรวมเอาข้อเท็จจริงต่างๆ มาใช้ในการประมาณความน่าจะเป็น

สำหรับการเรียนรู้เพื่อสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้า ในงานวิจัยนี้จะใช้ข้อเท็จจริงที่ได้จากการวิเคราะห์ความสัมพันธ์แบบพึ่งพาระหว่างคุณลักษณะสินค้ากับความเห็น ได้แก่ ประเภทของความสัมพันธ์และเส้นทางของความสัมพันธ์แบบพึ่งพา ทำให้แบบจำลองได้เรียนรู้ลักษณะของรูปแบบที่เกิดขึ้นระหว่างคุณลักษณะสินค้าและความเห็น สามารถตัดสินใจในการสกัดได้อย่างถูกต้องเหมาะสม

1.4 ขอบเขตของการวิจัย

การวิจัยครั้งนี้เป็นการออกแบบ และพัฒนาวิธีการในการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้า สำหรับประมวลผลข้อมูลการวิจารณ์สินค้าที่เป็นภาษาอังกฤษ ซึ่งมีรูปแบบของการแสดงความเห็นอยู่ในรูปแบบอิสระ (Free Format) มีการเขียนบรรยายเป็นประโยค ที่ยังไม่มีการจำแนกด้านความเห็นว่าเป็นด้านบวกหรือด้านลบ และไม่มีการแสดงคุณลักษณะสินค้าและความเห็นแบบแฝง

1.5 ขั้นตอนของการศึกษา

การวิจัยครั้งนี้สามารถแสดงลำดับขั้นตอนของการศึกษาตั้งแต่เริ่มต้นจนถึงสิ้นสุดได้ดังรายละเอียดต่อไปนี้

1. ศึกษาทฤษฎี และงานวิจัยจากบทความ และเอกสารต่างๆ ที่เกี่ยวข้องกับงานวิจัย
2. ออกแบบ และวิเคราะห์วิธีการในการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้าแบบอัตโนมัติ
3. พัฒนาโปรแกรมตามวิธีการที่นำเสนอ พร้อมทั้งแก้ไขข้อผิดพลาด
4. เตรียมข้อมูลการวิจารณ์สินค้าที่ใช้งานจริง เพื่อนำมาทดสอบวิธีการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้า
5. ทดลองสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้าจากข้อมูลการวิจารณ์สินค้าที่ใช้งานจริงด้วยวิธีการที่นำเสนอ พร้อมทั้งวัดประสิทธิภาพ และเปรียบเทียบประสิทธิภาพของการทำงานกับวิธีการที่ผ่านมา
6. วิเคราะห์และสรุปผลการทดลอง
7. เรียบเรียงเอกสารประกอบวิทยานิพนธ์

1.6 นิยามศัพท์เฉพาะ

ในงานวิจัยนี้ ได้ให้คำจำกัดความของคุณลักษณะสินค้า ความเห็น และคู่คุณลักษณะสินค้าและความเห็น ไว้ดังนี้

1. คุณลักษณะสินค้า หมายถึง คำหรือวลีที่อาจเป็น ยี่ห้อของสินค้า (เช่น Cannon และ Sony เป็นต้น) รุ่นของสินค้า (เช่น Powershot G3 เป็นต้น) คุณสมบัติ (เช่น ขนาดกล้อง และ สี เป็นต้น) ส่วนประกอบของสินค้า (เช่น เลนส์ และ แบตเตอรี่ เป็นต้น) คุณลักษณะของส่วนประกอบสินค้า (เช่น ระยะเวลาของแบตเตอรี่ เป็นต้น) หรือ คอนเซฟที่เกี่ยวข้อง (เช่น การออกแบบ เป็นต้น) โดยทั่วไปคุณลักษณะสินค้าอาจเป็นตัวสินค้าเอง หรือคุณลักษณะของสินค้า ตัวอย่างเช่น ประโยคการวิจารณ์สินค้า "*The picture quality is amazing*" ที่ "*picture quality*" คือ คุณลักษณะสินค้าที่กล่าวถึงคุณภาพของรูปภาพ

2. ความเห็น หมายถึง คำที่แสดงออกตามที่เห็น รู้ หรือคิด ซึ่งมี 2 ด้าน คือ ด้านบวก และด้านลบ ตัวอย่างเช่น ประโยคการวิจารณ์สินค้า "*The picture quality is amazing*" ที่มี "*amazing*" เป็นความเห็นด้านบวกที่มีต่อคุณภาพของรูปภาพ

3. คู่คุณลักษณะสินค้าและความเห็น หมายถึง คุณลักษณะสินค้าและความเห็นที่เป็นการแสดงความเห็นซึ่งอาจจะเป็นด้านบวกหรือด้านลบต่อคุณลักษณะสินค้านั้น ตัวอย่างเช่น ประโยคการวิจารณ์สินค้า "*The picture quality is amazing*" ที่มี *[picture quality, amazing]* เป็นคู่คุณลักษณะสินค้าและความเห็นที่กล่าวถึงคุณภาพของรูปภาพและมีการแสดงความเห็นด้านบวกต่อคุณภาพของรูปภาพ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีพื้นฐานต่างๆ ที่เกี่ยวข้องกับงานวิจัย รูปแบบการวิจารณ์สินค้า การสรุปความเห็น วิธีการสกัดคุณลักษณะสินค้าและความเห็นในงานวิจัยที่ผ่านมา พร้อมทั้งแสดงถึงงานวิจัยที่เกี่ยวข้องและปัญหาของการสกัดคุณลักษณะสินค้าและความเห็น

2.1 รูปแบบการวิจารณ์สินค้า

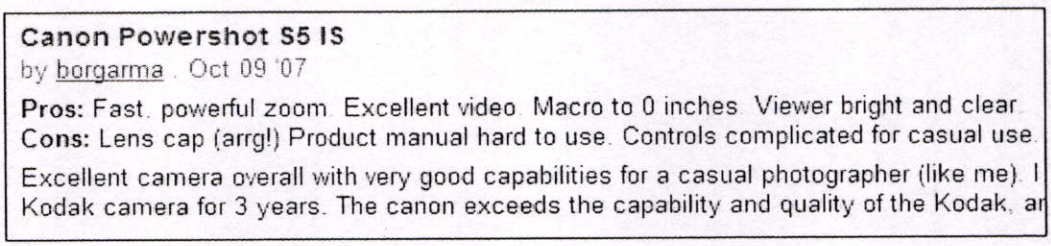
รูปแบบของการให้ลูกค้าวิจารณ์สินค้าผ่านเว็บไซต์สามารถแบ่งได้ 3 รูปแบบ [9] ดังนี้

1. รูปแบบการวิจารณ์สินค้าแบ่งด้านบวกและลบ (Pros and Cons) เป็นรูปแบบที่มีการแบ่งคำวิจารณ์ออกเป็นด้านบวกและลบ โดยผู้วิจารณ์จะเขียนประโยคการวิจารณ์แยกกันระหว่างด้านบวกและด้านลบ ตัวอย่างเว็บไซต์ที่มีรูปแบบการวิจารณ์สินค้ารูปแบบนี้ ได้แก่ เว็บไซต์ cnet.com โดยมีตัวอย่างการวิจารณ์สินค้าดังรูปที่ 2.1

> by dturano547 (see profile) - June 15, 2007
>
> **Pros:** Super Fast camera.great quality images as always from canon. hi res screen for viewing images in all lighting. 4x optical zoom. Image stabilization, very durable case, nothing feels cheap or flimsy.
>
> **Cons:** Im not a fan of the viewfinder but it can come in handy. controls were akward at first. no case or dock included. nothing major to gripe about with this camera.
>
> **51** out of **52** users found this opinion helpful (see all 3 comments)

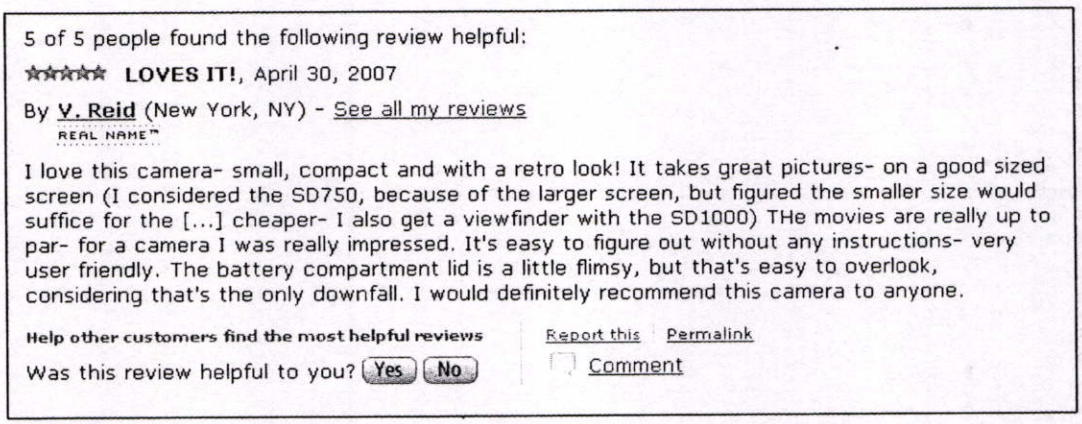
รูปที่ 2.1ตัวอย่างการวิจารณ์สินค้ารูปแบบที่ 1

2. รูปแบบการวิจารณ์สินค้าแบ่งด้านบวก ลบ และรายละเอียดของการวิจารณ์ (Pros, Cons and Detailed Review) เป็นการวิจารณ์สินค้าที่แบ่งออกเป็นด้านบวกและด้านลบ ซึ่งเป็นประโยคสั้นๆ หรือวลี แต่มีรายละเอียดของการวิจารณ์เพิ่มเติม ตัวอย่างเว็บไซต์ที่มีรูปแบบการวิจารณ์สินค้ารูปแบบนี้ ได้แก่ เว็บไซต์ epinion.com โดยมีตัวอย่างการวิจารณ์สินค้าดังรูปที่ 2.2



รูปที่ 2.2 ตัวอย่างการวิจารณ์สินค้ารูปแบบที่ 2

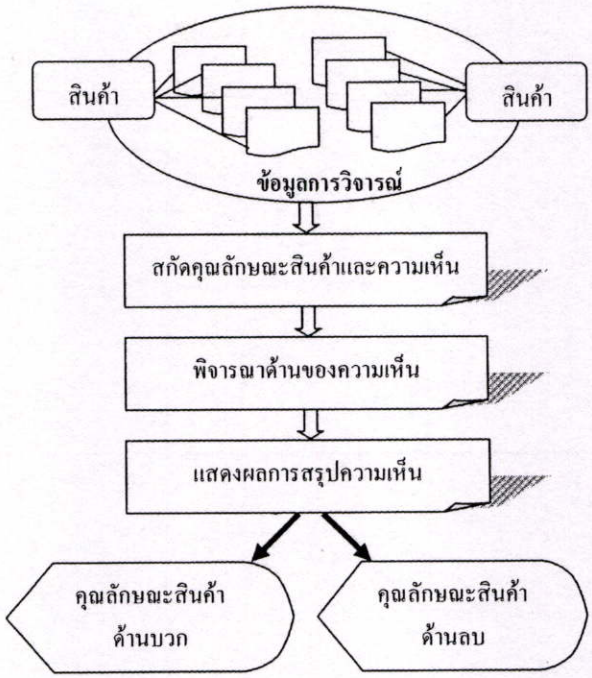
3. รูปแบบอิสระ (Free Format) เป็นการวิจารณ์สินค้าที่มีการเขียนในลักษณะของประโยคแบบสั้นและแบบประโยคยาว และไม่มีการแบ่งการวิจารณ์ออกเป็นด้านบวกและด้านลบ ตัวอย่างเว็บไซต์ที่มีรูปแบบการวิจารณ์สินค้ารูปแบบนี้ ได้แก่ เว็บไซต์ amazon.com โดยมีตัวอย่างการวิจารณ์สินค้าดังรูปที่ 2.3



รูปที่ 2.3 ตัวอย่างการวิจารณ์สินค้ารูปแบบที่ 3

2.2 การสรุปความเห็น

การสรุปความเห็น (Opinion summarization) จากข้อมูลการวิจารณ์สินค้า เป็นการวิเคราะห์ความเห็น โดยมุ่งเน้นที่จะสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะของสินค้า แล้วจำแนกด้านของความเห็น และนำเสนอผลลัพธ์ในรูปแบบที่นำไปใช้งานได้อย่างมีประสิทธิภาพ โดยทั่วไปการสรุปความเห็นจะแบ่งงานออกเป็น 3 งานหลักๆ (รูปที่ 2.4) ดังนี้



รูปที่ 2.4 กระบวนการสำหรับการสรุปความเห็น

1. การสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้า เป็นงานส่วนแรกที่สำคัญสำหรับการสรุปความเห็น โดยในงานส่วนนี้จะพยายามสกัดคุณลักษณะสินค้าในประโยคการวิจารณ์สินค้า และสกัดความเห็นที่มีต่อคุณลักษณะสินค้าที่สกัดได้นั้น ตัวอย่างเช่น

"This camera is very easy to use. The viewing screen is very clear. The pictures are clear and good color. To compare other digital cameras we have used, this one is definitely superior and we would highly recommend."

จากตัวอย่างข้อมูลการวิจารณ์กล้องดิจิทัล เราสามารถสกัดวลี ที่บ่งบอกถึงการแสดงความเห็นต่อสินค้า ได้แก่ "*very easy to use*" "*viewing screen is very clear*" และ "*pictures are clear and good color*" ในแต่ละวลีสามารถสกัดคุณลักษณะสินค้าได้ คือ "*use*" "*screen*" และ "*picture*" และคำที่แสดงถึงความเห็น คือ "*easy*" "*clear*" และ "*good*" ตามลำดับ

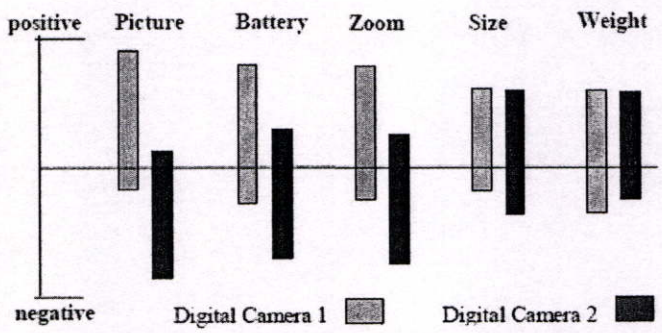
2. การพิจารณาด้านของความเห็น หลังจากการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้าแล้ว งานต่อไปจะเป็นการพิจารณาว่าความเห็นที่มีต่อคุณลักษณะสินค้านั้นเป็นความเห็นในด้านบวกหรือด้านลบ จากตัวอย่างข้อมูลการวิจารณ์กล้องดิจิทัล ที่กล่าวมาแล้ว จะเห็นว่ามีคำแสดงความเห็น คือ "*easy*" "*clear*" และ "*good*" ซึ่งเป็นการแสดงความเห็นที่มีต่อ

คุณลักษณะการใช้ คุณลักษณะหน้าจอ และคุณลักษณะภาพของกล้องดิจิทัล ซึ่งพิจารณาด้านของความเห็นแล้วเป็นความเห็นด้านบวกทั้งหมด

3. การแสดงผลลัพธ์ของการสรุปความเห็นตามคุณลักษณะสินค้า เป็นการนำเสนอผลลัพธ์ของการสรุปความเห็น ในรูปแบบที่นำไปใช้งานได้อย่างมีประสิทธิภาพ ซึ่งอาจจะมีการแสดงผลจำแนกตามคุณลักษณะของสินค้า และมีการจำแนกด้านของการแสดงความเห็น [2] แสดงดังรูปที่ 2.5 หรือมีการแสดงเป็นแผนภูมิแท่งเปรียบเทียบในแต่ละคุณลักษณะสินค้า [3] ตัวอย่างแสดงดังรูปที่ 2.6

Digital_camera_1:
Feature: **picture quality**
Positive: 253
<individual review sentences>
Negative: 6
< individual review sentences >
Feature: **size**
Positive: 134
<individual review sentences>
Negative: 10
< individual review sentences >
...

รูปที่ 2.5 ผลการสรุปความเห็นจำแนกตามคุณลักษณะสินค้า



รูปที่ 2.6 แผนภูมิแท่งแสดงผลสรุปความเห็นของแต่ละคุณลักษณะสินค้า

2.3 การวัดประสิทธิภาพของการสกัดคุณลักษณะสินค้าและความเห็น

การวัดประสิทธิภาพในการสกัดคุณลักษณะสินค้าและความคิดเห็น จะอาศัยวิธีพื้นฐานในการวัดประสิทธิภาพของการค้นคืนสารสนเทศ ซึ่งสามารถคำนวณค่าประสิทธิภาพได้ดังนี้ [10]

1. ค่าความระลึก (Recall) หมายถึง ความสามารถในการสกัดคุณลักษณะสินค้าและความคิดเห็น จากข้อมูลการวิจารณ์สินค้าได้ มีการคำนวณดังนี้

$$\text{Recall (R)} = \frac{\text{จำนวนของคุณลักษณะสินค้าและความเห็นที่สกัดได้ถูกต้องจากระบบ}}{\text{จำนวนของคุณลักษณะสินค้าและความเห็นทั้งหมด}}$$

ตัวอย่างเช่น ถ้ามีจำนวนคุณลักษณะสินค้าและความเห็นทั้งหมด 30 คู่ และระบบที่สร้างขึ้นสามารถสกัดคุณลักษณะสินค้าและความเห็นได้ถูกต้องจำนวน 25 คู่ แสดงว่า ระบบนี้มีค่าความระลึกคือ 25/30 เท่ากับ 0.83 หรือ 83%

2. ค่าความเที่ยงตรง (Precision) หมายถึง ความสามารถในการสกัดคุณลักษณะสินค้าและความเห็นได้ถูกต้อง มีการคำนวณดังนี้

$$\text{Precision (P)} = \frac{\text{จำนวนของคุณลักษณะสินค้าและความเห็นที่สกัดได้ถูกต้องจากระบบ}}{\text{จำนวนของคุณลักษณะสินค้าและความเห็น ที่สกัดได้ทั้งหมดจากระบบ}}$$

ตัวอย่างเช่น ถ้าระบบที่สร้างขึ้นมาสามารถสกัดคุณลักษณะสินค้าและความเห็นได้ทั้งหมด 26 คู่ และมีคุณลักษณะสินค้าและความเห็นที่ถูกต้องจำนวน 25 คู่ แสดงว่า ระบบนี้มีค่าความเที่ยงตรง คือ 25/26 เท่ากับ 0.96 หรือ 96%

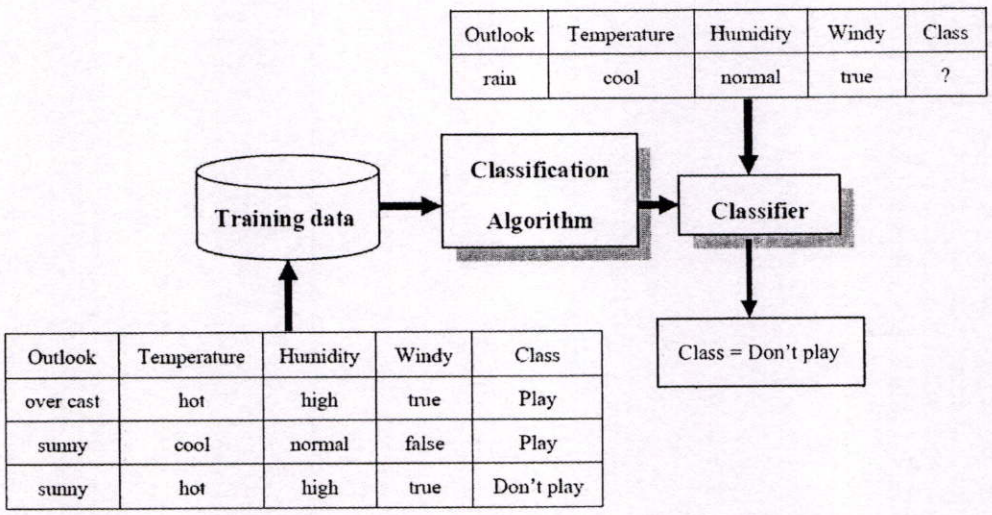
3. ค่าเอฟ (F-Measure) หรือ Harmonic Mean เป็นการเฉลี่ยค่าความเที่ยงตรงและค่าความระลึกในการสกัดคุณลักษณะสินค้าและความเห็นเข้าด้วยกัน จึงเป็นเหมือนค่าวัดความแม่นยำโดยรวม มีการคำนวณค่าดังนี้

$$\text{F-Measure (F)} = \frac{(\beta^2+1)PR}{\beta^2 R + P}$$

โดย β คือค่าพารามิเตอร์ที่แสดงสัดส่วนความสำคัญระหว่างค่าความเที่ยงตรง และค่าความระลึก โดยทั่วไป จะใช้ค่า β เท่ากับ 1 ตัวอย่างเช่น ถ้ามีค่าความระลึกเท่ากับ 83% และค่าความเที่ยงตรงเท่ากับ 96% คำนวณค่า F ได้ดังนี้ 2 x (0.83 x 0.96) / (0.83 + 0.96) เท่ากับ 0.89 หรือ 89%

2.4 การจำแนกประเภทข้อมูล

การจำแนกประเภทข้อมูล (Data Classification) เป็นการแบ่งหมวดหมู่ของข้อมูลให้อยู่ในกลุ่มหรือหมวดหมู่ที่มีการกำหนดไว้ก่อน โดยอาศัยโมเดลหรือแบบจำลองที่สร้างขึ้นจากการนำคุณสมบัติของข้อมูลส่วนหนึ่งมาสอนหรือให้การเรียนรู้ (Training Data) เรียกแบบจำลองที่ผ่านการเรียนรู้ว่าเป็นตัวจำแนก (Classifier) ซึ่งสามารถทำนายกลุ่มของข้อมูลใหม่ที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ โดยทำการเปรียบเทียบข้อมูลนั้นกับคุณสมบัติของต้นแบบในแต่ละหมวดหมู่ เพื่อตัดสินใจความเป็นไปได้ของประเภทหรือหมวดหมู่ของข้อมูลใหม่ เช่น ปัญหาการตัดสินใจในการเล่นกอล์ฟ มีผลการทำนาย คือ เล่นได้ (Play) หรือเล่นไม่ได้ (Don't Play) และมีรายละเอียดของข้อมูล คือ ทัศนียภาพ (Outlook) อุณหภูมิ (Temperature) ความชื้น (Humidity) และสถานะลม (Windy) ที่จะนำมาใช้ในการตัดสินใจ ขั้นตอนการจำแนกสามารถแสดงได้ดังรูปที่ 2.7



รูปที่ 2.7 การจำแนกประเภทข้อมูล

ในปัจจุบันวิธีการจำแนกประเภทข้อมูลมีหลายวิธีซึ่งมีหลักการทำงานแตกต่างกันออกไป ได้แก่

1. ต้นไม้ตัดสินใจ (Decision Tree) เป็นวิธีในการจำแนกประเภทข้อมูลในลักษณะของโครงสร้างต้นไม้โดยประกอบไปด้วยกฎที่ใช้ในการตัดสินใจ การจำแนกด้วยต้นไม้ตัดสินใจนี้สามารถทำความเข้าใจได้ง่ายเมื่อเทียบกับวิธีการจำแนกประเภทข้อมูลแบบอื่น แต่วิธีการนี้ส่วนใหญ่จะไม่รองรับข้อมูลแบบต่อเนื่อง

2. โครงข่ายประสาทเทียม (Artificial Neural Network) มีพื้นฐานมาจากแบบจำลองการทำงานของสมองมนุษย์ โดยวิธีการนี้มีความซับซ้อนมากกว่าวิธีการจำแนกประเภทข้อมูล

แบบอื่นๆ ค่อนข้างมาก นอกจากนี้ ผลลัพธ์ที่ได้ยังยากต่อการทำความเข้าใจและยากต่ออธิบาย วิธีการนี้จึงมักถูกเรียกว่า Black Box

3. แบบจำลองความน่าจะเป็น (Probabilistic Model) เป็นวิธีการจำแนกประเภทข้อมูลที่อาศัยหลักการของความน่าจะเป็น สามารถแบ่งย่อยได้เป็น 2 ลักษณะ คือ ความน่าจะเป็นร่วม (Joint Probability) ซึ่งเป็นความน่าจะเป็นร่วมกันระหว่าง 2 เหตุการณ์และความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability) ที่เหตุการณ์ 2 เหตุการณ์ มีความสัมพันธ์กันในลักษณะที่การเกิดหรือไม่เกิดของเหตุการณ์หนึ่งมีผลต่อความน่าจะเป็นของอีกเหตุการณ์หนึ่ง วิธีการจำแนกประเภทข้อมูลที่ใช้หลักการของความน่าจะเป็นร่วม จะถูกเรียกว่า Generative Model ซึ่งได้แก่ Naïve Bayes Model (NB) และ Hidden Markov Model (HMM) สำหรับวิธีการจำแนกประเภทข้อมูลที่ใช้หลักการของความน่าจะเป็นแบบมีเงื่อนไข จะถูกเรียกว่า Discriminative Model ซึ่งได้แก่ Maximum Entropy Model (ME) และ Conditional Random Field (CRF)

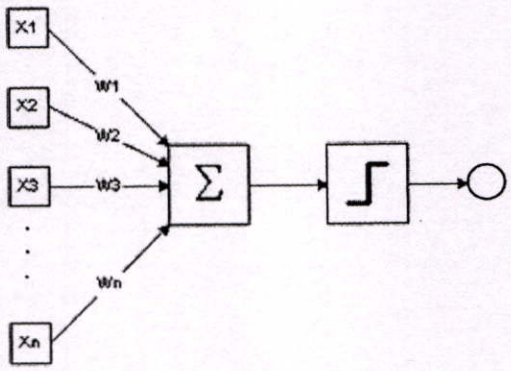
สำหรับ HMM และ CRF เป็นวิธีการจำแนกประเภทข้อมูลสำหรับข้อมูลแบบเรียงลำดับต่อเนื่อง (Sequential Data) และในการทบทวนวรรณกรรมที่ผ่านมา วิธีการจำแนกประเภทข้อมูลที่อาศัยหลักการของความน่าจะเป็นร่วม $P(x|t)$ (โดยที่ x คือบริบทหรือคุณสมบัติที่ใช้จำแนก และ t คือ หมวดหมู่ของการจำแนก) เมื่อต้องการคำนวณค่าความน่าจะเป็นเพื่อจะทำให้แบบจำลองในการจำแนกได้ผลดีกว่าวิธีการจำแนกประเภทข้อมูลที่อาศัยหลักการของ ความน่าจะเป็นแบบมีเงื่อนไข $P(t|x)$ จำเป็นต้องอาศัย x จำนวนมาก ซึ่งมีข้อจำกัดในสภาพความเป็นจริงของข้อมูล จึงเป็นการยากที่วิธีการจำแนกประเภทข้อมูล Generative Model ที่ใช้หลักการของความน่าจะเป็นร่วม จะให้ผลในการจำแนกดีกว่าวิธีการจำแนกประเภทข้อมูล Discriminative Model ที่ใช้หลักการของความน่าจะเป็นแบบมีเงื่อนไข [11] ในปัจจุบันได้มีการนำเสนอวิธีการจำแนกประเภทข้อมูลใหม่ที่ผสมผสานกันระหว่าง Maximum Entropy Model และ Hidden Markov Model เป็น Maximum Entropy Markov Model (MEMM) ซึ่งเป็นแบบจำลองที่มีประสิทธิภาพมากขึ้นสำหรับข้อมูลแบบเรียงลำดับต่อเนื่อง [12]

การวิจัยนี้สามารถนำวิธีการจำแนกประเภทมาใช้ในการจำแนกข้อมูลว่าเป็นหรือไม่เป็นคู่คุณลักษณะสินค้าและความเห็น โดยเลือกใช้แบบจำลองแมกซิมั่มเอนโทรปีเป็นตัวจำแนกประเภท ซึ่งเป็นวิธีการจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่งที่นิยมใช้แก้ปัญหาในงานทางด้านการประมวลผลภาษาธรรมชาติ เช่น ปัญหาการกำกับหมวดคำ (POS Tagging) ปัญหาการจำแนกประเภทข้อความ (Text Categorization) ปัญหาการสกัดนิพจน์ระบุนาม (Name Entity Detection) และปัญหาการแจงส่วน (Parsing) เนื่องจากเป็นตัวจำแนกประเภทข้อมูลที่มีประสิทธิภาพ คุณลักษณะเด่นของแบบจำลองนี้ คือเป็นแบบจำลองที่สามารถรวมเอาข้อมูลบริบทต่างๆจำนวนมากมาใช้ประมาณค่าความน่าจะเป็นของประเภทข้อมูล a ในแต่ละบริบท b ใดๆได้ และในสภาพความเป็นจริง บริบทที่ใช้สำหรับการสกัดคุณลักษณะสินค้าและความเห็นในงานวิจัยนี้ ได้แก่ คำ

เส้นทางของความสัมพันธ์ และประเภทความสัมพันธ์ของคุณลักษณะสินค้าและความเห็น ซึ่งขึ้นอยู่กับลักษณะของประโยคและคำที่ใช้ในการวิจารณ์สินค้าของผู้วิจารณ์ จึงมีความหลากหลายและมีการกระจายของข้อมูลมาก ทำให้ไม่สามารถประมาณค่าความน่าจะเป็น $p(a,b)$ ที่แท้จริงได้ ดังนั้นแบบจำลองแมกซิมั่มเอนโทรปีซึ่งใช้แนวคิดของการแจกแจงแบบสม่ำเสมอ จะพยายามสร้างแบบจำลองให้มีลักษณะที่สม่ำเสมอตามข้อเท็จจริงทั้งหมดที่มีอยู่ และไม่ตั้งข้อสันนิษฐานใดๆ เกี่ยวกับข้อเท็จจริงที่ไม่รู้ ทำให้สามารถประมาณค่าความน่าจะเป็น $p(a,b)$ ของแบบจำลองได้อย่างน่าเชื่อถือและใกล้เคียงกับความเป็นจริงมากที่สุด การวิจัยนี้ได้ทำการวัดประสิทธิภาพในการจำแนกคุณลักษณะสินค้าและความเห็นของตัวจำแนกโครงข่ายประสาทเทียมและตัวจำแนกเบส์อย่างง่าย เพื่อทำการเปรียบเทียบประสิทธิภาพกับแบบจำลองแมกซิมั่มเอนโทรปี รายละเอียดนำเสนอในบทที่ 4

2.5 โครงข่ายประสาทเทียม

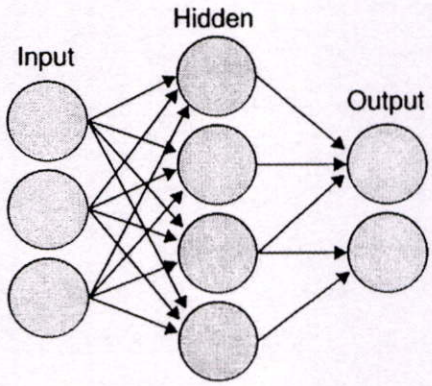
โครงข่ายประสาทเทียม คือโมเดลทางคณิตศาสตร์สำหรับประมวลผลข้อมูลด้วยคอมพิวเตอร์ที่จำลองมาจากกระบวนการทำงานบางส่วนของสมองมนุษย์ ซึ่งประกอบด้วยเซลล์ประสาท และจุดประสานประสาท แต่ละเซลล์ประสาทประกอบด้วย ตัวเซลล์ ใยประสาทนำเข้า ซึ่งเป็นเหมือนส่วนนำเข้า และใยประสาทนำออก ซึ่งเป็นเหมือนส่วนนำออกของเซลล์ สำหรับโครงข่ายประสาทเทียมมีความสามารถในการจดจำรูปแบบ (Pattern Recognition) และการอุปมานความรู้ (Knowledge Deduction) เช่นเดียวกับความสามารถของสมองมนุษย์ การเรียนรู้ของโครงข่ายประสาทเทียมจะเรียนรู้จากชุดการสอน โดยการส่งข้อมูลเข้ามายังส่วนที่เรียกว่าเพอร์เซปตรอน (Perceptron) ที่จำลองลักษณะของเซลล์ประสาทในสมองของมนุษย์แสดงดังรูปที่ 2.8 [13]



รูปที่ 2.8 เพอร์เซปตรอนของโครงข่ายประสาทเทียม

เพอร์เซปตรอนรับอินพุตเป็นเวกเตอร์จำนวนจริง แล้วคำนวณหาผลรวมเชิงเส้นของค่าน้ำหนัก $w_1, w_2, w_3,\ldots, w_n$ ของอินพุต $x_1, x_2, x_3,\ldots, x_n$ และเปรียบเทียบผลรวมกับค่าขีดแบ่ง (Threshold) โดยถ้าผลรวมที่ได้มีค่าเกินค่าขีดแบ่งเอาต์พุต O ที่ได้จะเป็น 1 และถ้ามีค่าไม่เกินค่าขีดแบ่งเอาต์พุต O จะเป็น -1

อัลกอริทึมการแพร่กระจายย้อนกลับ (Back Propagation Algorithm) [12] เป็นอัลกอริทึมที่นิยมใช้วิธีหนึ่งในการเรียนรู้ของโครงข่ายประสาทเทียมหลายชั้น ในการปรับค่าน้ำหนักในเส้นเชื่อมต่อระหว่างโหนดให้เหมาะสม เพื่อหาค่าต่ำสุดของค่าผิดพลาดระหว่างเอาต์พุตของโครงข่ายที่คำนวณได้กับเอาต์พุตเป้าหมาย โครงข่ายประกอบด้วยชั้น 3 ชั้น คือ ชั้นนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) แสดงดังรูปที่ 2.9



รูปที่ 2.9 โครงข่ายประสาทเทียมหลายชั้นแบบหนึ่งชั้นซ่อน

2.6 ตัวจำแนกเบส์อย่างง่าย

ตัวจำแนกเบส์อย่างง่าย (Naïve Bayes Classifier) เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่งที่มีพื้นฐานมาจากทฤษฎีของเบส์ เพื่อสร้างแบบจำลองที่อยู่ในรูปแบบความน่าจะเป็น ตัวจำแนกเบส์อย่างง่ายมีอัลกอริทึมในการทำงานที่ไม่ซับซ้อนและเหมาะกับกรณีที่ข้อมูลเซตตัวอย่างมีจำนวนมากและมีคุณสมบัติไม่ขึ้นต่อกัน จึงมีการนำไปประยุกต์ใช้งานด้านการจำแนกประเภทข้อความ สมมุติให้ $A_1, A_2, A_3,\ldots,A_n$ เป็นคุณสมบัติของตัวอย่าง x การจำแนกของตัวจำแนกเบส์อย่างง่ายแสดงได้ดังสมการข้างล่างนี้ [13]

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \times \prod_{i=1}^{n} P(a_i \mid v_j)$$

โดยที่ a_i ในสมการเป็นค่าของคุณสมบัติ A_i และ V เป็นเซตของประเภทที่เป็นไปได้ของตัวอย่าง x

ตัวอย่างการใช้ตัวจำแนกเบส์อย่างง่ายในการจำแนกคู่ที่เป็นคุณลักษณะสินค้าและความเห็น สมมุติข้อมูลตัวอย่างสอนสำหรับการเรียนรู้เบส์อย่างง่ายแสดงดังตารางที่ 2.1

ตารางที่ 2.1 ตัวอย่างสอนสำหรับการเรียนรู้เบส์อย่างง่ายในการจำแนกคู่ที่เป็นคุณลักษณะสินค้า

คำคุณลักษณะสินค้า	คำความเห็น	เส้นทางความสัมพันธ์	ประเภทความสัมพันธ์	ผลลัพธ์
camera	bad	NNJJ	parent	yes
camera	slow	NNVBJJ	grandchild	no
picture	good	NNJJ	child	yes
battery	like	NNVB	child	yes
battery	good	NNVBJJ	grandchild	no
camera	slow	NNJJ	child	no
picture	good	NNVBJJ	grandchild	no
battery	bad	NNVBJJ	grandchild	no

สมมุติว่าตัวอย่างที่ต้องการจำแนก คือ [battery, good, NNJJ, child, ?] โดยที่ V = {yes, no} สามารถคำนวณ

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \times \prod_{i=1}^{n} P(a_i \mid v_j)$$

กรณี v_j = yes ได้ว่า

$$P(yes)P(battery|yes)P(good|yes)P(NNJJ|yes)P(child|yes) = \frac{3}{8}\times\frac{1}{3}\times\frac{1}{3}\times\frac{2}{3}\times\frac{2}{3} = \frac{1}{54}$$

กรณี v_j = no ได้ว่า

$$P(no)P(battery|no)P(good|no)P(NNJJ|no)P(child|no) = \frac{5}{8}\times\frac{2}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{1}{5} = \frac{1}{250}$$

ดังนั้นได้ v_{NB} = yes หมายความว่าตัวอย่างที่ต้องการจำแนกจะถูกสกัดว่าเป็นคู่คุณลักษณะสินค้าและความเห็น

2.7 แบบจำลองแมกซิมั่มเอนโทรปี

แมกซิมั่มเอนโทรปีถูกนำมาประยุกต์ใช้ และนำเสนอครั้งแรกโดย Jaynes [14] ในปี ค.ศ. 1957 ปัจจุบันถูกนำมาใช้อย่างกว้างขวางในปัญหาของการจำแนกประเภท โดยเฉพาะปัญหาในการประมวลผลภาษาธรรมชาติ ซึ่งสามารถมองเสมือนเป็นปัญหาของการจำแนกประเภทได้ เช่น ปัญหาการตรวจหาขอบเขตของประโยค [15][16] ปัญหาการสกัดข้อสนเทศ [17] และปัญหาการกำกับหมวดคำ [18] เป็นต้น กระบวนการทำงานของการจำแนกประเภทจะเป็นการประมาณค่าความน่าจะเป็นของ "ประเภท" หรือ คลาส (Class) a ที่เกิดขึ้นร่วมกับ "บริบท" (Context) b นั่นคือการหาค่า $p(a,b)$ [19] แต่อย่างไรก็ดีในสภาพความเป็นจริง ข้อมูลที่มีอยู่มักไม่เพียงพอที่จะสามารถคำนวณหาค่าความน่าจะเป็นที่แท้จริงของ $p(a,b)$ ได้ เนื่องจากค่าของบริบท b มักกระจายและมีความหลากหลาย จึงต้องหาวิธีเหมาะสมที่สามารถประมาณค่าความน่าจะเป็น $p(a,b)$ จากข้อมูลอันจำกัดที่มีอยู่เกี่ยวกับ a และ b ให้ใกล้เคียงกับความเป็นจริงมากที่สุด

แบบจำลองแมกซิมั่มเอนโทรปีเป็นแบบจำลองความน่าจะเป็นที่มีการสร้างแบบจำลองในลักษณะที่เป็นการแจกแจงแบบสม่ำเสมอให้ได้มากที่สุดเท่าที่จะเป็นไปได้ โดยแบบจำลองจะจำลองทุกสิ่งตามข้อเท็จจริงทั้งหมดที่มีอยู่ แต่ไม่ตั้งข้อสันนิษฐานใดๆ เกี่ยวกับข้อเท็จจริงที่ไม่รู้ และอาศัยหลักการของแมกซิมั่มเอนโทรปี กล่าวคือ แบบจำลองที่ดีจะทำให้ค่าเอนโทรปีหรือค่าความไม่แน่นอนสูงที่สุด เมื่อพิจารณาภายใต้เงื่อนไขที่เป็นตัวแสดงถึงข้อเท็จจริงที่มีอยู่ ค่าเอนโทรปี H สามารถคำนวณได้จากสมการ

$$H(X) = -\sum p(x)\log p(x) \tag{2.1}$$

2.7.1 การแทนข้อเท็จจริง

หลักสำคัญในการทำงานของแบบจำลองแมกซิมั่มเอนโทรปีสำหรับการทำนายนั้น คือ การใช้ข้อเท็จจริงที่มีอยู่เพื่อกำหนดเงื่อนไขที่เป็นข้อบังคับให้แบบจำลอง ดังนั้น ผลลัพธ์การทำนายของแบบจำลองจะดีหรือไม่นั้น ขึ้นอยู่กับความสามารถในการได้มาซึ่งข้อเท็จจริงที่เกี่ยวข้องกับปัญหานั้นๆ โดยทั่วไปแล้ว การแทนข้อเท็จจริง หรือข้อมูลหลักฐาน เพื่อเป็นเงื่อนไขข้อบังคับที่แบบจำลองใช้ในการตัดสินใจนั้น สามารถทำได้โดยการใช้ฟังก์ชันเข้ารหัสข้อมูล ที่เรียกว่า Contextual Predicate หรือ Feature

ถ้า $A = \{a_1, \ldots, a_q\}$ เป็นเซตของคลาส ทั้งหมดที่เป็นไปได้ของผลลัพธ์ในการทำนาย และ B เป็นเซตของบริบทที่เป็นไปได้จากการสังเกต แล้ว Contextual Predicate แสดงเป็นฟังก์ชันได้ดังนี้

$$cp : B \rightarrow \{true, false\}$$

ค่า true หรือ false หมายถึงการมี หรือไม่มีข้อมูลบริบท $b \in B$ โดยเซตของ Contextual Predicate $\{cp_1 \ldots cp_m\}$ ที่เป็นประโยชน์จะมีได้หลากหลายขึ้นอยู่กับปัญหา สำหรับการแทนข้อเท็จจริงเพื่อเป็นเงื่อนไขบังคับแบบจำลองนั้น Contextual Predicate จะถูกใช้ในรูปแบบของ ฟังก์ชันคุณสมบัติ (Feature Function) ซึ่งเป็นไบนารีฟังก์ชัน (Binary Function) ที่ให้ค่าเป็น 0 หรือ 1

$$f : A \times B \rightarrow \{0,1\}$$

คุณสมบัติ (Feature) ใดๆ จะถูกแสดงได้ในรูปแบบฟังก์ชันคุณสมบัติดังนี้ คือ

$$f_{cp,a'}(a,b) = \begin{cases} 1 & \text{if } a = a' and\ cp(b) = true \\ 0 & \text{otherwise} \end{cases} \qquad (2.2)$$

โดยทั่วไป ฟังก์ชันคุณสมบัติจะแสดงให้เห็นถึงความสัมพันธ์ระหว่างบริบทกับผลการทำนาย ตัวอย่างเช่น หากต้องการประมาณค่าแบบจำลอง $p(a,b)$ โดยที่ a เป็น ประเภทของคู่คุณลักษณะสินค้าและความเห็น ซึ่งมี 2 ประเภท คือ เป็นคู่คุณลักษณะสินค้าและความเห็น และ ไม่เป็นคู่คุณลักษณะสินค้าและความเห็น และ b เป็นบริบท คือ Syntactic Relationship ระหว่างคำที่ต้องการทำนายว่าเป็นคู่คุณลักษณะสินค้าและความเห็นหรือไม่ โดยถ้าทั้งสองคำมีความสัมพันธ์กันในรูปแบบพ่อ (Parent Relationship) สามารถสร้างฟังก์ชันคุณสมบัติได้เป็น

$$f_{cp,a'}(a,b) = \begin{cases} 1 & \text{if } a = YES\ and\ Syn(PARENT) = true \\ 0 & \text{otherwise} \end{cases}$$

2.7.2 ตัวอย่างการใช้แมกซิมั่มเอนโทรปี

ในหัวข้อนี้จะแสดงตัวอย่างการใช้แมกซิมั่มเอนโทรปี กับปัญหาการสกัดคุณลักษณะสินค้าและความเห็นอย่างง่ายที่พิจารณาเพียงความสัมพันธ์ที่เกิดขึ้นระหว่างคำที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น คือ การทำนายว่าคู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็นที่มีความสัมพันธ์กันในลักษณะความสัมพันธ์แบบพ่อจะเป็นคุณลักษณะสินค้าและความเห็นที่แท้จริง กี่เปอร์เซ็นต์ โดยกำหนดให้ปริภูมิเหตุการณ์ (Event Space) คือ $\{1,0\} \times \{x_1, x_2, x_3, x_4, x_5, x_6\}$ ซึ่งนิยามได้ดังนี้

x_1 = ความสัมพันธ์แบบลูก (Child)

x_2 = ความสัมพันธ์แบบพ่อ (Parent)

x_3 = ความสัมพันธ์แบบพี่น้อง (Sibling)

x_4 = ความสัมพันธ์แบบหลาน (Grandchild)

x_5 = ความสัมพันธ์แบบปู่ย่า (Grandparent)

x_6 = ความสัมพันธ์แบบทางอ้อม (Indirect)

1 = เป็นคุณลักษณะสินค้าและความเห็นแท้จริง (Yes)

0 = ไม่เป็นคุณลักษณะสินค้าและความเห็นแท้จริง (No)

จากปัญหาการทำนายดังกล่าว เป็นการแจกแจงความน่าจะเป็นร่วม คือ $p(a,b)$ เมื่อ $a \in \{1,0\}$ และ $b \in \{x_1, x_2, x_3, x_4, x_5, x_6\}$ โดย x_1, x_2, x_3, x_4, x_5 และ x_6 เป็นข้อสังเกตที่ไม่เกิดร่วม (Mutually Exclusive Observations) และ 1 และ 0 เป็นผลลัพธ์ที่ไม่เกิดร่วม (Mutually Exclusive Outcome) ถ้าทราบข้อเท็จจริงว่า มีคู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็น 60% ที่เป็นคุณลักษณะสินค้าและความเห็นที่แท้จริง ข้อเท็จจริงนี้สามารถใช้เป็นเงื่อนไขบังคับได้ คือ

$$p(1,x_1) + p(1,x_2) + p(1,x_3) + p(1,x_4) + p(1,x_5) + p(1,x_6) = 0.6$$

เนื่องจากคู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็นทั้งหมด 100% จะเป็นคุณลักษณะสินค้าและความเห็นแท้จริง และไม่เป็นคุณลักษณะสินค้าและความเห็นที่แท้จริง ดังนั้น จึงได้ $\sum_{a,b} p(a,b) = 1$ นั่นคือ

$$p(1,x_1) + p(0,x_1) + p(1,x_2) + p(0,x_2) + p(1,x_3) + p(0,x_3) + p(1,x_4) + p(0,x_4) +$$
$$p(1,x_5) + p(0,x_5) + p(1,x_6) + p(0,x_6) = 1$$

การหาคำตอบว่า คู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็นที่มีความสัมพันธ์กันในลักษณะความสัมพันธ์แบบพ่อ กี่เปอร์เซ็นต์ที่จะเป็นคุณลักษณะสินค้าและความเห็นที่แท้จริง ทำได้โดยการคำนวณการแจกแจงความน่าจะเป็น p ที่เกิดภายใต้เงื่อนไขทั้ง 2 แสดงดังตารางข้างล่าง

$p(a,b)$	1	0	
x_1	?	?	
x_2	?	?	
x_3	?	?	
x_4	?	?	
x_5	?	?	
x_6	?	?	
รวม	0.6	0.4	1.0

จะเห็นว่าความน่าจะเป็นในแต่ละเซลล์สามารถเป็นไปได้มากมายหลายค่าที่ทำให้สมการทั้ง 2 เป็นจริง ตัวอย่างเช่น

$$p(1,x_1)=0.1,\ p(1,x_2)=0.2,\ p(1,x_3)=0.1,\ p(1,x_4)=0.1,\ p(1,x_5)=0.05,\ p(1,x_6)=0.05$$
$$p(0,x_1)=0.07,\ p(0,x_2)=0.03,\ p(0,x_3)=0.04,\ p(0,x_4)=0.06,\ p(0,x_5)=0.1,\ p(0,x_6)=0.1$$

หรือ

$$p(1,x_1)=0.2,\ p(1,x_2)=0.2,\ p(1,x_3)=0.05,\ p(1,x_4)=0.05,\ p(1,x_5)=0.05,\ p(1,x_6)=0.05$$
$$p(0,x_1)=0.04,\ p(0,x_2)=0.03,\ p(0,x_3)=0.1,\ p(0,x_4)=0.06,\ p(0,x_5)=0.1,\ p(0,x_6)=0.07$$

หรือ

$$p(1,x_1)=0.1,\ p(1,x_2)=0.1,\ p(1,x_3)=0.1,\ p(1,x_4)=0.1,\ p(1,x_5)=0.1,\ p(1,x_6)=0.1$$
$$p(0,x_1)=0.07,\ p(0,x_2)=0.03,\ p(0,x_3)=0.04,\ p(0,x_4)=0.06,\ p(0,x_5)=0.1,\ p(0,x_6)=0.1$$

อย่างไรก็ตาม จากหลักการของแมกซิมั่มเอนโทรปีที่ต้องการแบบจำลองที่มีการแจกแจงแบบสม่ำเสมอ โดยการกำหนดค่าความน่าจะเป็นที่จะเกิดเท่าๆกัน เมื่อพิจารณาเฉพาะจากข้อเท็จจริงที่มีอยู่ ทำให้ค่าความน่าจะเป็นในแต่ละเซลล์ เป็นดังนี้

$p(a,b)$	1	0	
x_1	0.1	0.0667	
x_2	0.1	0.0667	
x_3	0.1	0.0667	
x_4	0.1	0.0667	
x_5	0.1	0.0667	
x_6	0.1	0.0667	
รวม	0.6	0.4	1.0

ภายใต้กรอบการทำงานของหลักการแมกซิมั่มเอนโทรปี $p(1,x_1)+p(1,x_2)+p(1,x_3)+p(1,x_4)+p(1,x_5)+p(1,x_6)=0.6$ จะถูกพิจารณาให้เป็นเงื่อนไขบังคับ สำหรับค่าคาดหมาย (Expected Value) ของคุณสมบัติ f_1

$$E_p f_1 = 0.6 \tag{2.3}$$

เมื่อ $$E_p f_1 = \sum_{a\in\{x,y\},b\in\{1,0\}} p(a,b) f_1(a,b)$$

และ f_1 ถูกนิยามได้ดังนี้

$$f_1(a,b) = \begin{cases} 1 & \text{if } a=1 \\ 0 & \text{otherwise} \end{cases}$$

เช่นเดียวกับข้อเท็จจริง $p(1,x_1) + p(0,x_1) + p(1,x_2) + p(0,x_2) + p(1,x_3) + p(0,x_3) + p(1,x_4) + p(0,x_4) + p(1,x_5) + p(0,x_5) + p(1,x_6) + p(0,x_6) = 1$ จะถูกพิจารณาให้เป็นเงื่อนไขบังคับสำหรับค่าคาดคะเนของคุณสมบัติ f_2

$$E_p f_2 = 1.0 \tag{2.4}$$

เมื่อ $$E_p f_2 = \sum_{a\in\{x,y\},b\in\{1,0\}} p(a,b) f_2(a,b)$$

$$f_2(a,b) = 1.0$$

ค่าคาดหมายจากการสังเกตคุณสมบัติ f_1 หรือ $E_{\tilde{p}} f_1$ คือ 0.6 และค่าคาดหมายจากการสังเกตคุณสมบัติ f_2 หรือ $E_{\tilde{p}} f_2$ คือ 1.0 จากข้อกำหนดทั้ง 2 นี้ จะต้องหาแบบจำลองที่เป็นไปตามเงื่อนไข ในสมการที่ 2.3 และ 2.4 ที่ทำให้ค่าเอนโทรปี H(p) มีค่าสูงสุด

$$H(p) = -\sum_{a\in\{x,y\},b\in\{1,0\}} p(a,b)\log p(a,b)$$

จากตัวอย่างเมื่อค่าความน่าจะเป็นในแต่ละเซลล์ เป็นดังนี้

$p(1,x_1) = 0.1,\ p(1,x_2) = 0.1,\ p(1,x_3) = 0.1,\ p(1,x_4) = 0.1,\ p(1,x_5) = 0.1,\ p(1,x_6) = 0.1$

$p(0,x_1) = 0.0667,\ p(0,x_2) = 0.0667,\ p(0,x_3) = 0.0667,\ p(0,x_4) = 0.0667,\ p(0,x_5) = 0.0667$

$p(0,x_6) = 0.0667$

จะทำให้ค่าเอนโทรปี H(p) มีค่าสูงสุด คือ 2.4651 ดังนั้น การทำนายว่าคู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็นที่มีความสัมพันธ์กันในลักษณะความสัมพันธ์แบบพ่อจะเป็นคุณลักษณะสินค้าและความเห็นที่แท้จริง 10%

2.7.3 แบบจำลองแมกซิมั่มเอนโทรปีแบบมีเงื่อนไข

จากตัวอย่างการใช้แบบจำลองแมกซิมั่มเอนโทรปีในหัวข้อย่อยที่ผ่านมา เห็นได้ว่ามีการใช้เงื่อนไขข้อบังคับแค่ 2 ตัว แต่สำหรับปัญหาที่ซับซ้อนจำเป็นต้องใช้เงื่อนไขข้อบังคับที่มากขึ้น ในการศึกษาครั้งนี้ ถ้าสมมุติให้คุณสมบัติที่ใช้มี k ตัว ให้ $a \in A$ คือ ผลการทำนาย และ $b \in B$ คือบริบทที่สามารถสังเกตได้ เป้าหมาย คือ ต้องการหาค่าประมาณสำหรับความน่าจะเป็นแบบมีเงื่อนไข $p(a|b)$ โดยแบบจำลอง p^* ของแบบจำลองแมกซิมั่มเอนโทรปีแบบมีเงื่อนไขที่เหมาะสมที่สุด คือ แบบจำลองที่มีการแจกแจงค่าความน่าจะเป็นที่มีค่าความไม่แน่นอนสูงที่สุด โดยเป็นไปตามเงื่อนไขข้อบังคับทั้ง k ตัว บนค่าคาดหมายของคุณสมบัติทั้งหมด [19]

คุณสมบัติ k ตัว ที่เป็นเงื่อนไขบังคับ จะมีรูปแบบ ดังนี้

$$E_{\tilde{p}} f_j = E_p f_j \tag{2.5}$$

เมื่อ 1 <= j <= k และ $E_p f_j$ คือ ค่าคาดหมายของ f_j ของแบบจำลอง p

โดย

$$E_p f_j = \sum_{a,b} \tilde{p}(b)\, p(a \mid b) f_j(a,b) \tag{2.6}$$

โดยมีเงื่อนไขคือ ค่า $E_p f_j$ ต้องเท่ากับค่าคาดหมายจากการสังเกต (Observed Expectation) $E_{\tilde{p}} f_j$

โดย

$$E_{\tilde{p}} f_j = \sum_{a,b} \tilde{p}(a,b) f_j(a,b) \tag{2.7}$$

เมื่อ $\tilde{p}$ คือ ค่าความน่าจะเป็นจากการสังเกต หรือ Empirical Probability Distribution ของ (a,b) บนชุดตัวอย่างสำหรับการเรียนรู้ ซึ่งสามารถคำนวณได้ดังนี้

$$p(a,b) \equiv \frac{1}{N} * \text{จำนวนครั้งที่ } (a,b) \text{ เกิดขึ้นในชุดตัวอย่าง}$$

แบบจำลอง p จะถูกต้องตรงกับข้อเท็จจริงที่สังเกตได้ ก็ต่อเมื่อ แบบจำลอง p เป็นไปตามเงื่อนไขบังคับ k ข้อ ดังที่แสดงในสมการ 2.5 หลักการของแมกซิมั่มเอนโทรปี จะเลือกแบบจำลอง p^* จากเซต P ของแบบจำลองทั้งหมดที่เป็นไปตามเงื่อนไขบังคับ ที่ทำให้ค่าเอนโทรปี H(p) สูงสุด

$$P = \{p \mid E_p f_j = E_{\tilde{p}} f_j, j = \{1 \ldots k\}\}$$

$$p^* = \arg\max_{p \in P} H(p)$$

$$H(p) = -\sum_{a,b} \tilde{p}(b)\, p(a \mid b) \log p(a \mid b)$$

ให้ P เป็นเซตของแบบจำลองที่เป็นไปตามเงื่อนไขบังคับ จากหลักการของแมกซิมั่มเอนโทรปี ให้เลือกแบบจำลองที่ทำให้ค่าเอนโทรปี H(p) สูงสุด : $p^* = \arg\max_{p \in P} H(p)$

โดย p^* คำนวณได้ดังนี้

$$p_{\max}(a \mid b) = \frac{1}{Z(b)} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)} \tag{2.8}$$

โดย

$$Z(b) = \sum_{f} \prod_{j} \alpha_j^{f_j(a,b)} \tag{2.9}$$

เมื่อ $Z(b)$ คือ ค่าตัวประกอบการทำให้เป็นบรรทัดฐาน (Normalization Factor) เพื่อให้เป็นไปตามเงื่อนไข $\sum_a p(a \mid b) = 1$ และ α_j คือ ค่าพารามิเตอร์ของแบบจำลอง โดย $0 < \alpha_j < \infty$ แต่ละค่า α_j จะสัมพันธ์กับแต่ละฟังก์ชันคุณสมบัติ f_j ซึ่ง ค่า α_j นี้อาจมองเสมือนเป็นค่าถ่วงน้ำหนักของฟังก์ชันคุณสมบัตินั้น

2.7.4 การประมาณค่าพารามิเตอร์

สำหรับวิธีในการประมาณค่าพารามิเตอร์ของแบบจำลองแมกซิมั่มเอนโทรปีที่เป็นที่นิยมมี 2 วิธี คือ อัลกอริทึม Improved Iterative Scaling หรือ IIS และ อัลกอริทึม Generalized Iterative Scaling หรือ GIS สำหรับอัลกอริทึม IIS มักจะเกิดปัญหาการล้น (Overflow) มากกว่าอัลกอริทึม GIS นอกจากนี้ GIS ยังมีประสิทธิภาพมากกว่า IIS [20] การศึกษาในครั้งนี้ใช้อัลกอริทึม GIS [21] เป็นกระบวนการที่ทำหน้าที่ในการประมาณค่าพารามิเตอร์ $\{\alpha_1, \ldots, \alpha_k\}$ สำหรับ p^* [19]

โดยอัลกอริทึม GIS กำหนดเงื่อนไข เมื่อ $(a,b) \in A \times B$ คือ

$$\sum_{j=1}^{k} f_j(a,b) = C \tag{2.10}$$

แต่ถ้าหากไม่เป็นไปตามเงื่อนไขให้เลือกใช้ค่า C จากการใช้ชุดตัวอย่างสำหรับการเรียนรู้ ดังนี้คือ

$$C = \max_{a \in A, b \in \Gamma} \sum_{j=1}^{k} f_j(a,b)$$

และให้มีการเพิ่ม Correction Feature คือ f_l เมื่อ l = k+1 ที่ทำให้

$$f_j(a,b) = C - \sum_{j=1}^{k} f_j(a,b)$$

สำหรับคู่ของ (a,b) ใดๆ f_l จะแตกต่างจากฟังก์ชันคุณสมบัติอื่น โดยมีค่าได้ตั้งแต่ 0 ถึง C เมื่อ C สามารถมีค่ามากกว่า 1 ได้ นอกจากนี้ GIS ยังได้กำหนดสมมุติฐานไว้ว่า สำหรับทุกเหตุการณ์จะต้องมีอย่างน้อย 1 ฟังก์ชันคุณสมบัติที่เป็นจริง

$$\exists f_j : f_j(a,b) = 1$$

ขั้นตอนการคำนวณ และปรับค่า α_j ของอัลกอริทึม GIS มีดังนี้

$$\alpha_j^{(0)} = 1$$

$$\alpha_j^{(n+1)} = \alpha_j^n \left[\frac{\tilde{E} f_j}{E^{(n)} f_j} \right]^{\frac{1}{C}}$$

$$E_{p^{(n)}} f_j = \sum_{a,b} \tilde{p}(b) p^{(n)}(a \mid b) f_j(a,b)$$

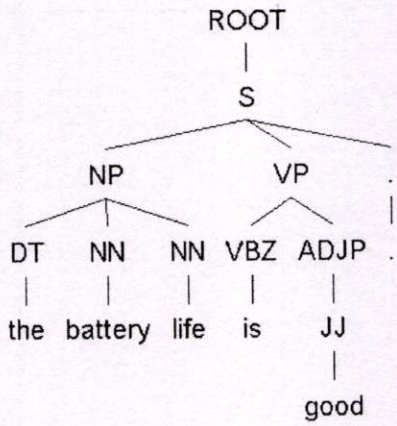
$$p^{(n)}(a,b) = \frac{1}{Z(b)} \prod_{j=1}^{l} (\alpha_j^{(n)})^{f_j(a,b)}$$

กระบวนการประมาณค่าพารามิเตอร์นี้ จะสิ้นสุดเมื่อจำนวนรอบของการวนซ้ำครบตามที่กำหนดไว้ หรือเมื่อการเปลี่ยนแปลงของค่า Log-likelihood ต่ำมาก คือ มีการลู่เข้าหรือใกล้เคียงกับการลู่เข้ามากที่สุด ในการศึกษาครั้งนี้ใช้จำนวน 100 รอบสำหรับสิ้นสุดการประมาณค่าพารามิเตอร์

2.8 การแจงส่วน

การแจงส่วน หรือ พาสซิง (Parsing) คือ การแจกแจงส่วนประกอบของโครงสร้างประโยค โดยใช้กฎไวยากรณ์ในการวิเคราะห์ประโยค และกำหนดโครงสร้างประโยคตามที่ไวยากรณ์ระบุ การวิเคราะห์โครงสร้างประโยค จะเป็นการวิเคราะห์หาความสัมพันธ์ของส่วนประกอบในประโยค ว่าคำใดทำหน้าที่เป็นประธานและกรรมของคำกริยา และหาความสัมพันธ์ของคำขยายและคำที่ถูกขยาย โดยทั่วไป โครงสร้างของประโยคจะแทนด้วยรูปต้นไม้ที่เป็นลำดับชั้น ดังนั้น การแจงส่วนจึงเป็นกระบวนการที่ค้นหาโครงสร้างของประโยคจากกฎวิเคราะห์ประโยค และโปรแกรมที่ใช้

สำหรับการแจงส่วนประโยคถูกเรียกว่า พาสเซอร์ (Parser) กรอบการทำงานของโปรแกรมนี้จะใช้ไวยากรณ์ไม่พึ่งบริบท (Context-Free Grammar) ผลลัพธ์ที่ได้จากการแจงส่วนของประโยคจะแสดงในรูปของต้นไม้แจงส่วน เพื่อแสดงความสัมพันธ์ของคำแต่ละคำในประโยค เช่น ตัวใดเป็นประธานหรือกรรมของกริยา คำใดทำหน้าที่ขยายคำอื่น เป็นต้น ตัวอย่างเช่น การแจงส่วนของประโยค "*The battery life is good.*" โดยใช้โปรแกรมสแตนฟอร์ดพาสเซอร์ (Stanford Parser) [22] ผลลัพธ์ที่ได้ แสดงในรูปที่ 2.10



รูปที่ 2.10 ผลลัพธ์ที่ได้จากการแจงส่วนของประโยค "*The battery life is good.*"

จากรูปที่ 2.10 ตัวบนสุดในรูปเขียนแทนด้วยสัญลักษณ์ S แสดงถึงประโยค (Sentence) ต้นไม้แจงส่วนนี้แสดงโครงสร้างของประโยค โดยแบ่งเป็นสองส่วน คือ (1) นามวลี ซึ่งแทนด้วย NP ประกอบด้วยคำนำหน้านาม แทนด้วย DT และต่อด้วยคำนาม แทนด้วย NN และ (2) กริยาวลี ซึ่งแทนด้วย VP ประกอบด้วยคำกริยา และคำคุณศัพท์ โดยทั่วไปเทคนิคของการแจงส่วน ถือได้ว่าเป็นรูปแบบการค้นหารูปแบบหนึ่ง โดยการค้นหาโครงสร้างประโยคที่เป็นไปตามกฎไวยากรณ์ ซึ่งมีวิธีหลักๆ อยู่ 2 วิธีคือ [23]

2.8.1 การแจงส่วนจากล่างขึ้นบน (Bottom-up parsing) เป็นการค้นหาโดยเริ่มจากคำศัพท์ซึ่งถือเป็นสัญลักษณ์สิ้นสุด ถูกแทนที่ด้วยชนิดของคำ เช่น Jane ถูกแทนด้วย NAME หรือ ate ถูกแทนด้วย V โดยคำศัพท์ทุกคำจะถูกแทนที่ด้วยชนิดของคำจนหมด หลังจากนั้นใช้สัญลักษณ์ที่อยู่ฝั่งซ้ายของกฎเขียนแทนกลุ่มชนิดของคำ เช่น NP แทนที่ NAME หรือ NP แทนที่กลุ่มชนิดของคำ ART N ไปเรื่อยๆจนกระทั่งพบสัญลักษณ์ S ดังแสดงในตัวอย่างการแจงส่วนประโยค "*Jane ate a hamburger.*"

Jane ate a hamburger

➡ NAME ate a hamburger (rewriting Jane)

➡ NAME V a hamburger (rewriting ate)

➡ NAME V ART hamburger (rewriting a)

➡ NAME V ART N (rewriting hamburger)

➡ NP V ART N (rewriting NAME)

➡ NP V NP (rewriting ART N)

➡ NP VP (rewriting V NP)

➡ S (rewriting NP VP)

2.8.2 การแจงส่วนจากบนลงล่าง (Top-down parsing) เป็นการค้นหาโดยเริ่มจากสัญลักษณ์เริ่มต้น S แล้วเขียนใหม่ด้วยสัญลักษณ์ที่อยู่ฝั่งขวาของกฎ เช่น สัญลักษณ์ NP VP โดยสัญลักษณ์ NP VP สามารถเขียนแทนได้ใหม่เป็น NAME VP ทำจนกระทั่งสัญลักษณ์ไม่สิ้นสุดทุกตัวถูกแทนที่ด้วยสัญลักษณ์สิ้นสุดหรือคำศัพท์ ดังแสดงในตัวอย่างการแจงส่วนประโยค "*Jane ate a hamburger.*"

S

➡ NP VP (rewriting S)

➡ NAME VP (rewriting NP)

➡ Jane VP (rewriting NAME)

➡ Jane V NP (rewriting VP)

➡ Jane ate NP (rewriting V)

➡ Jane ate ART N (rewriting NP)

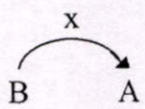
➡ Jane ate a N (rewriting ART)

➡ Jane ate a hamburger (rewriting N)

2.9 ไวยากรณ์พึ่งพา

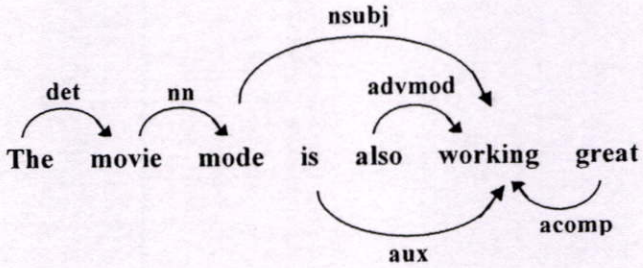
ทฤษฎีภาษาศาสตร์จำนวนมากถูกสร้างโดยนักภาษาศาสตร์ที่มีมุมมองในเรื่องโครงสร้างภาษาเป็นแบบหน่วยประกอบ (Constituent) ซึ่งเป็นแนวคิดพื้นฐานที่มีมาตั้งแต่สมัยไวยากรณ์โครงสร้าง แต่มีนักภาษาศาสตร์อีกกลุ่มหนึ่งที่มีมุมมองในเรื่องโครงสร้างภาษาเป็นแบบหน่วยพึ่งพา (Dependency) ในลักษณะของเครือข่ายความสัมพันธ์แบบพึ่งพาระหว่างคำที่เป็นคำหลัก และคำพึ่งพา ความสัมพันธ์แบบพึ่งพาเป็นความสัมพันธ์แบบทวิภาคอสมมาตร (Asymmetric Binary

Relationship) ระหว่างคำสองคำ ที่เรียกว่า หน่วยหลัก (Head หรือ Governor) และหน่วยพึ่งพา (Modifier หรือ Dependent) [24]



A เป็นหน่วยหลัก
B เป็นหน่วยพึ่งพาของ A
A มีความสัมพันธ์ "x" กับ B

คำแต่ละคำสามารถมีหน่วยพึ่งพาได้มากกว่าหนึ่ง เช่น *big black dog* ทั้ง *big* และ *black* เป็นหน่วยพึ่งพาของ *dog* โดยทั่วไป ทฤษฎีการพึ่งพาจะถือว่าคำแต่ละคำในประโยคจะมีหน่วยหลักได้เพียงหน่วยเดียว กล่าวโดยสรุปว่า ความสัมพันธ์ระหว่างคำหลักและคำพึ่งพาในกรณีปกติมีลักษณะดังนี้ คือ แต่ละคำจะเป็นคำหลักหรือไม่ก็เป็นคำพึ่งพาของอีกคำ โดยแต่ละคำจะมีคำหลักเพียงหนึ่งคำ และคำหลักสามารถมีคำพึ่งพาได้มากกว่าหนึ่งคำ ดังตัวอย่างรูปที่ 2.11 ซึ่งแสดงความสัมพันธ์แบบพึ่งพาระหว่างคำในประโยค "*The movie mode is also working great.*" จากรูปที่ 2.11 คำที่เป็นคำพึ่งพาสามารถเป็นคำหลักของคำอื่นๆได้ จึงเกิดเป็นสายโซ่การพึ่งพา (Chain of Dependency) โดยมีความสัมพันธ์แบบต่างๆ เช่น ประธาน, กริยาช่วย, คำนำหน้านาม, ส่วนเติมเต็ม และส่วนขยาย



รูปที่ 2.11 ความสัมพันธ์แบบพึ่งพาระหว่างคำในประโยค "*The movie mode is also working great.*"

สำหรับงานวิจัยนี้ใช้โปรแกรมสแตนฟอร์ดพาสเซอร์ในการแจงส่วนประโยคเพื่อหาความสัมพันธ์แบบพึ่งพา ผลลัพธ์ที่ได้จากการแจงส่วนประโยค "*The movie mode is also working great.*" คือ

det(mode-3, The-1)

nn(mode-3, movie-2)

nsubj(working-6, mode-3)

aux(working-6, is-4)

advmod(working-6, also-5)

acomp(working-6, great-7)

ผลลัพธ์ที่ได้จะแสดงในรูปแบบ *abbreviated_relation_name(head, dependent)* ซึ่งมีส่วนประกอบ 3 ส่วน ได้แก่

1. ชื่อย่อของความสัมพันธ์ (Abbreviated_relation_name) หมายถึงชื่อย่อของความสัมพันธ์ระหว่างหน่วยหลักและหน่วยพึ่งพา เช่น คำ "*is*" และคำ "*working*" ในประโยคตัวอย่างในรูปที่ 2.8 จะมีชื่อย่อของความสัมพันธ์คือ *aux* ซึ่งหมายถึงกริยาช่วย สามารถดูรายละเอียดได้ในภาคผนวก ค.

2. หน่วยหลัก (Head) หมายถึงคำที่อยู่ด้านหัวลูกศร เช่นความสัมพันธ์แบบพึ่งพาระหว่าง คำ "*is*" และคำ "*working*" ในประโยคตัวอย่างในรูปที่ 2.8 คำ "*working*" คือ หน่วยหลักของคำ "*is*"

3. หน่วยพึ่งพา (Dependent) หมายถึงคำที่อยู่ด้านปลายของลูกศร เช่น ความสัมพันธ์แบบพึ่งพาระหว่าง คำ "*is*" และคำ "*working*" ในประโยคตัวอย่างในรูปที่ 2.8 คำ "*is*" คือ หน่วยพึ่งพาของคำ "*working*"

2.10 ออนโทโลยี

ในปัจจุบันออนโทโลยี หรือภววิทยา (Ontology) ได้ถูกนำไปใช้ในงานวิจัยหลายๆด้าน เช่น ด้านปัญญาประดิษฐ์ วิศวกรรมความรู้ และงานด้านการประมวลผลภาษาธรรมชาติ ได้มีผู้ให้คำนิยามความหมายของออนโทโลยีไว้หลายความหมาย ดังต่อไปนี้

นักปราชญ์อริสโตเติล ได้กำหนดวิชาออนโทโลยี ซึ่งมีรากศัพท์จาก onto รวมกับ logy โดย onto หมายถึงสิ่งที่มีอยู่ และ logy หมายถึงศาสตร์ เมื่อรวมกันจะมีความหมายว่า เป็นศาสตร์ที่กล่าวถึงสิ่งที่มีอยู่ ส่วนความหมายอื่นของออนโทโลยีจะหมายถึง นิยามที่เป็นทางการและมีการประกาศอย่างชัดแจ้งของคำศัพท์หรือแนวคิดที่ใช้ร่วมกัน ประกอบไปด้วยเซตของแนวคิดที่มีการกำหนดความหมาย และความสัมพันธ์ของแนวคิด ตัวอย่างออนโทโลยีเกี่ยวกับกีฬา แสดงดังรูปที่ 2.12 [25]

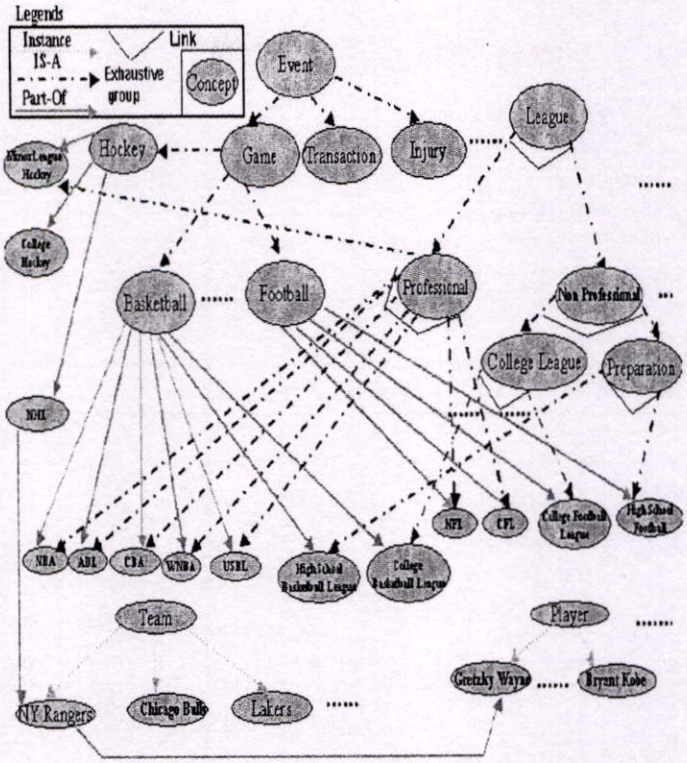
หลักการของออนโทโลยีจะมีลักษณะคล้ายกับหลักการเชิงวัตถุ การใช้ออนโทโลยีมีประโยชน์ทำให้เข้าใจความหมายที่ตรงกันระหว่างมนุษย์กับคอมพิวเตอร์ และสามารถนำกลับมาใช้

งานซ้ำได้อีก ดังนั้น ในปัจจุบันจึงมีการนำไปประยุกต์ใช้งานในงานด้านต่างๆ ได้แก่ การค้นคืนสารสนเทศ การจำแนกประเภทข้อความ และการแปลภาษา เป็นต้น นอกจากนี้ ยังมีการนำออนโทโลยีมาช่วยในการแก้ปัญหาต่างๆ ได้แก่ ปัญหาด้านการสรุปข้อความ โดย Wu และ Liu [26] ได้นำออนโทโลยีมาช่วยในการสรุปข้อความสำหรับข่าวทางธุรกิจ ออนโทโลยีที่สร้างขึ้นแบบขึ้นอยู่กับขอบเขตที่เกี่ยวกับบริษัทโซนี และนำเสนอออนโทโลยีในโครงสร้างต้นไม้ โดยใช้ออนโทโลยีในการพิจารณาหัวข้อย่อยที่เป็นหลักในการนำไปใช้สรุปข้อความ ผลการทดลองพบว่า การใช้ออนโทโลยีทำให้การหาหัวข้อย่อยในการสรุปมีความถูกต้องมากขึ้น และไม่ขึ้นอยู่กับความถี่ของการพบหัวข้อย่อยนั้น ปัญหาด้านการดึงข้อมูลจากเอกสาร Boufaden [27] ได้นำออนโทโลยีมาช่วยในการทำป้ายความหมาย (Semantic Tagger) ของคำเพื่อทำให้การดึงข้อมูลนั้นมีความทนทานและถูกต้องมากขึ้น ปัญหาในการค้นหาข้อมูล Gao et al [28] ได้นำออนโทโลยีเข้ามาช่วยในการค้นสารสนเทศของโปรแกรมเสิร์จเอ็นจินสำหรับวิเคราะห์ความหมายของคำหลักและข้อมูลที่ค้นได้ทำให้ได้ผลลัพธ์ที่ตรงกับความต้องการมากขึ้น นอกจากนี้แล้ว Cheng et al [29] ได้นำออนโทโลยีมาใช้ในการจำแนกประเภทข้อความ เพื่อแก้ปัญหาการใช้คำที่แตกต่างกันแต่หมายถึงเรื่องเดียวกันทำให้การจำแนกประเภทข้อความมีความถูกต้องมากขึ้น

โดยทั่วไป ออนโทโลยีแบ่งออกเป็น 2 ประเภท คือ ออนโทโลยีทั่วไป (Generic Ontology) ที่สร้างขึ้นมาใช้งานทั่วไป ไม่ได้เฉพาะเจาะจงกับปัญหาใดปัญหาหนึ่ง เช่น CYC, WordNet และ Sensus เป็นต้น ออนโทโลยีแบบนี้มักมีขนาดใหญ่ ขาดรายละเอียด และสร้างได้ยาก ออนโทโลยีอีกประเภท คือ ออนโทโลยีแบบขึ้นอยู่กับโดเมน (Domain Dependent Ontology) ออนโทโลยีประเภทนี้จะสร้างขึ้นมาเพื่อใช้เฉพาะเจาะจงกับปัญหาใดปัญหาหนึ่ง ดังนั้น จะมีความละเอียดมากกว่าออนโทโลยีทั่วไป ขั้นตอนในการออกแบบและสร้างออนโทโลยียังไม่มีมาตรฐาน แต่พอสรุปได้ดังนี้ [26]

1. กำหนดวัตถุประสงค์ในการใช้ออนโทโลยี
2. กำหนดขอบเขตของออนโทโลยี
3. สร้างออนโทโลยี

สำหรับการศึกษาครั้งนี้ ใช้ออนโทโลยีแบบขึ้นอยู่กับโดเมนเพื่อช่วยแก้ปัญหาในกรณีที่มีการใช้คำที่แตกต่างกันในการแสดงถึงคุณลักษณะสินค้า



รูปที่ 2.12 ตัวอย่างออนโทโลยีเกี่ยวกับกีฬา

2.11 วิธีการสกัดคุณลักษณะสินค้าและความเห็น

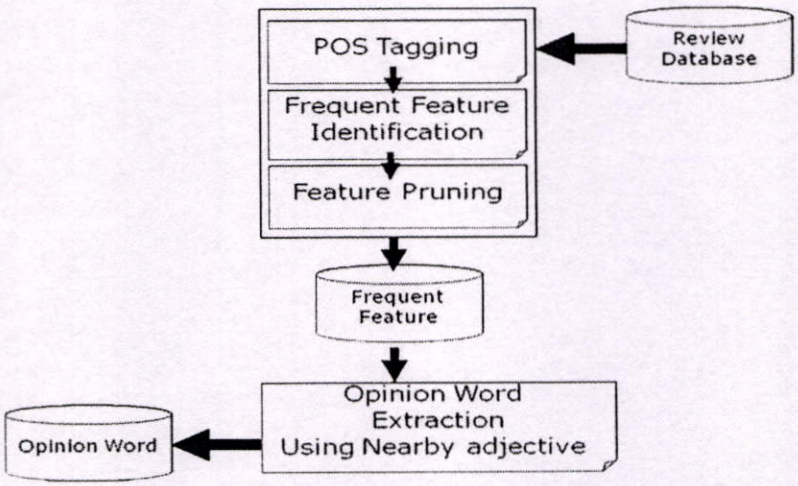
การสรุปความเห็นมีความแตกต่างจากการสรุปข้อความทั่วไป คือ การสรุปข้อความมีวัตถุประสงค์เพื่อสรุปข้อความให้สั้น กระชับ และได้ใจความหลักที่สำคัญจากข้อความเดิมที่มีขนาดยาว แต่การสรุปความเห็นมีวัตถุประสงค์เพื่อสรุปความเห็นตามหัวข้อและด้านที่มีการแสดงความเห็นจากข้อมูลการวิจารณ์ของลูกค้าที่มีจำนวนมาก งานวิจัยเกี่ยวกับการสรุปความเห็นส่วนใหญ่จะทำการศึกษากับข้อมูลการวิจารณ์สินค้า โดยจะทำการสรุปความเห็นตามคุณลักษณะของสินค้า พร้อมทั้งมีการระบุด้านความเห็นที่มีต่อคุณลักษณะสินค้านั้นด้วย เมื่อพิจารณาปัญหาในการสรุปความเห็นแล้ว พบว่า ปัญหาหลักที่สำคัญคือ ปัญหาของการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้า เนื่องจากการสกัดความเห็นที่มีต่อคุณลักษณะสินค้าในการวิจารณ์สินค้านั้น จะส่งผลต่อขั้นตอนการระบุด้านความเห็นนั้นๆด้วย ดังนั้น การสรุปความเห็นจะมีประสิทธิภาพดีเพียงใดนั้น จะขึ้นอยู่กับความถูกต้องหรือประสิทธิภาพในการสกัดคุณลักษณะสินค้าและความเห็น

2.11.1 วิธีการสกัดคุณลักษณะสินค้าและความเห็นของงานวิจัยที่ผ่านมา

งานวิจัยที่ผ่านมาวิธีการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้าจะเริ่มจากการสกัดคุณลักษณะสินค้า ซึ่งส่วนใหญ่จะใช้วิธีการทางด้านสถิติ แล้วค้นหาความเห็นที่สัมพันธ์กับคุณลักษณะสินค้าที่สกัดได้ วิธีการที่ใช้ในงานวิจัยที่ผ่านมาสามารถสรุปได้ดังนี้

2.11.1.1 วิธีการใช้คำคุณศัพท์ใกล้กัน (Nearby Adjective Based Approach)

งานวิจัยของ Hu and Liu [2] ถือเป็นงานวิจัยแรกที่บุกเบิกงานวิจัยด้านการสรุปความเห็นจากข้อมูลการวิจารณ์สินค้า วิธีการสกัดคุณลักษณะสินค้าและความเห็นของงานวิจัยนี้แสดงได้ดังรูปที่ 2.13



รูปที่ 2.13 วิธีการสกัดคุณลักษณะสินค้าและความเห็นในงานวิจัยของ Hu and Liu

วิธีการสกัดคุณลักษณะสินค้าและความเห็นของวิธีการใช้คำคุณศัพท์ใกล้กัน จะเริ่มจากการกำกับหมวดคำ และทำการสกัดคุณลักษณะสินค้าซึ่งใช้หลักการของการปรากฏของคำ ถ้าคำใดมีความถี่ในการปรากฏสูงจะถือว่าเป็นคำแสดงคุณลักษณะสินค้า (Frequent Feature) ที่ลูกค้าส่วนใหญ่แสดงความเห็น ซึ่งจะพิจารณาจากคำนามและนามวลี ดังนั้นวิธีการสกัดคุณลักษณะสินค้าของงานวิจัยนี้ใช้ Association Rule Mining ในการหาคุณลักษณะสินค้า โดยจะพิจารณาจากความถี่ในการปรากฏของคำถ้ามีค่าสูงกว่าค่า Minimum Support ที่กำหนดไว้ คือ 1% จะถือว่าคำนั้นเป็นคุณลักษณะสินค้า แล้วทำการตัดคุณลักษณะสินค้าที่ไม่มีความหมายและที่ซ้ำซ้อนทิ้ง หลังจากสกัดคุณลักษณะสินค้าได้แล้ว จะนำคุณลักษณะสินค้าที่สกัดได้ไปสกัดความเห็นที่สัมพันธ์กับคุณลักษณะสินค้านั้นๆ โดยใช้การพิจารณาคำคุณศัพท์ใกล้กัน ซึ่งมีหลักการว่า คำที่แสดงความเห็นจะอยู่ใกล้กับคุณลักษณะสินค้าที่แสดงความเห็นนั้น วิธีการนี้จะสกัดคำคุณศัพท์ที่อยู่ใกล้กันกับคำนามหรือนามวลีที่เป็นคุณลักษณะสินค้า ซึ่งคำคุณศัพท์จะทำหน้าที่ขยายคำที่เป็นคุณลักษณะสินค้านั้น รายละเอียดวิธีการสกัดความเห็นสามารถแสดงได้ดังนี้

FOR each sentence in the review database

IF (it contains a frequent feature, extract all the adjective words as opinion words)

FOR each feature in the sentence

the nearby adjective is recorded as its *effective opinion*.

/* A nearby adjective refers to the adjacent adjective that

modifies the noun/noun phrase that is a frequent feature. */

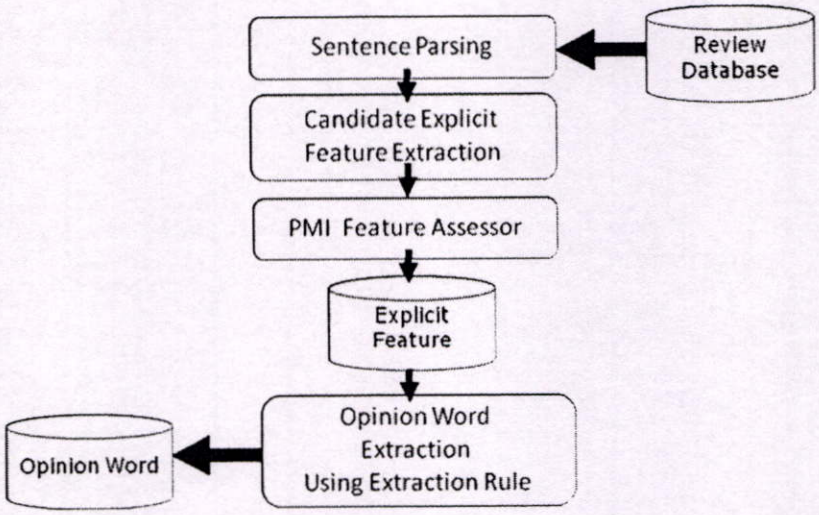
ตัวอย่างการสกัดคุณลักษณะสินค้าและความเห็นจากประโยคการวิจารณ์กล้องดิจิทัล

1. "*The strap is **horrible** and gets in the way of parts of the camera you need access to.*"
2. "*After nearly 800 pictures I have found that this camera takes **incredible** pictures*"

การสกัดคุณลักษณะสินค้าและความเห็นจากประโยคตัวอย่างทั้งสอง จะเริ่มจากการสกัดคุณลักษณะสินค้าก่อน ถ้าสมมุติว่า "*strap*" และ "*pictures*" มีความถี่ในการปรากฏสูงกว่าค่า Minimum Support จะถือว่าสองคำนี้เป็นคุณลักษณะสินค้า หลังจากนั้นจะสกัดความเห็นโดยพิจารณาคำคุณศัพท์ที่ทำหน้าที่ขยาย "*strap*" และ "*pictures*" ซึ่งพบว่า ในประโยคที่ 1 "*horrible*" เป็นคำคุณศัพท์ที่ขยาย "*strap*" และประโยคที่ 2 "*incredible*" เป็นคำคุณศัพท์ที่ขยาย "*pictures*" ดังนั้น จะถือว่า "*horrible*" และ "*incredible*" เป็นคำแสดงความเห็นที่มีต่อ "*strap*" และ "*pictures*" ตามลำดับ

2.11.1.2 วิธีการใช้กฎในการสกัด (Extraction Rule Based Approach)

Popescu [4] ได้นำเสนอระบบ OPINE ในการสกัดข้อมูลแบบไม่มีการสอนในการวิเคราะห์ความเห็น วิธีการสกัดคุณลักษณะและความเห็นของงานวิจัยนี้แสดงได้ดังรูปที่ 2.14



รูปที่ 2.14 วิธีการสกัดคุณลักษณะสินค้าและความเห็นในงานวิจัยของ Popescu

วิธีการสกัดคุณลักษณะและความเห็นของวิธีการใช้กฎในการสกัด จะเริ่มจากการแจงส่วนประโยค แล้วดึงคำที่เป็นนามวลีมาประเมินว่าเป็นคุณลักษณะของสินค้าหรือไม่ ในส่วนของการประเมินจะใช้ PMI Feature Assessor โดยการคำนวณค่า PMI (Point-Wise Mutual Information) ระหว่างนามวลีและคำที่มีความสัมพันธ์ในลักษณะของ Meronymy Discriminator กับกลุ่มสินค้า เช่น *of scanner*, *scanner has*, *scanner comes* ซึ่งเป็นคำที่มีความสัมพันธ์ในลักษณะของ Meronymy Discriminator สำหรับเครื่องสแกนเนอร์ หลังจากสกัดคุณลักษณะสินค้าได้แล้ว จะนำคุณลักษณะสินค้าที่สกัดได้ไปสกัดความเห็นที่สัมพันธ์กับคุณลักษณะสินค้านั้นๆ แนวทางในการสกัดความเห็นจะใช้รูปแบบการเกิดขึ้นร่วมกันระหว่างคุณลักษณะและคำแสดงความเห็น ซึ่งมีหลักการคล้ายกับวิธีการของ Hu and Liu ที่ว่า คำแสดงความเห็นควรอยู่ใกล้กับคำที่เป็นคุณลักษณะสินค้าแต่ใช้โครงสร้างทางไวยากรณ์ที่เกิดขึ้นในประโยคนั้น สร้างเป็นกฎสำหรับสกัดคำแสดงความเห็นแทนการสกัดจากคำที่อยู่ใกล้คำที่เป็นคุณลักษณะสินค้านั้น สำหรับกฎที่ใช้สกัดความเห็นในงานวิจัยของ Popescu แสดงได้ดังตารางที่ 2.2

ตารางที่ 2.2 กฎที่ใช้สกัดความเห็นในงานวิจัยของ Popescu

กฎที่ใช้สกัดความเห็น	ตัวอย่างในการสกัดความเห็น
if $\exists(M, NP = f) \rightarrow po = M$	(expensive) scanner
if $\exists(S = f, P, O) \rightarrow po = O$	Lamp has (problems)
if $\exists(S, P, O = f) \rightarrow po = P$	I (hate) this scanner
if $\exists(S = f, P) \rightarrow po = P$	Program (crashed)

เมื่อ po = potential opinion, M =modifier, NP = noun phrase, S = subject, P = predicate, O = object และ f = feature

ตัวอย่างการสกัดคุณลักษณะสินค้าและความเห็นจากประโยคการวิจารณ์กล้องดิจิทัล "*This camera takes* ***excellent*** *pictures*" การสกัดคุณลักษณะสินค้าและความเห็นจากประโยคตัวอย่างจะเริ่มจากการสกัดคุณลักษณะสินค้าก่อน ถ้าสมมุติว่า "*pictures*" ถูกประเมินว่าเป็นคุณลักษณะสินค้า หลังจากนั้นจะทำการสกัดความเห็นโดยใช้กฎ ดังนั้น จากประโยคตัวอย่างสามารถสกัดความเห็นที่สัมพันธ์กับคุณลักษณะสินค้า "*pictures*" ได้เป็น *"excellent"* เมื่อพิจารณาจากกฎ if $\exists(M, NP = f) \rightarrow po = M$ ที่ว่าถ้านามวลีเป็นคุณลักษณะสินค้าแล้วคำที่ทำหน้าที่ขยายนามวลีจะถือว่าเป็นความเห็น

2.11.2 วิธีการสกัดคุณลักษณะสินค้าของงานวิจัยที่ผ่านมา

โดยทั่วไปการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้าจะเริ่มจากการสกัดคุณลักษณะสินค้าก่อน ดังนั้น งานวิจัยบางส่วนจะเน้นการแก้ปัญหาเฉพาะเรื่องการสกัด

คุณลักษณะสินค้า ซึ่งสามารถสรุปได้ดังนี้

2.11.2.1 Beginning Definite Base Noun Phrases Heuristic Approach

Yi and Niblack [5] ได้นำเสนอวิธีการสกัดคุณลักษณะสินค้าโดยใช้วิธีการทางด้านการประมวลผลภาษาธรรมชาติและการคำนวณทางด้านสถิติ เริ่มจากการใช้นามวลีที่มีลักษณะเป็น bBNP (Beginning Definite Base Noun Phrases) ที่เริ่มต้นประโยคและมีกริยาวลีตามหลัง นามวลีจะมีรูปแบบดังนี้ NN, NN NN, JJ NN, NN NN NN, JJ NN NN และ JJ JJ NN เมื่อ NN แทนคำนาม และ JJ แทนคำคุณศัพท์ แล้วทำการพิจารณานามวลีแต่ละตัวว่าเป็นหัวข้อหรือคุณลักษณะสินค้าที่แสดงความเห็นหรือไม่ จากค่าอัตราส่วนควรจะเป็น (Likelihood Ratio) ที่คำนวณจากสมการ $-2\log\lambda$ ว่ามีค่ามากกว่าค่าระดับความเชื่อมั่นที่กำหนดไว้หรือไม่ การทดลองใช้ข้อมูลการวิจารณ์กล้องดิจิทัลและข้อมูลการวิจารณ์เพลง ผลการทดลองพบว่ามีความถูกต้องมากกว่าระบบ ReviewSeer [30] ที่ใช้วิธีทางสถิติในการสกัดความเห็น

2.11.2.2 Language Pattern Matching Approach

Lui et al [3] ได้ทำการวิจัยเรื่อง Opinion Observer: Analyzing and Comparing Opinions on the Web นำเสนอวิธีการสกัดคุณลักษณะสินค้า โดยใช้เทคนิคการทำเมืองข้อมูลรูปแบบภาษา (Language Pattern Mining) ซึ่งเป็นวิธีแบบมีการสอนในการดึงคุณลักษณะที่ถือว่าเป็นคุณลักษณะสินค้าที่แสดงความเห็น ซึ่งรูปแบบภาษาที่ใช้จะสร้างจากกฎที่ได้จากการทำไมนิ่งหากฎความสัมพันธ์ ที่ใช้ค่าสนับสนุน 1% และค่าความเชื่อมั่น 50% เช่น กฎ <N1>,<N2> → [feature] ซึ่งนำมาสร้างเป็นรูปแบบภาษาได้เป็น <N1>[feature]<N2> ส่วนคำที่ถือว่าเป็นคุณลักษณะจะใช้ทั้ง คำนาม นามวลี คำกริยา และคำคุณศัพท์ หลังจากสกัดคุณลักษณะสินค้าได้แล้วจะทำการจัดกลุ่มคำคุณลักษณะของสินค้าที่มีความหมายเหมือนกันเข้าด้วยกันโดยใช้ WordNet ผลการทดลองพบว่าการสกัดคุณลักษณะสินค้าจากการวิจารณ์สินค้าด้านบวก ให้ค่าความเที่ยงตรงสูงกว่าวิธีการของ Hu and Liu [2] แต่การสกัดคุณลักษณะสินค้าจากการวิจารณ์สินค้าด้านลบให้ค่าที่ต่ำกว่า เพราะว่าคำแสดงคุณลักษณะสินค้าในการวิจารณ์สินค้าด้านลบไม่ค่อยมีรูปแบบที่ชัดเจนเหมือนการวิจารณ์สินค้าด้านบวก

2.11.2.3 Probability-based Approach

Scaffidi et al [8] ได้นำเสนอวิธีการสกัดคุณลักษณะสินค้าที่แตกต่างจากงานวิจัยที่ผ่านมา โดยมีสมมุติฐานว่าคำที่จะเป็นคุณลักษณะสินค้าจะถูกอ้างถึงในการวิจารณ์สินค้ามากว่าที่จะถูกอ้างถึงในการเขียนทั่วๆไป ดังนั้น จึงใช้ค่าความถี่ของการปรากฏของคำที่คาดว่าจะเป็นคุณลักษณะสินค้าในฐานข้อมูลการวิจารณ์สินค้าและค่าความน่าจะเป็นที่เกิดขึ้นในการเขียนทั่วไป โดยใช้ความน่าจะเป็นที่แจกแจงแบบไบโนเมียล สมมุติว่า x เป็นคำนามซึ่งคาดว่าจะเป็นคุณลักษณะสินค้า และถ้า $n_x > p_x N$ และ $\ln(P(n_x))$ มีค่าน้อยมาก ระบบจะพิจารณาว่า x เป็นคุณลักษณะสินค้า ตัวอย่างเช่น สมมุติว่าข้อมูลของการวิจารณ์สินค้ามีจำนวนทั้งหมด

1,690 ประโยค มีจำนวนคำนามทั้งหมด N = 58,543 และกำหนดให้ x = "lens" ซึ่งเป็นคำนามที่คาดว่าจะเป็นคุณลักษณะสินค้าและเกิดขึ้นจำนวน n_x = 1,174 ครั้งในข้อมูลการวิจารณ์สินค้า ค่าความน่าจะเป็น p_x = 3.1E-5 ที่จะเกิดขึ้นของคำนามทั้งหมดในการเขียนภาษาอังกฤษทั่วไป และค่า $\ln(P(n_x))$ = -6,429 จากเงื่อนไขในการพิจารณาคุณลักษณะสินค้าที่กล่าวแล้วข้างต้น คือ $n_x > p_x N$ และ $\ln(P(n_x))$ มีค่าน้อยมาก สามารถระบุได้ว่า "lens" เป็นคุณลักษณะสินค้า

2.11.2.4 Probabilistic Model Approach

แกมกาญจน์ สมประเสริฐศรี และภัทรชัย ลลิตโรจน์วงศ์ [31] ได้นำเสนอวิธีการสกัดคุณลักษณะสินค้า โดยใช้แมกซิมั่มเอนโทรปีซึ่งเป็นแบบจำลองความน่าจะเป็นร่วมกับคำและไวยากรณ์ในประโยค โดยมีสมมุติฐานที่ว่า บริบทของคำในประโยคน่าจะเป็นข้อสนเทศที่สำคัญต่อการสกัดคุณลักษณะสินค้า วิธีการนี้ใช้การเรียนรู้แบบมีการสอนในการจำแนกประเภทคำว่าเป็นคุณลักษณะสินค้าหรือไม่เป็นคุณลักษณะสินค้า โดยมีคุณสมบัติในการเรียนรู้ คือ (1) คำที่คาดว่าจะเป็นคุณลักษณะสินค้า และหมวดคำของคำนั้น เช่น "picture" ซึ่งเป็นคำที่คาดว่าจะเป็นคุณลักษณะสินค้า และมีหมวดคำเป็นคำนาม "NN" (2) ความถี่ที่เกิดขึ้นของคำว่ามีความถี่มากหรือน้อย ถ้าคำที่คาดว่าจะเป็นคุณลักษณะสินค้านั้นปรากฏน้อยกว่า 5 ครั้ง แสดงว่าคำนั้นเป็นคำที่มีความถี่น้อย ดังนั้น คุณสมบัติความถี่น้อยจะมีค่าเป็นจริง (3) ลักษณะของคำที่ประกอบด้วยตัวอักษรและตัวเลข ถ้าคำที่คาดว่าจะเป็นคุณลักษณะสินค้านั้นประกอบด้วยตัวอักษรและตัวเลข คุณสมบัติลักษณะคำจะมีค่าเป็นจริง (4) คำและหมวดของคำที่มีความสัมพันธ์แบบพึ่งพากับคำที่คาดว่าจะเป็นคุณลักษณะสินค้า ผลการทดลองพบว่าวิธีการนี้ให้ค่าความเที่ยงตรงสูงกว่าวิธีการของ Hu and Liu [2]

2.12 สรุปปัญหาของวิธีการสกัดคุณลักษณะสินค้าและความเห็น

จากงานวิจัยที่ผ่านมาสรุปได้ว่า การสกัดคุณลักษณะสินค้าและความเห็น จะเริ่มจากขั้นตอนการสกัดคุณลักษณะสินค้า แล้วนำคุณลักษณะสินค้าที่สกัดได้ไปค้นหาหรือสกัดความเห็นที่มีต่อคุณลักษณะสินค้านั้นๆ ซึ่งมีข้อเด่นตรงที่ไม่จำเป็นต้องทำการกำกับข้อมูลก่อนเหมือนวิธีแบบมีผู้สอน แต่พบว่ายังมีข้อจำกัดในการสกัดคุณลักษณะสินค้าและความเห็นจากประโยคการวิจารณ์สินค้าในบางกรณีอยู่ ซึ่งสามารถสรุปปัญหาการสกัดคุณลักษณะสินค้าและความเห็นจากวิธีการที่ผ่านมา ได้ดังนี้

1. วิธีการสกัดคุณลักษณะสินค้าด้วยการพิจารณาคำที่คาดว่าจะเป็นคุณลักษณะสินค้าเพียงอย่างเดียว โดยไม่มีการพิจารณาคำที่แสดงความเห็นในประโยคนั้นร่วมด้วย ทำให้คุณลักษณะสินค้าที่สกัดได้อาจจะไม่ใช่คุณลักษณะสินค้าที่ลูกค้ามีการแสดงความเห็น เช่น ประโยคการวิจารณ์สินค้า "*Canon launched many new A series cameras in August 2007 with Image Stabilization and improved zoom.*" จากประโยคการวิจารณ์กล้องดิจิทัลนี้ ถึงแม้ว่า วลี "*Image Stabilization*" และคำ

"*zoom*" จะเป็นคุณลักษณะสินค้า แต่จะพบว่าในประโยคการวิจารณ์สินค้านี้ไม่ได้มีการแสดงความเห็นต่อคุณลักษณะสินค้าทั้ง 2 คุณลักษณะ

2. วิธีการสกัดความเห็นที่มีต่อคุณลักษณะสินค้าที่พิจารณาจากคำคุณศัพท์ที่อยู่ใกล้กันกับคุณลักษณะสินค้า และการที่ใช้กฎของรูปแบบการเกิดขึ้นร่วมกันระหว่างคุณลักษณะสินค้าและความเห็นนั้น ยังมีข้อจำกัดสำหรับประโยคบางลักษณะที่คุณลักษณะสินค้าไม่ได้อยู่ใกล้กับความเห็น ตัวอย่างประโยค "*I bought my canon g3 about a month ago and i have to say i am very **satisfied**.*" ซึ่งไม่สามารถสกัดด้วยวิธีการใช้คำคุณศัพท์ที่อยู่ใกล้กัน และวิธีการใช้กฎได้

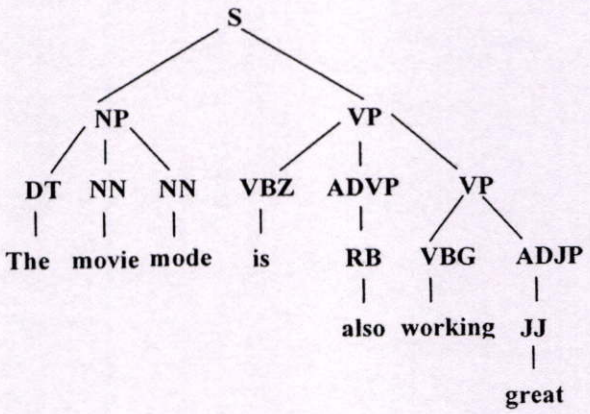
จากปัญหาของวิธีการสกัดคุณลักษณะสินค้าและความเห็นจากข้อมูลการวิจารณ์สินค้าของงานวิจัยที่ผ่านมา ที่การสกัดคุณลักษณะสินค้าไม่ได้มีการพิจารณาคำที่แสดงความคิดเห็นร่วมด้วย ทำให้คุณลักษณะสินค้าที่สกัดได้นั้นอาจเป็นคุณลักษณะสินค้าที่ไม่ได้มีการแสดงความเห็น และการสกัดความเห็นโดยวิเคราะห์จากการเกิดร่วมกันระหว่างคุณลักษณะสินค้าและความเห็น ยังมีข้อจำกัดสำหรับประโยคที่มีโครงสร้างแบบซับซ้อนที่คุณลักษณะสินค้าไม่ได้อยู่ใกล้กับความเห็น จากปัญหาดังกล่าว งานวิจัยนี้จึงนำเสนอวิธีการในการสกัดคุณลักษณะสินค้าและความเห็น โดยใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับการวิเคราะห์ความสัมพันธ์แบบพึ่งพา ซึ่งมีความยืดหยุ่นของตำแหน่งระหว่างคู่ของคำมากกว่าการใช้การเกิดขึ้นร่วมกันของคำที่ใกล้กันของงานวิจัยที่ผ่านมา นอกจากนี้วิธีการที่นำเสนอยังมีการพิจารณาคุณลักษณะสินค้าร่วมกับความเห็นที่มีต่อคุณลักษณะสินค้านั้นด้วย ซึ่งจะช่วยให้การสกัดประโยคความเห็นมีประสิทธิภาพดีกว่าวิธีการที่ผ่านมาที่พิจารณาเฉพาะคำที่คาดว่าจะเป็นคุณลักษณะสินค้าเพียงอย่างเดียว โดยไม่มีการพิจารณาคำที่แสดงความเห็นในประโยคนั้นร่วมด้วย

บทที่ 3
การสกัดคุณลักษณะสินค้าและความเห็นจากการวิจารณ์สินค้า

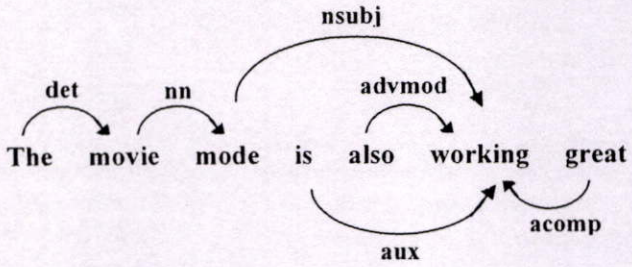
ในบทนี้จะกล่าวถึงรายละเอียดของความสัมพันธ์แบบพึ่งพารวมถึงข้อสนเทศจากความสัมพันธ์แบบพึ่งพาที่ใช้สำหรับสกัดคุณลักษณะสินค้าและความเห็น และขั้นตอนการสกัดคุณลักษณะสินค้าและความเห็นจากการวิจารณ์สินค้าโดยใช้แบบจำลองแมกซิมั่มเอนโทรปีซึ่งเป็นแบบจำลองความน่าจะเป็นร่วมกับการวิเคราะห์ความสัมพันธ์แบบพึ่งพา

3.1 ข้อสนเทศจากความสัมพันธ์แบบพึ่งพาสำหรับสกัดคุณลักษณะสินค้าและความเห็น

ปัจจุบันแม้ว่าจะมีการวิจัยด้านการสรุปความเห็นมากขึ้น แต่วิธีการสกัดคุณลักษณะสินค้าซึ่งเป็นหัวข้อที่ลูกค้าแสดงความเห็นในการวิจารณ์สินค้า และการสกัดความเห็นที่มีต่อคุณลักษณะสินค้านั้น ยังมีข้อจำกัดดังที่กล่าวแล้วในบทที่ 2 ในการศึกษาครั้งนี้พยายามแก้ปัญหาในการสกัดคุณลักษณะสินค้าและความเห็นที่ถูกต้อง โดยมีสมมุติฐานที่ว่าความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็นสามารถพิจารณาได้จากข้อสนเทศทางไวยากรณ์ที่วิเคราะห์จากความสัมพันธ์แบบพึ่งพาระหว่างคุณลักษณะสินค้าและความเห็นในประโยค การวิเคราะห์ความสัมพันธ์แบบพึ่งพาจะวิเคราะห์จากสายโซ่ความสัมพันธ์แบบพึ่งพาที่สร้างจากผลลัพธ์ของการแจงส่วน ตัวอย่างต้นไม้แจงส่วน และสายโซ่ความสัมพันธ์แบบพึ่งพาสำหรับประโยค "*The movie mode is also working great.*" แสดงได้ดังรูปที่ 3.1 และรูปที่ 3.2



รูปที่ 3.1 ตัวอย่างต้นไม้แจงส่วน



รูปที่ 3.2 ตัวอย่างสายโซ่ความสัมพันธ์แบบพึ่งพา

ในการหาความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็น โดยใช้คำศัพท์เพื่อระบุความสัมพันธ์จะมีความหลากหลาย ไม่สามารถระบุได้เหมือนการหาความความสัมพันธ์ระหว่างเอนทิตีในงานสกัดสารสนเทศ ที่สามารถหาความความสัมพันธ์ได้จากการสร้างศัพทวากยสัมพันธ์ (Lexico-syntactic) ที่เกี่ยวข้อง ตัวอย่างเช่น การหาความสัมพันธ์ระหว่างเอนทิตี 2 เอนทิตีที่มีความสัมพันธ์ works_at ตัวอย่างรูปแบบศัพทวากยสัมพันธ์ที่ใช้ในการสกัด เช่น

Person → subj → is employed by ←obj ← Organization

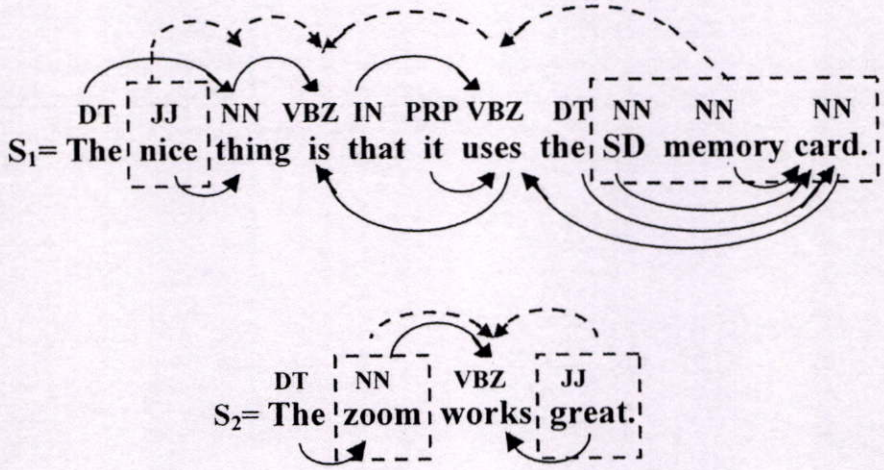
โดยมีคำกริยา "employ" เป็นข้อมูลหลักสำหรับสกัดเอนทิตี Person ที่เป็นคำที่อยู่ในตำแหน่งประธานแต่เป็นผู้ถูกกระทำของประโยค และเอนทิตี Organization ที่เป็นคำที่อยู่ในตำแหน่งกรรมของประโยค โดยที่ เอนทิตี Person มีความสัมพันธ์ works_at กับ เอนทิตี Organization

จากปัญหาความหลากหลายด้านภาษา การหาความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็น งานวิจัยนี้จึงนำเสนอข้อสนเทศที่สามารถแสดงถึงความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็น ได้แก่ เส้นทางของความสัมพันธ์แบบพึ่งพาระหว่างคุณลักษณะสินค้าและความเห็น และประเภทความสัมพันธ์ของคุณลักษณะสินค้าและความเห็น

3.1.1 เส้นทางของความสัมพันธ์แบบพึ่งพา

เส้นทางของความสัมพันธ์แบบพึ่งพา เป็นเส้นทางของหมวดคำที่เริ่มตั้งแต่หมวดคำของคุณลักษณะสินค้าไปจนถึงหมวดคำของความเห็น ซึ่งสามารถสร้างได้จากสายโซ่ความสัมพันธ์แบบพึ่งพา การหาเส้นทางจะเริ่มจากคำคุณลักษณะสินค้าไปตามสายโซ่ความสัมพันธ์แบบพึ่งพาเรื่อยไปถึงคำความเห็น ตัวอย่างการหาเส้นทางของความสัมพันธ์แบบพึ่งพาแสดงดังรูปที่ 3.3 จากตัวอย่างประโยค "*The nice thing is that it uses the SD memory card.*" มีคำแสดงความเห็น "*nice*" ที่แสดงความเห็นต่อคุณลักษณะสินค้า "*SD memory card*" การหาเส้นทางจะเริ่มจากคุณลักษณะสินค้า "*SD memory card*" ไปตามสายโซ่ความสัมพันธ์แบบพึ่งพาจนถึงคำแสดงความเห็น "*nice*" ซึ่งจะได้เส้นทางของความสัมพันธ์แบบพึ่งพา คือ ***NN VB VB NN JJ***ส่วนตัวอย่างประโยค "*The*

zoom works great." มีคำแสดงความเห็น "*great*" ที่แสดงความเห็นต่อคุณลักษณะสินค้า "*zoom*" การหาเส้นทางจะเริ่มจากคุณลักษณะสินค้า "*zoom*" ไปตามสายโซ่ความสัมพันธ์แบบพึ่งพาจนถึงคำแสดงความเห็น "*great*" ซึ่งจะได้เส้นทางของความสัมพันธ์แบบพึ่งพา คือ ***NN VB JJ*** โดย NN หมายถึง คำนาม VB หมายถึง คำกิริยา และ JJ หมายถึง คำคุณศัพท์ ซึ่งชุดหมวดคำที่ใช้ในงานวิจัยนี้แสดงในภาคผนวก ก.



S_1: Feature= "SD memory card" **Opinion word = "nice"**	**NN VB VB NN JJ**
S_2: Feature = "zoom" **Opinion word = "great"**	**NN VB JJ**

รูปที่ 3.3 ตัวอย่างสายโซ่ความสัมพันธ์แบบพึ่งพาและเส้นทางความสัมพันธ์

3.1.2 ประเภทความสัมพันธ์ของคุณลักษณะสินค้าและความเห็น

ในข้อมูลการวิจารณ์สินค้าที่ลูกค้าเขียนบรรยายเป็นประโยค งานวิจัยนี้จะแบ่งลักษณะความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็นเป็น 6 ประเภท ดังนี้

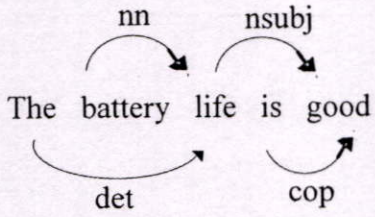
3.1.2.1 ความสัมพันธ์แบบลูก (Child Relationship) เป็นความสัมพันธ์แบบพึ่งพาระหว่างคุณลักษณะสินค้าและความเห็น ที่คุณลักษณะสินค้าเป็นหน่วยคำพึ่งพาของความเห็น แสดงดังรูปที่ 3.4 (a) ลักษณะความสัมพันธ์แบบลูกมี 2 กรณี คือ กรณีที่คุณลักษณะสินค้าทำหน้าที่เป็นประธานในประโยค และความเห็นเป็นส่วนเติมเต็มที่ตามหลังวิกตรรถกริยา ที่ทำให้ประโยคมีความหมายสมบูรณ์ และกรณีที่คุณลักษณะสินค้าทำหน้าที่เป็นกรรมในประโยคและความเห็นจะทำ

หน้าที่เป็นกริยาที่บ่งบอกถึงความรู้สึก เช่น love, like, hate และ appreciate เป็นต้น ตัวอย่างประโยคการวิจารณ์สินค้าที่คุณลักษณะสินค้าและความเห็นมีความสัมพันธ์แบบลูก เช่น

(1) *The battery life is good.*

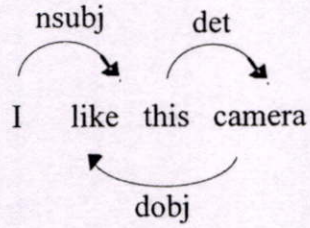
(2) *I like this camera.*

จากตัวอย่างประโยค *"The battery life is good."* ความสัมพันธ์แบบพึ่งพาของคำในประโยคสามารถแสดงได้ดังนี้ *{det(life-3, The-1), nn(life-3, battery-2), nsubj(good-5, life-3), cop(good-5, is-4)}*



ซึ่งมี *"battery life"* เป็นคำแสดงคุณลักษณะสินค้าและทำหน้าที่เป็นประธานในประโยค ส่วน *"good"* เป็นคำแสดงความเห็นและทำหน้าที่เป็นส่วนเติมเต็มในประโยคที่ตามหลัง *"is"* ซึ่งเป็นวิกตรรถกริยา ดังนั้นคำแสดงคุณลักษณะสินค้า *"battery life"* จึงมีความสัมพันธ์แบบลูกกับคำแสดงความเห็น *"good"*

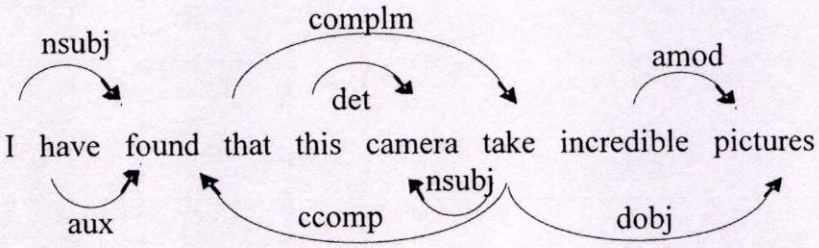
จากตัวอย่างประโยค *"I like this camera."* ความสัมพันธ์แบบพึ่งพาของคำในประโยคสามารถแสดงได้ดังนี้ *{nsubj(like-2, I-1), det(camera-4, this-3), dobj(like-2, camera-4)}*



ซึ่งมี *"camera"* เป็นคำแสดงคุณลักษณะสินค้าและทำหน้าที่เป็นกรรมในประโยค ส่วน *"like"* เป็นคำแสดงความเห็นและทำหน้าที่เป็นกริยาในประโยค ดังนั้นคำแสดงคุณลักษณะสินค้า *"camera"* จึงมีความสัมพันธ์แบบลูกกับคำแสดงความเห็น *"like"*

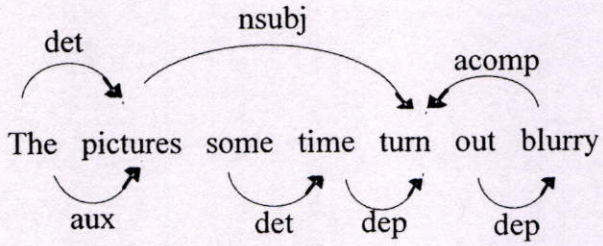
3.1.2.2 ความสัมพันธ์แบบพ่อ (Parent Relationship) เป็นความสัมพันธ์แบบพึ่งพาระหว่างคุณลักษณะสินค้าและความเห็นที่คุณลักษณะสินค้าเป็นหน่วยคำหลักของความเห็น แสดงดังรูปที่ 3.4 (b) โดยความเห็นทำหน้าที่เป็นตัวขยายของคุณลักษณะสินค้า ซึ่งได้แก่ amod (Adjectival Modifier) และ rcmod (Relative Clause Modifier) ตัวอย่างประโยคการวิจารณ์สินค้าที่คุณลักษณะสินค้าและความเห็นมีความสัมพันธ์แบบพ่อ เช่น ประโยค *"I have found that this camera take incredible pictures."* ซึ่งมีความสัมพันธ์แบบพึ่งพาของคำในประโยคดังนี้

{nsubj(found-3, I-1), aux(found-3, have-2), complm(take-7, that-4),det(camera-6, this-5), nsubj(take-7, camera-6), ccomp(found-3, take-7),amod(pictures-9, incredible-8), dobj(take-7, pictures-9)}



จากตัวอย่างประโยค *"I have found that this camera take incredible pictures."* ซึ่งมี *"pictures"* เป็นคำแสดงคุณลักษณะสินค้า และ *"incredible"* เป็นคำแสดงความเห็นซึ่งทำหน้าที่เป็นคุณศัพท์ขยายคำแสดงคุณลักษณะ *"pictures"* ดังนั้นคำแสดงคุณลักษณะสินค้า *"pictures"* จึงมีความสัมพันธ์แบบพ่อกับคำแสดงความเห็น *"incredible"*

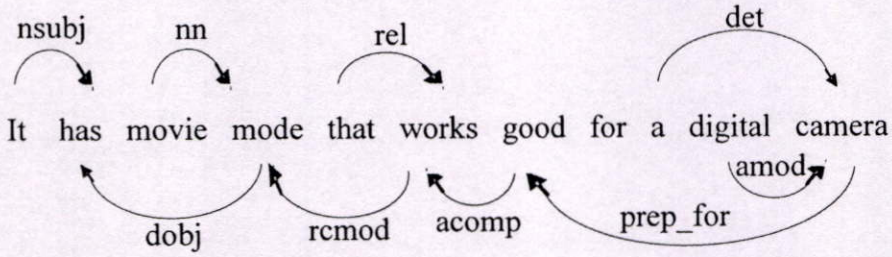
3.1.2.3 ความสัมพันธ์แบบพี่น้อง (Sibling Relationship) เป็นความสัมพันธ์แบบพึ่งพาระหว่างคุณลักษณะสินค้าและความเห็นที่คุณลักษณะสินค้าและความเห็นพึ่งพาหน่วยคำหลักเดียวกันแสดงดังรูปที่ 3.4 (c) โดยความเห็นทำหน้าที่เป็นส่วนเติมเต็มในประโยค ซึ่งได้แก่ advmod (Adverbial Modifier), acomp (Adjectival Complement), ccomp (Clausal Complement) และ xcomp (Open Clausal Complement) ตัวอย่างประโยคการวิจารณ์สินค้าที่คุณลักษณะสินค้าและความเห็นมีความสัมพันธ์แบบพี่น้อง เช่นประโยควิจารณ์สินค้า *"The pictures some time turn out blurry."* ซึ่งมีความสัมพันธ์แบบพึ่งพาของคำในประโยคดังนี้ *{det(picture-2, The-1), nsubj(turns-5, picture-2), det(time-4, some-3), dep(turns-5, time-4), dep(blurry-7, out-6), acomp(turns-5, blurry-7)}*



จากตัวอย่างประโยค *"The pictures some time turn out blurry."* ซึ่งมี *"pictures"* เป็นคำแสดงคุณลักษณะสินค้า และ *"blurry"* เป็นคำแสดงความเห็น ซึ่งทั้งคำแสดงคุณลักษณะ *"pictures"* และ

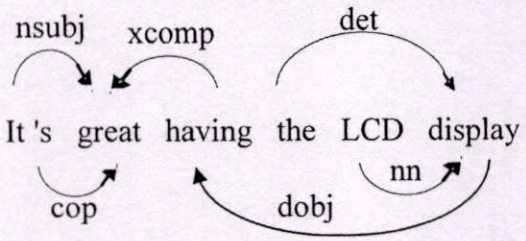
คำแสดงความเห็น *"blurry"* มีคำหลักคำเดียวกัน และ *"blurry"* ทำหน้าที่เป็นกริยาวิเศษณ์ขยายคำแสดงคุณลักษณะ *"pictures"* ดังนั้น คำแสดงคุณลักษณะสินค้า *"pictures"* จึงมีความสัมพันธ์แบบพี่น้องกับคำแสดงความเห็น *"blurry"*

3.1.2.4 ความสัมพันธ์แบบปู่ย่า (Grandparent Relationship) เป็นความสัมพันธ์แบบพึ่งพาระหว่างคุณลักษณะสินค้าและความเห็นที่ความเห็นพึ่งพาหน่วยคำที่พึ่งพาคุณลักษณะสินค้าแสดงดังรูปที่ 3.4 (d) โดยความเห็นทำหน้าที่เป็นส่วนขยายหรือส่วนเติมเต็มสมบูรณ์ของส่วนที่ขยายคุณลักษณะสินค้า ซึ่งได้แก่ amod (Adjectival Modifier), rcmod (Relative Clause Modifier), acomp (Adjectival Complement) และ xcomp (Open Clausal Complement) ตัวอย่างประโยคการวิจารณ์สินค้าที่คุณลักษณะสินค้าและความเห็นมีความสัมพันธ์แบบปู่ย่า เช่น ประโยค "*It has movie mode that works good for a digital camera.*" ซึ่งมีความสัมพันธ์แบบพึ่งพาของคำในประโยคดังนี้ *{nsubj(has-2, It-1), nn(mode-4, movie-3), dobj(has-2, mode-4),rel(works-6, that-5), rcmod(mode-4, works-6), acomp(works-6, good-7),det(camera-11, a-9), amod(camera-11, digital-10), prep_for(good-7,camera-11)}*

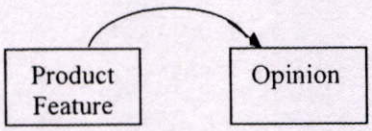


จากตัวอย่างประโยคซึ่งมี *"movie mode"* เป็นคุณลักษณะสินค้า และ *"good"* เป็นคำแสดงความเห็น ซึ่งคำแสดงความเห็น *"good"* ทำหน้าที่เป็นส่วนเติมเต็มของส่วนขยายคำแสดงคุณลักษณะสินค้า *"movie mode"* ดังนั้น คำแสดงคุณลักษณะสินค้า *"movie mode"* จึงมีความสัมพันธ์แบบปู่ย่ากับคำแสดงความเห็น *"good"*

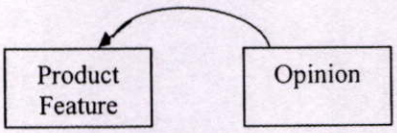
3.1.2.5 ความสัมพันธ์แบบหลาน (Grandchild Relationship) เป็นความสัมพันธ์แบบพึ่งพาระหว่างคุณลักษณะสินค้าและความเห็นที่คุณลักษณะสินค้าพึ่งพาหน่วยคำที่พึ่งพาความเห็นแสดงดังรูปที่ 3.4 (e) โดยคุณลักษณะสินค้าทำหน้าที่เป็นกรรม \ของส่วนเติมเต็มและความเห็นจะทำหน้าที่เป็นส่วนเติมเต็มที่ตามหลังวิกตรรถกริยา ซึ่งได้แก่ ccomp (Clausal Complement) และ xcomp (Open Clausal Complement) ตัวอย่างประโยคการวิจารณ์สินค้าที่คุณลักษณะสินค้าและความเห็นมีความสัมพันธ์แบบหลาน เช่น ประโยค "*It's great having the LCD display.*" ซึ่งมีความสัมพันธ์แบบพึ่งพาของคำในประโยคดังนี้ *{nsubj(great-3, It-1), cop(great-3, 's-2), xcomp(great-3, having-4), det(display-7, the-5), nn(display-7, LCD-6), dobj(having-4, display-7)}*



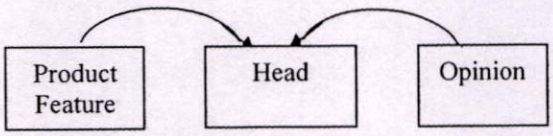
จากตัวอย่างประโยคซึ่งมี *"LCD display"* เป็นคุณลักษณะสินค้า และ *"great"* เป็นคำแสดงความเห็น ซึ่งคำแสดงคุณลักษณะสินค้าทำหน้าที่เป็นกรรมของส่วนเติมเต็ม ส่วน *"great"* ทำหน้าที่เป็นส่วนเติมเต็มในประโยคที่ตามหลัง *"is"* ดังนั้น คำแสดงคุณลักษณะสินค้า *"LCD display"* จึงมีความสัมพันธ์แบบหลานกับคำแสดงความเห็น *"great"*



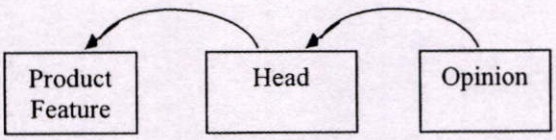
(a) ความสัมพันธ์แบบลูก



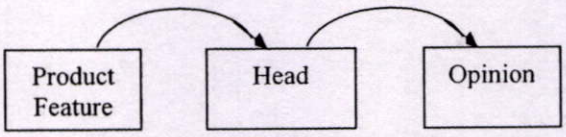
(b) ความสัมพันธ์แบบพ่อ



(c) ความสัมพันธ์แบบพี่น้อง



(d) ความสัมพันธ์แบบปู่ย่า



(e) ความสัมพันธ์แบบหลาน

รูปที่ 3.4 ประเภทความสัมพันธ์ของคุณลักษณะสินค้าและความเห็น

3.1.2.6 ความสัมพันธ์แบบทางอ้อม (Indirect Relationship) คุณลักษณะสินค้าและความเห็นไม่ได้มีความสัมพันธ์กันโดยตรงตามประเภทความสัมพันธ์ทั้ง 5 ประเภทที่กล่าวมา

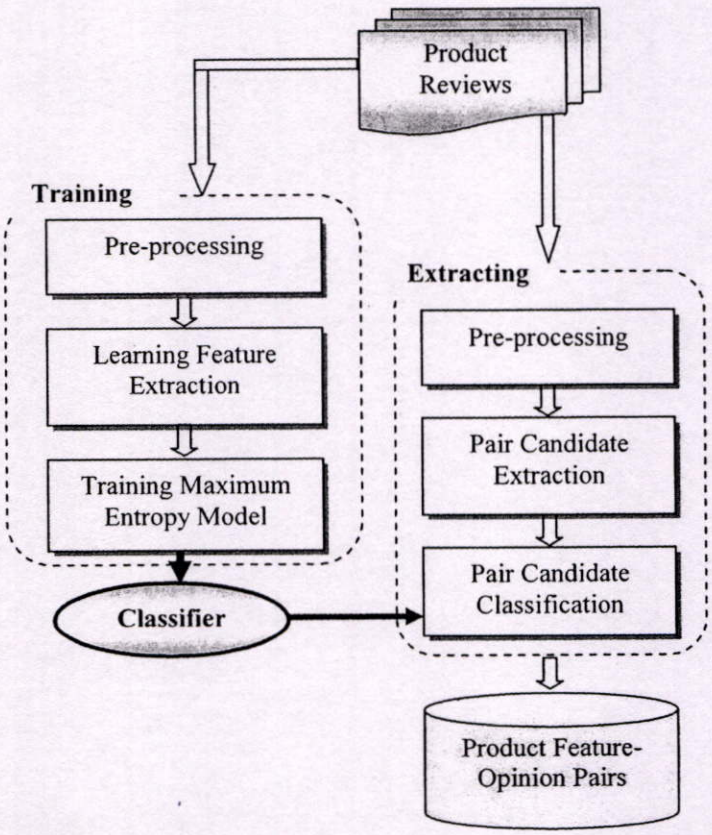
ประเภทความสัมพันธ์ของคุณลักษณะสินค้าและความเห็นในการสกัดคุณลักษณะสินค้าและความเห็นที่ใช้ในงานวิจัยนี้สามารถสรุปได้ดังตารางที่ 3.1

ตารางที่ 3.1 ประเภทความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็น

ประเภทความสัมพันธ์	คำอธิบาย
แบบลูก (Child)	คุณลักษณะสินค้าเป็นหน่วยคำพึ่งพาความเห็น
แบบพ่อ (Parent)	ความเห็นเป็นหน่วยคำพึ่งพาคุณลักษณะสินค้า
แบบพี่น้อง (Sibling)	คุณลักษณะสินค้าและความเห็นพึ่งพาหน่วยคำหลักเดียวกัน
แบบหลาน (Grandchild)	คุณลักษณะสินค้าพึ่งพาหน่วยคำที่พึ่งพาความเห็น
แบบปู่ย่า (Grandparent)	ความเห็นพึ่งพาหน่วยคำที่พึ่งพาคุณลักษณะสินค้า
แบบทางอ้อม (Indirect)	คุณลักษณะสินค้าและความเห็นไม่มีความสัมพันธ์กันโดยตรง

3.2 กระบวนการของการสกัดคุณลักษณะสินค้าและความเห็น

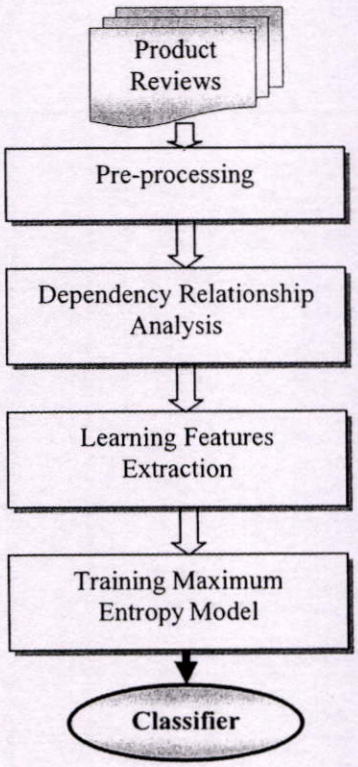
กระบวนการในการสกัดคุณลักษณะสินค้าและความเห็นในงานวิจัยนี้ ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับการวิเคราะห์ความสัมพันธ์แบบพึ่งพา ซึ่งภาพรวมของกระบวนการในการสกัดคุณลักษณะสินค้าและความเห็นแสดงดังรูปที่ 3.5 โดยแบ่งออกได้เป็น 2 ส่วน คือ ส่วนการเรียนรู้ (Training) และส่วนการสกัดคุณลักษณะสินค้าและความเห็น (Extraction) ดังมีรายละเอียดต่อไปนี้



รูปที่ 3.5 กระบวนการในการสกัดคุณลักษณะสินค้าและความเห็น

3.2.1 ส่วนการเรียนรู้

ในส่วนนี้เป็นส่วนของการเรียนรู้เพื่อสร้างตัวจำแนกที่จะใช้ในส่วนการสกัดคุณลักษณะสินค้าและความเห็น เพื่อจำแนกคู่ของคำว่าเป็นคุณลักษณะสินค้าและความเห็นหรือไม่ ในส่วนการเรียนรู้นี้ใช้แบบจำลองแมกซิมั่มเอนโทรปีในการเรียนรู้เป็นตัวจำแนก โดยขั้นตอนในส่วนของการเรียนรู้แบ่งได้เป็น 4 ขั้นตอน (รูปที่ 3.6) แสดงรายละเอียดได้ดังนี้



รูปที่ 3.6 ขั้นตอนในส่วนของการเรียนรู้

3.2.1.1 การเตรียมข้อมูลสำหรับการเรียนรู้ (Pre-processing)

ในขั้นตอนนี้ประกอบด้วยการรวบรวมประโยคการวิจารณ์สินค้า ซึ่งในวิทยานิพนธ์นี้ใช้การวิจารณ์สินค้าอิเล็กทรอนิกส์รูปแบบอิสระ ซึ่งส่วนใหญ่เป็นประโยคยาว ไม่มีการแบ่งคำวิจารณ์ออกเป็นด้านบวกและด้านลบดังตัวอย่างรูปที่ 3.7 แล้วทำการกำกับคู่ของคุณลักษณะสินค้าและความเห็นในประโยค ตามที่แสดงด้วยการเน้นตัวเข้มต้นประโยค จากนั้นนำข้อมูลการวิจารณ์สินค้าที่รวบรวมได้มาทำการแจงส่วนประโยค ในงานวิจัยนี้ใช้โปรแกรมสแตนฟอร์ดพาสเซอร์สำหรับแจงส่วนประโยค โดยผลลัพธ์ที่ได้จากการแจงส่วนแสดงดังรูปที่ 3.8 และ 3.9

3.2.1.2 การวิเคราะห์ความสัมพันธ์แบบพึ่งพาในส่วนการเรียนรู้ (Dependency Relationship Analysis)

ในขั้นตอนนี้จะนำประโยคการวิจารณ์สินค้าที่มีการทำป้ายกำกับคำคุณลักษณะสินค้า

และความเห็น และมีการแจกส่วนของประโยคจากขั้นตอนการเตรียมข้อมูลสำหรับการเรียนรู้มาวิเคราะห์ความสัมพันธ์แบบพึ่งพาของคำในประโยค โดยสร้างเป็นสายโซ่ความสัมพันธ์แบบพึ่งพาเพื่อใช้หาเส้นทางระหว่างคุณลักษณะสินค้าและความเห็น และกำหนดประเภทของความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็น การสร้างสายโซ่ความสัมพันธ์แบบพึ่งพาจะสร้างจากผลลัพธ์การแจงส่วนที่ได้จากโปรแกรมสแตนฟอร์ดพาสเซอร์ ซึ่งแสดงได้ดังรูปที่ 3.9 เมื่อทำการสร้างสายโซ่ความสัมพันธ์แบบพึ่งพาแล้ว ขั้นตอนต่อไปในส่วนของการเรียนรู้คือการสกัดคุณสมบัติสำหรับการเรียนรู้

3.2.1.3 การสกัดคุณสมบัติสำหรับการเรียนรู้ (Learning Features Extraction)

ในขั้นตอนนี้เป็นการสกัดคุณสมบัติของแต่ละคู่ของคุณลักษณะสินค้าและความเห็นเพื่อเป็นข้อสนเทศในการเรียนรู้ของแบบจำลองแมกซิมั่มเอนโทรปี รายละเอียดของการสกัดคุณสมบัตินี้จะขอกล่าวถึงในหัวข้อ 3.3 โดยผลลัพธ์ที่ได้จากขั้นตอนนี้ คือ ชุดของฟังก์ชันคุณสมบัติ (Feature function) ที่มีลักษณะเป็นไบนารีฟังก์ชัน ซึ่งมีค่าเป็น 0 หรือ 1 ดังสมการ

$$f_{cp,a'}(a,b) = \begin{cases} 1 & \text{if } a = a' and\ cp(b) = true \\ 0 & \text{otherwise} \end{cases}$$

ตัวอย่างการสร้างฟังก์ชันคุณสมบัติจากประโยค *"Battery is very good even when using flash and lcd."* แสดงได้ดังนี้

$$f(a,b) = \begin{cases} 1 & \text{ถ้าผลทำนาย a = yes และคำแสดงความเห็น b = good} \\ 0 & \text{อื่นๆ} \end{cases}$$

3.2.1.4 การเรียนรู้ (Training Maximum Entropy Model)

ในขั้นตอนนี้เป็นการเรียนรู้ของแบบจำลองแมกซิมั่มเอนโทรปี จากคุณสมบัติของแต่ละคู่ของคุณลักษณะสินค้าและความเห็นที่สกัดได้ในขั้นตอนที่ผ่านมา สำหรับการประมาณค่าพารามิเตอร์ α_j ของแบบจำลองแมกซิมั่มเอนโทรปีในงานวิจัยนี้ใช้อัลกอริทึม Generalized Iterative Scaling (GIS) รายละเอียดดังที่กล่าวไว้ในบทที่ 2 ส่วนผลลัพธ์ที่ได้จากขั้นตอนการเรียนรู้นี้คือ ชุดของค่าพารามิเตอร์ α_j หรือค่าถ่วงน้ำหนักสำหรับแต่ละฟังก์ชันคุณสมบัติ ซึ่งจะเป็นตัวจำแนกที่จะนำไปใช้ในส่วนของการสกัดคุณลักษณะสินค้าและความเห็นต่อไป

[canon, satisfied] ## I recently purchased the Canon and I am extremely satisfied with the purchase.

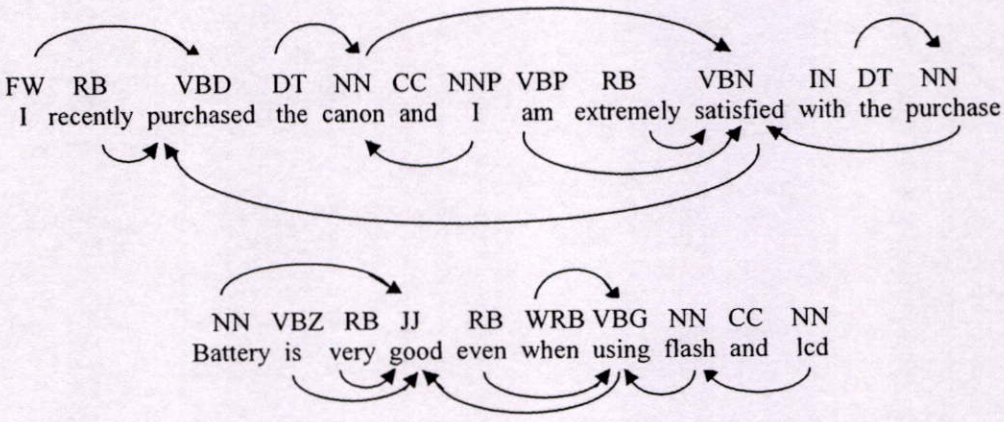
[battery, good] ## Battery is very good even when using flash and lcd.

รูปที่ 3.7 ตัวอย่างข้อมูลการวิจารณ์สินค้าที่ใช้ในการเรียนรู้

```
i/FW recently/RB purchased/VBD the/DT Canon/NNP and/CC i/NNP
am/VBP extremely/RB satisfied/VBN with/IN the/DT purchase/NN ./.

battery/NN is/VBZ very/RB good/JJ even/RB when/WRB using/VBG
flash/NN and/CC lcd/NN ./.
```

รูปที่ 3.8 ตัวอย่างผลลัพธ์การกำกับหมวดคำจากโปรแกรมสแตนฟอร์ดพาสเซอร์

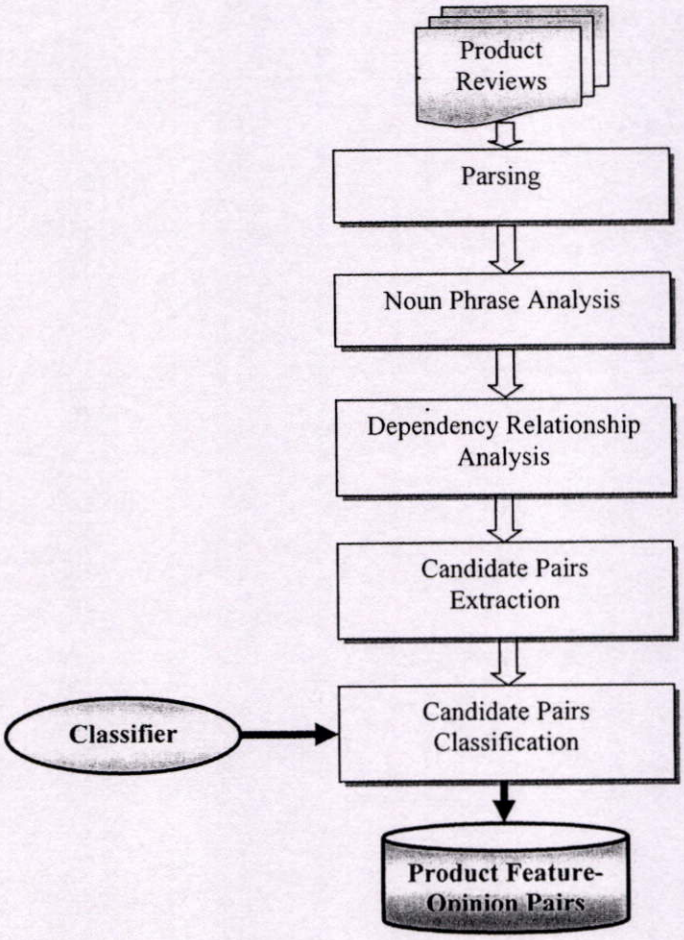
```
nsubj(purchased-3, I-1)
advmod(purchased-3, recently-2)
det(Canon-5, the-4)
nsubjpass(satisfied-10, Canon-5)
conj_and(Canon-5, i-7)
auxpass(satisfied-10, am-8)
advmod(satisfied-10, extremely-9)
ccomp(purchased-3, satisfied-10)
det(purchase-13, the-12)
prep_with(satisfied-10, purchase-13)

nsubj(good-4, battery-1)
cop(good-4, is-2)
advmod(good-4, very-3)
advmod(using-7, even-5)
advmod(using-7, when-6)
ccomp(good-4, using-7)
dobj(using-7, flash-8)
conj_and(flash-8, lcd-10)
```



รูปที่ 3.9 ตัวอย่างผลลัพธ์ความสัมพันธ์แบบพึ่งพาและสายโซ่ความสัมพันธ์แบบพึ่งพา

3.2.2 ส่วนการสกัดคุณลักษณะสินค้าและความเห็น

ในส่วนนี้เป็นการสกัดคุณลักษณะสินค้าและความเห็นจากข้อมูลการวิจารณ์สินค้า โดยใช้แบบจำลองแมกซิมั่มเอนโทรปีที่ผ่านการเรียนรู้ในส่วนของการเรียนรู้แล้วเป็นตัวจำแนกว่าคู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็น เป็นคู่คุณลักษณะสินค้าและความเห็นหรือไม่ ขั้นตอนในส่วนการสกัดคุณลักษณะสินค้าและความเห็นแบ่งได้เป็น 5 ขั้นตอน ดังรูปที่ 3.10 โดยมีรายละเอียดได้ดังนี้



รูปที่ 3.10 ขั้นตอนในส่วนของการสกัดคุณลักษณะสินค้าและความเห็น

3.2.2.1 การแจงส่วนประโยค (Parsing)

การแจงส่วนประโยคเป็นขั้นตอนแรกในส่วนของการสกัดคุณลักษณะสินค้าและความเห็น โดยใช้โปรแกรมสแตนฟอร์ดพาสเซอร์ ผลลัพธ์ที่ได้จากขั้นตอนนี้คือประโยคที่มีการกำกับหมวดคำในประโยคและความสัมพันธ์แบบพึ่งพาของคำในประโยค จะเหมือนกับผลลัพธ์ที่ได้จากขั้นตอนการแจงส่วนประโยคในส่วนการเรียนรู้ที่ผ่านมา

3.2.2.2 การวิเคราะห์นามวลี (Noun Phrase Analysis)

โดยทั่วไปแล้ว คำที่กล่าวถึงคุณลักษณะสินค้าส่วนใหญ่มักจะเป็นคำนาม ซึ่งอาจจะ

เป็นคำเดี่ยว หรือเป็นวลีที่ประกอบด้วยคำหลายคำ ตัวอย่างเช่น ประโยคการวิจารณ์กล้องดิจิทัล "*The zoom works great.*" ซึ่งมีคุณลักษณะสินค้า คือ "*zoom*" ซึ่งเป็นคำเดี่ยว และ "*The nice thing is that it uses the SD memory card.*" มีคุณลักษณะสินค้า คือ "*SD memory card*" ซึ่งเป็นนามวลีที่ประกอบไปด้วยคำ 3 คำ ดังนั้น การสกัดคำที่คาดว่าจะเป็นคุณลักษณะสินค้าในงานวิจัยนี้ จึงพิจารณาจากคำนามหรือนามวลี ซึ่งนามวลีที่คาดว่าจะเป็นคุณลักษณะสินค้านี้ไม่สามารถใช้นามวลีที่ได้จากการแจงส่วนของประโยคโดยตรง เนื่องจากนามวลีที่ได้ อาจจะประกอบไปด้วยความเห็นและคุณลักษณะสินค้า เช่น "*good pictures*" ดังนั้น เพื่อลดปัญหาที่จะเกิดขึ้นในการสกัดคุณลักษณะสินค้าและความเห็น งานวิจัยนี้จึงใช้รูปแบบของนามวลีร่วมกับพจนานุกรม GI (General Inquirer's Harvard-4 Dictionary) [32] ในการวิเคราะห์นามวลีแทนการใช้นามวลีที่ได้จากการแจงส่วนประโยค โดยอัลกอริทึมที่ใช้สำหรับการวิเคราะห์นามวลีแสดงดังรูปที่ 3.11 และรูปแบบของนามวลีที่ใช้มี 9 รูปแบบ ดังนี้

รูปแบบที่ 1 NN	รูปแบบที่ 2 NN NN	รูปแบบที่ 3 JJ NN
รูปแบบที่ 4 NN NN NN	รูปแบบที่ 5 JJ NN NN	รูปแบบที่ 6 JJ JJ N
รูปแบบที่ 7 NN IN NN	รูปแบบที่ 8 NN IN DT NN	รูปแบบที่ 9 NN NN NN NN

เมื่อ NN คือ คำนาม JJ คือ คำคุณศัพท์ DT คือ คำนำหน้านาม และ IN คือ คำบุพบท

3.2.2.3 การวิเคราะห์ความสัมพันธ์แบบพึ่งพาในส่วนการสกัด (Dependency Relationship Analysis)

ในขั้นตอนนี้เป็นการสร้างสายโซ่ความสัมพันธ์แบบพึ่งพา เพื่อใช้หาเส้นทางระหว่างคุณลักษณะสินค้าและความเห็น และประเภทความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็น โดยสร้างจากผลลัพธ์การกำกับหมวดคำ และความสัมพันธ์แบบพึ่งพาที่ได้จากโปรแกรมสแตนฟอร์ดพาสเซอร์ ซึ่งมีรายละเอียดเหมือนที่ได้กล่าวมาแล้วในส่วนการเรียนรู้

3.2.2.4 การสกัดคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น (Candidate Pairs Extraction)

ขั้นตอนนี้เป็นการสกัดคำที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็น ซึ่งรวมถึงการสกัดคุณสมบัติจากคู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็นด้วย ข้อสนเทศสำคัญที่ใช้ในการจำแนกคือ เส้นทางและประเภทความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็นที่ได้จากสายโซ่ความสัมพันธ์แบบพึ่งพา แต่ในหัวข้อนี้จะขอกล่าวถึงรายะเอียดของการสกัดคู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็น ส่วนการสกัดคุณสมบัติจะกล่าวรายละเอียดในหัวข้อที่ 3.3 การสกัดคำที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็นในขั้นตอนนี้ จะเริ่มจากการสกัดคำที่คาดว่าจะเป็นคุณลักษณะสินค้าก่อน โดยพิจารณาคำที่เป็นคำนาม หรือนามวลีที่ได้จากขั้นตอนการวิเคราะห์นามวลีที่ไม่ใช่คำหยุด (Stop Word) หลังจากนั้น จะสกัดคำที่คาดว่าจะเป็นความเห็น โดยพิจารณาคำคุณศัพท์ หรือคำกริยาที่ปรากฏในพจนานุกรม GI ที่มีความสัมพันธ์แบบพึ่งพากับคำที่

คาดว่าจะเป็นคุณลักษณะสินค้า ซึ่งในแต่ละประโยค อาจจะมีคู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็นได้มากกว่า 1 คู่ ซึ่งอัลกอริทึมของการสกัดคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็นแสดงได้ดังรูปที่ 3.12

```
Input   S  – Set of tagged sentences; s = s1,s2,…,sm
        P  – Set of noun phrase patterns
        GI – Set of words in GI dictionary
Output  NP – Set of noun phrases of each sentence
        NP = ∅
        For each tagged sentence sn ∈ S
           NPS = ∅
                For i=1 to end of sentence sn
                  If  i < Length(sn) – 2  Then  x = 3
                    Else If  i = Length(sn) – 2  Then  x = 2
                        Else If  i = Length(sn) – 1  Then  x = 1
                            Else x = 0
                             End
                        End
                  End
                  For j = x to 0
                     GT = Ti to Ti+j    /* POS Tag of wordi to wordi+j of sn */
                     GW = wordi to wordi+j
                     If  GT ∈ P and GW ∉ GI
                        i = i+j
                        NPS = NPS ∪ GW
                        Break
                     End
                  End
                End
           NP = NP ∪ NPS
        End
```

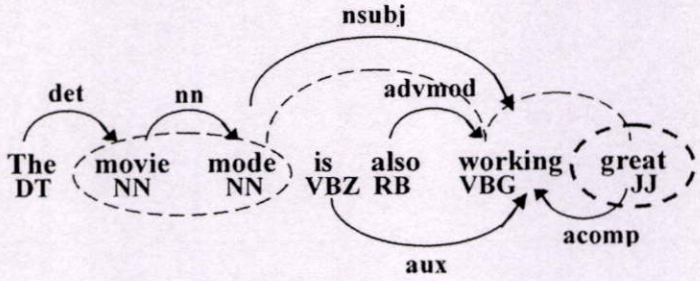
รูปที่ 3.11 การวิเคราะห์นามวลีสำหรับการสกัดคุณลักษณะสินค้า

```
Input     DT – Set of dependency trees
          PS – Set of product feature candidates in each sentence
          GI – Set of words in GI dictionary
Output    FOS – Set of product feature-opinion pairs

PairExtract(dt_i, ps_i) /* Return the set of product feature-opinion pairs of each
                           dependency tree */
FO = ø
  For m=1 to end of product feature candidate ps_i
     Rnode = ps_i(m)              /* Initial product feature candidate to root node */
     f = FirstVisit(dt_i,Rnode)   /* Find first neighbor, return -1 if no neighbor */
     While (f ~= -1)
        If (neighbor is adjective) or (neighbor is adveb and ∈ GI) then
          pair = [Rnode,neigbor]  /* product feature-opinion pair */
          FO = FO ∪ pair
          f = -1
        Else
          f = NextVisit(dt_i,Rnode)  /* Find next neighbor, return -1 if no neighbor */
        End
     End
  End
PairExtract(DT, PS, GI)           /* Return set of product feature-opinion pairs */
  For each dependency tree dt_i ∈ DT
      FO = PairExtract(dt_i, ps_i)
  End
```

รูปที่ 3.12 การสกัดคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น

ตัวอย่างการสกัดคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็นจากประโยค "*The movie mode is also working great.*" แสดงได้ดังรูปที่ 3.13 เริ่มจากการสกัดคำที่คาดว่าเป็นคุณลักษณะสินค้า คือ "*movie mode*" ซึ่งเป็นนามวลีที่ได้จากขั้นตอนการวิเคราะห์นามวลี หลังจาก

นั้นดึงคำคุณศัพท์ที่มีความสัมพันธ์แบบพึ่งพากับ "*movie mode*" ในที่นี้ก็คือ "*great*" ซึ่งถือว่าเป็นคำที่คาดว่าเป็นความเห็นที่มีต่อ"*movie mode*" ดังนั้น จากประโยคตัวอย่างนี้เราสามารถสกัดคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็นได้คือ [*movie mode*, *great*]



รูปที่ 3.13 ตัวอย่างการสกัดคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น

3.2.2.5 การจำแนกประเภทคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น (Candidate pairs Classification)

ในขั้นตอนนี้เป็นการใช้แบบจำลองแมกซิมั่มเอนโทรปีที่ฝึกฝนแล้วในส่วนของการเรียนรู้ มาจำแนกคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น ซึ่งผลการจำแนก a มี 2 ประเภท คือ เป็นคู่คุณลักษณะสินค้าและความเห็น และไม่เป็นคู่คุณลักษณะสินค้าและความเห็น ในขั้นตอนการจำแนกนี้แบบจำลองแมกซิมั่มเอนโทรปีซึ่งเป็นตัวจำแนก จะอาศัยคุณสมบัติของคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น ในการจำแนกว่าเป็นคู่คุณลักษณะสินค้าและความเห็นหรือไม่เป็นคู่คุณลักษณะสินค้าและความเห็น โดยจะพิจารณาจากค่าความน่าจะเป็นแบบมีเงื่อนไขที่ให้ค่าสูงสุดในบริบท b ซึ่งความน่าจะเป็นแบบมีเงื่อนไขคำนวณได้จากสมการดังนี้

$$p_{\max}(a \mid b) = \frac{1}{Z(b)} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)}$$

Classifier cl: X → Y

และ

$$cl(b) = \arg\max_{a} p(a \mid b)$$

ตัวอย่างการจำแนกคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น

สมมุติให้ ฟังก์ชันคุณสมบัติและค่าถ่วงน้ำหนักของแต่ละฟังก์ชันคุณสมบัติที่ได้จากขั้นตอนการเรียนรู้ของแบบจำลองแมกซิมั่มเอนโทรปีมีดังนี้

f_1 (คำคุณลักษณะสินค้า = "phone", yes) ค่าถ่วงน้ำหนัก = 7.599

f_2 (คำคุณลักษณะสินค้า = "zoom", yes) ค่าถ่วงน้ำหนัก = 3.663

f_3 (คำคุณลักษณะสินค้า = “camera”, no) ค่าถ่วงน้ำหนัก = 1.728

f_4 (คำคุณลักษณะสินค้า = “weather”, no) ค่าถ่วงน้ำหนัก = 9.714

f_5 (คำแสดงความเห็น = “great”, yes) ค่าถ่วงน้ำหนัก = 5.605

f_6 (คำแสดงความเห็น = “great”, no) ค่าถ่วงน้ำหนัก = 9.312

f_7 (คำแสดงความเห็น = “awesome”, yes) ค่าถ่วงน้ำหนัก = 3.795

f_8 (คำแสดงความเห็น = “best”, no) ค่าถ่วงน้ำหนัก = 3. 723

f_9 (เส้นทางความสัมพันธ์ = “NNVBJJ”, yes) ค่าถ่วงน้ำหนัก = 2.728

f_{10} (เส้นทางความสัมพันธ์ = “NNVBJJ”, no) ค่าถ่วงน้ำหนัก = 2.795

f_{11} (เส้นทางความสัมพันธ์ = “NNJJ”, yes) ค่าถ่วงน้ำหนัก = 3.566

f_{12} (เส้นทางความสัมพันธ์ = “NNJJ”, no) ค่าถ่วงน้ำหนัก = 2. 395

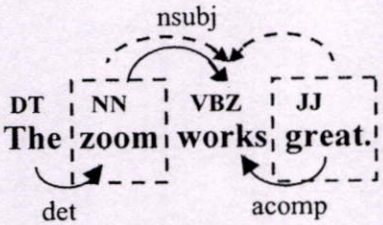
f_{13} (ประเภทความสัมพันธ์ = “sibling”, yes) ค่าถ่วงน้ำหนัก = 3.663

f_{14} (ประเภทความสัมพันธ์ = “sibling”, no) ค่าถ่วงน้ำหนัก = 1.795

f_{15} (ประเภทความสัมพันธ์ = “child”, yes) ค่าถ่วงน้ำหนัก = 5.605

f_{16} (ประเภทความสัมพันธ์ = “child”, no) ค่าถ่วงน้ำหนัก = 6.599

ตัวอย่าง “*The zoom works great.*” ซึ่งมีความสัมพันธ์แบบพึ่งพาของคำในประโยคดังนี้



จากประโยคดังกล่าว มีคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น = [*zoom, great*] สามารถสกัดคุณสมบัติเพื่อใช้ในการสกัดคุณลักษณะสินค้าและความเห็นได้ดังนี้

f_1 (คำคุณลักษณะสินค้า = “zoom”)

f_2 (คำแสดงความเห็น = “great”)

f_3 (เส้นทางความสัมพันธ์แบบพึ่งพา = “NNVBJJ”)

f_4 (ประเภทความสัมพันธ์ = “sibling”)

สามารถคำนวณความน่าจะเป็นแบบมีเงื่อนไขได้ดังนี้

$$p_{\max}(a \mid b) = \frac{1}{Z(b)} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)}$$

เมื่อ

$$Z(b) = \sum_{f} \prod_{j} \alpha_j^{f_j(a,b)}$$

$$p(yes \mid b) = \frac{1}{Z(b)} \prod_{j=1}^{4} \alpha_j^{f_j(yes,b)}$$

$$P(yes|b) = \frac{3.663^1 \text{ x } 5.605^1 \text{ x } 2.728^1 \text{ x } 3.663^1}{Z(b)}$$

$$p(no \mid b) = \frac{1}{Z(b)} \prod_{j=1}^{4} \alpha_j^{f_j(no,b)}$$

$$P(no|b) = \frac{0^0 \text{ x } 9.312^1 \text{ x } 2.795^1 \text{ x } 1.795^1}{Z(b)}$$

เมื่อ

$$Z(b) = \prod_{j} \alpha_j^{f_j(yes,b)} + \prod_{j} \alpha_j^{f_j(no,b)}$$

$$Z(b) = (3.663^1 \text{ x } 5.605^1 \text{ x } 2.728^1 \text{ x } 3.663^1) + (0^0 \text{ x } 9.312^1 \text{ x } 2.795^1 \text{ x } 1.795^1)$$

$$P(yes|b) = \frac{205.1605}{(205.1605 + 46.7185)}$$

$$P(yes|b) = 0.8145$$

$$P(no|b) = \frac{46.7185}{(205.1605 + 46.7185)}$$

$$P(no|b) = 0.1855$$

เมื่อพิจารณาจากค่าความน่าจะเป็นแบบมีเงื่อนไขที่ให้ค่าสูงสุดในบริบท *b* คือ 0.8145 ดังนั้น [*zoom, great*] จะถูกจำแนกว่าเป็นคู่คุณลักษณะสินค้าและความเห็น

3.3 คุณสมบัติสำหรับการสกัดคุณลักษณะสินค้าและความเห็น

ในหัวข้อนี้จะอธิบายรายละเอียดของคุณสมบัติและการสกัดคุณสมบัติที่ใช้ในส่วนของการเรียนรู้และส่วนของการสกัดคุณลักษณะสินค้าและความเห็น ซึ่งมีรายละเอียดดังนี้

3.3.1 ประเภทของคุณสมบัติ

สำหรับคุณสมบัติที่ใช้ในการจำแนกคู่ของคุณลักษณะสินค้าและความเห็นสำหรับงานวิจัยในครั้งนี้ใช้ข้อสนเทศที่สกัดได้จากความสัมพันธ์แบบพึ่งพาของคำในประโยค ได้แก่

1. คำที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น คือ คำที่ประกอบด้วยคำที่พิจารณาว่าเป็นคุณลักษณะสินค้า และคำที่พิจารณาว่าเป็นความเห็น

ตัวอย่างการสร้างฟังก์ชันคุณสมบัติคำ เช่น

$$f(a,b) = \begin{cases} 1 & \text{ถ้า ผลทำนาย } a = \text{yes และคำคุณลักษณะสินค้า } b = \text{picture} \\ 0 & \text{อื่นๆ} \end{cases}$$

$$f(a,b) = \begin{cases} 1 & \text{ถ้า ผลทำนาย } a = \text{yes และคำแสดงความเห็น } b = \text{good} \\ 0 & \text{อื่นๆ} \end{cases}$$

2. เส้นทางความสัมพันธ์แบบพึ่งพา คือ เส้นทางระหว่างความสัมพันธ์ของคำที่ต้องการสกัด ถือเป็นข้อสนเทศที่สำคัญสำหรับการสกัดคุณลักษณะสินค้าและความเห็น สำหรับเส้นทางที่ใช้เป็นเส้นทางของหมวดคำจากคุณลักษณะสินค้าไปยังความเห็น ซึ่งพิจารณาจากความสัมพันธ์แบบพึ่งพา

ตัวอย่างการสร้างฟังก์ชันคุณสมบัติเส้นทางความสัมพันธ์แบบพึ่งพา เช่น

$$f(a,b) = \begin{cases} 1 & \text{ถ้า ผลทำนาย } a = \text{yes และเส้นทางระหว่างคุณลักษณะสินค้าและความเห็น } b = \text{NNJJ} \\ 0 & \text{อื่นๆ} \end{cases}$$

3. ประเภทความสัมพันธ์ เป็น ลักษณะความสัมพันธ์ที่เกิดขึ้นระหว่างคุณลักษณะสินค้าและความเห็น ในงานวิจัยนี้ใช้ประเภทความสัมพันธ์ที่สรุปดังตารางที่ 3.1 ตัวอย่างการสร้างฟังก์ชันคุณสมบัติประเภทความสัมพันธ์ เช่น

$$f(a,b) = \begin{cases} 1 & \text{ถ้า ผลทำนาย } a = \text{yes และความสัมพันธ์ของคุณลักษณะสินค้าและความเห็น } b = \text{parent} \\ 0 & \text{อื่นๆ} \end{cases}$$

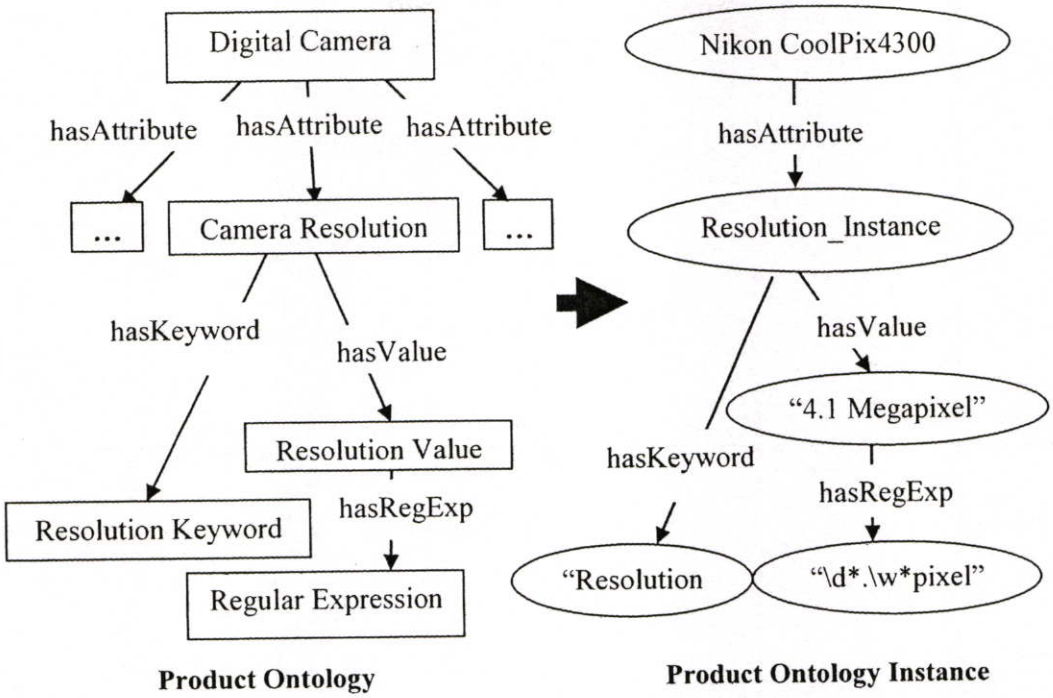
3.3.2 การสกัดคุณสมบัติ

การสกัดคุณสมบัติในส่วนของการเรียนรู้ จะสกัดจากคู่คุณลักษณะสินค้าและความเห็นในคลังข้อมูลการวิจารณ์สินค้าที่มีการกำกับคู่คุณลักษณะสินค้าและความเห็นแล้ว แต่การสกัดคุณสมบัติในส่วนของการสกัดคุณลักษณะสินค้าและความเห็น จะสกัดจากคู่ที่คาดว่าจะเป็นคุณลักษณะสินค้าและความเห็น ซึ่งคุณสมบัติที่สกัดได้จะประกอบด้วย คำที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น คุณสมบัติเส้นทางความสัมพันธ์แบบพึ่งพา และคุณสมบัติประเภทความสัมพันธ์ซึ่งได้กล่าวในหัวข้อ 3.3.1 รายละเอียดของการสกัดคุณสมบัติ สรุปได้ดังนี้

3.3.2.1 การสกัดคำที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น โดยทั่วไปการวิจารณ์สินค้าอาจจะใช้คำที่ต่างกันในกล่าวถึงคุณลักษณะสินค้าเดียวกัน เช่น ใช้คำว่า *"pic", "picture", "photo"* หรือ *"image"* ในการกล่าวถึงคุณลักษณะสินค้าที่เกี่ยวกับรูปภาพ ดังนั้นเพื่อแก้ปัญหาที่เกิดขึ้นในลักษณะนี้ ในการสกัดคำที่คาดว่าเป็นคุณลักษณะสินค้า งานวิจัยนี้ใช้ออนโทโลยีในการจัดกลุ่มคำที่ต่างกันแต่มีความหมายเหมือนกัน ตัวอย่างเช่น คุณลักษณะสินค้าที่เกี่ยวกับรูปภาพ จะใช้คำเพียงคำเดียว คือ "*picture*" แทนคำ *"pic", "picture", "photo"* และ *"image"* โดยออนโทโลยีที่ใช้เป็นออนโทโลยีแบบขึ้นอยู่กับโดเมน ที่ผู้วิจัยสร้างขึ้น โดยประยุกต์มาจาก Product Ontology [33] และสร้างจากข้อมูลคำอธิบายสินค้าที่ได้จากเว็บไซต์ขายสินค้าโดยพิจารณาร่วมกับคำในประโยคการวิจารณ์สินค้า โครงสร้างของออนโทโลยีที่ใช้ในงานวิจัยนี้ แสดงดังรูปที่ 3.14 คำที่คาดว่าเป็นคุณลักษณะสินค้า จะใช้คำสำคัญที่ได้จากออนโทโลยีแทนการใช้คำคุณลักษณะสินค้าที่ได้จากขั้นตอนการสกัดคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น ส่วนคำที่คาดว่าเป็นความเห็นจะใช้คำแสดงความเห็นที่ได้จากขั้นตอนการสกัดคู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น

3.3.2.2 การสกัดเส้นทางความสัมพันธ์แบบพึ่งพา การสกัดเส้นทางความสัมพันธ์แบบพึ่งพา จะใช้สายโซ่ความสัมพันธ์แบบพึ่งพาที่ได้จากขั้นตอนการวิเคราะห์ความสัมพันธ์แบบพึ่งพา โดยคุณสมบัติเส้นทางความสัมพันธ์แบบพึ่งพา คือหมวดคำตั้งแต่คำคุณลักษณะสินค้าไปจนถึงคำแสดงความเห็น โดยหมวดคำที่ใช้จะใช้เพียงหมวดคำหลักๆ เช่น NN, JJ, RB และ VB

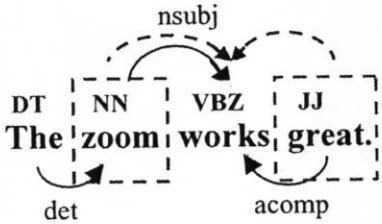
3.3.2.3 การสกัดประเภทความสัมพันธ์ สำหรับการสกัดประเภทความสัมพันธ์ จะเป็นการระบุประเภทความสัมพันธ์ระหว่างคุณลักษณะสินค้าและความเห็น โดยจะเปรียบเทียบเส้นทางความสัมพันธ์แบบพึ่งพากับกฎที่สร้างขึ้นของความสัมพันธ์แต่ละประเภท



รูปที่ 3.14 โครงสร้างของออนโทโลยีและตัวอย่างอินสแตนซ์

3.3.3 ตัวอย่างการสกัดคุณสมบัติ

การสกัดคุณสมบัติจะดำเนินการทั้งในส่วนการเรียนรู้ รวมไปถึงส่วนการสกัดคุณลักษณะสินค้าและความเห็นตัวอย่างการสกัดคุณสมบัติจากประโยคการวิจารณ์สินค้า *"The zoom works great."* คู่ที่คาดว่าเป็นคุณลักษณะสินค้าและความเห็น = [*zoom, great*]



จากประโยคการวิจารณ์สินค้าข้างบนสามารถสกัดคุณสมบัติเพื่อใช้ในการสกัดคุณลักษณะสินค้าและความเห็น ได้ดังนี้

f_1 (คำคุณลักษณะสินค้า = "zoom")

f_2 (คำแสดงความเห็น = "great")

f_3 (เส้นทางความสัมพันธ์แบบพึ่งพา = "NNVBJJ")

f_4 (ประเภทความสัมพันธ์ = "sibling")

บทที่ 4
การทดลองประสิทธิภาพการทำงาน

ในการพิจารณาประสิทธิภาพของการสกัดคุณลักษณะสินค้าและความเห็นจากการวิจารณ์สินค้า งานวิจัยนี้วัดประสิทธิภาพของวิธีการ โดยพิจารณาจากค่าความระลึก ค่าความเที่ยงตรง และค่าเอฟ ทำการทดลองเปรียบเทียบประสิทธิภาพของวิธีการสกัดคุณลักษณะสินค้าและความเห็นที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพากับวิธีการที่ผ่านมาที่ใช้คำคุณศัพท์ที่ใกล้กัน [2] และการใช้กฎ [4] โดยมีรายละเอียดในการออกแบบการทดลอง ข้อมูลที่ใช้ในการทดลอง และผลของการทดลองดังต่อไปนี้

4.1 การออกแบบการทดลอง

งานวิจัยนี้ใช้วิธีการทดสอบแบบการตรวจสอบไขว้ (Cross-Validation) ซึ่งเป็นวิธีการทดสอบที่จะทำการแบ่งชุดเอกสารออกเป็น k ชุด แล้วทำการทดสอบ โดยการทดสอบแต่ละครั้งจะใช้เอกสาร 1 ชุด เป็นชุดทดสอบ ส่วนที่เหลือ k-1 ชุด จะใช้เป็นชุดสำหรับการเรียนรู้ สับเปลี่ยนชุดทดสอบโดยใช้ทุกชุดเป็นชุดทดสอบจนครบ การทดสอบแต่ละครั้งถือเป็นการทดลองหนึ่งครั้ง ซึ่งเท่ากับได้ทำการทดลอง k ครั้ง แล้วนำผลที่วัดได้จากการทดลองทั้ง k ครั้งมาหาค่าเฉลี่ยของตัววัดประสิทธิภาพ โดยการทดลองสกัดคุณลักษณะสินค้าและความคิดเห็นจากข้อมูลการวิจารณ์สินค้าของงานวิจัยนี้ ดำเนินการทดลองโดยใช้ k เท่ากับ 5 คือการตรวจสอบไขว้ 5 ทบ ซึ่งทำการทดลองสกัดคุณลักษณะสินค้าและความคิดเห็น 5 รอบ แต่ละรอบจะใช้ชุดข้อมูลย่อย 4 ชุดเป็นชุดสำหรับการเรียนรู้ และใช้ชุดข้อมูลย่อยที่เหลืออีกหนึ่งชุดเป็นชุดข้อมูลทดสอบ ซึ่งในแต่ละรอบจะทำการวัดประสิทธิภาพซึ่งได้แก่ ค่าความระลึก ค่าความเที่ยงตรง และค่าเอฟ หลังจากนั้นจะคำนวณหาค่าเฉลี่ยรวมของค่าความระลึก ค่าความเที่ยงตรง และค่าเอฟ โดยมีรายละเอียดดังนี้

รอบที่ 1: ข้อมูลย่อยชุดที่ 1 เป็นข้อมูลทดสอบ ส่วนข้อมูลย่อยที่เหลือเป็นข้อมูลสำหรับการเรียนรู้

Dataset1 Dataset2 Dataset3 Dataset4 Dataset5

ค่าความระลึก = r_1, ค่าความเที่ยงตรง = p_1, ค่าเอฟ = f_1

รอบที่ 2: ข้อมูลย่อยชุดที่ 2 เป็นข้อมูลทดสอบ ส่วนข้อมูลย่อยที่เหลือเป็นข้อมูลสำหรับการเรียนรู้

Dataset1 **Dataset2** Dataset3 Dataset4 Dataset5

ค่าความระลึก = r_2, ค่าความเที่ยงตรง = p_2, ค่าเอฟ = f_2

รอบที่ 3: ข้อมูลย่อยชุดที่ 3 เป็นข้อมูลทดสอบ ส่วนข้อมูลย่อยที่เหลือเป็นข้อมูลสำหรับการเรียนรู้

Dataset1 Dataset2 **Dataset3** Dataset4 Dataset5

ค่าความระลึก = r_3, ค่าความเที่ยงตรง = p_3, ค่าเอฟ = f_3

รอบที่ 4: ข้อมูลย่อยชุดที่ 4 เป็นข้อมูลทดสอบ ส่วนข้อมูลย่อยที่เหลือเป็นข้อมูลสำหรับการเรียนรู้

Dataset1 Dataset2 Dataset3 **Dataset4** Dataset5

ค่าความระลึก = r_4, ค่าความเที่ยงตรง = p_4, ค่าเอฟ = f_4

รอบที่ 5: ข้อมูลย่อยชุดที่ 5 เป็นข้อมูลทดสอบ ส่วนข้อมูลย่อยที่เหลือเป็นข้อมูลสำหรับการเรียนรู้

Dataset1 Dataset2 Dataset3 Dataset4 **Dataset5**

ค่าความระลึก = r_5, ค่าความเที่ยงตรง = p_5, ค่าเอฟ = f_5

หลังจากที่ได้ค่าประสิทธิภาพในแต่ละรอบ ซึ่งประกอบด้วยค่าความระลึก r_1, r_2, r_3, r_4, r_5 ค่าความเที่ยงตรง p_1, p_2, p_3, p_4, p_5 และค่าเอฟ f_1, f_2, f_3, f_4, f_5 แล้วทำการคำนวณหาค่าเฉลี่ยของแต่ละค่า ซึ่งสามารถคำนวณได้ดังสมการต่อไปนี้

$$\text{ค่าเฉลี่ยของค่าความระลึก} = (r_1 + r_2 + \ldots + r_K) / k$$
$$\text{ค่าเฉลี่ยของค่าความเที่ยงตรง} = (p_1 + p_2 + \ldots + p_K) / k$$
$$\text{ค่าเฉลี่ยของค่าเอฟ} = (f_1 + f_2 + \ldots + f_K) / k$$

ในการหาประสิทธิภาพของวิธีการที่นำเสนอในงานวิจัยนี้จะพิจารณาทั้งในแง่ความถูกต้องในการสกัดคุณลักษณะสินค้าและความเห็น และในแง่ของจำนวนประโยคที่สามารถสกัดได้ ดังนั้นจึงแบ่งการทดลองออกเป็น 2 ส่วน คือ การทดลองสกัดคู่คุณลักษณะสินค้าและความเห็นจากข้อมูลการวิจารณ์สินค้า และการทดลองสกัดประโยคความเห็นจากข้อมูลการวิจารณ์สินค้า

4.2 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในการทดลองสำหรับการวิจัยในครั้งนี้ คือข้อมูลการวิจารณ์สินค้าที่อยู่ในรูปแบบอิสระ ซึ่งจะเขียนบรรยายเป็นประโยค โดยไม่มีการแบ่งด้านของความเห็นว่าเป็นด้านบวกหรือด้านลบ ตัวอย่างแสดงดังรูปที่ 4.1

ข้อมูลการวิจารณ์สินค้าที่ใช้ในการทดลองประกอบด้วยข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่และข้อมูลการวิจารณ์กล้องดิจิทัล ซึ่งแบ่งออกเป็น 4 ชุด มีรายละเอียดดังนี้

This review is from: **Canon PowerShot SD1200IS 10 MP Digital Camera with 3x Optical Image Stabilized Zoom and 2.5-inch LCD (Silver) (Electronics)**

I bought the SD110 (3.2MP) back in 2004 and decided it was time to upgrade this year as it was showing some age. The latest version of that same line is the SD1200. I have had it for a few weeks now and have a few observations. The SD1200 is much faster from the time you turn it on until it is ready to shoot as compared to the SD110. Next, the screen is far better than the one on the SD110. Very bright and readable even in sunlight. The camera itself is smaller than the SD110 as you might expect but does feel a little cheaper. So far I have taken the SD1200 to three main events. A birthday party, K-4 graduation, backyard play time. The birthday part pictures did not turn out well in automatic mode. We were indoors (Pump-it-up) and the lighting was probably not the best. The images were blurry for the most part. I had the same issue at the second event (K-4 graduation) but this time I switched to manual mode and used the "indoor" setting. This greatly improved the picture. Finally, the outdoor shots turned out looking wonderful in automatic mode. With the SD110 the automatic mode was always better than any of the manual settings. It seems with the SD1200 that is not the case. I will continue to explore the settings/features of the camera. I was expecting the SD1200 to be far superior to my 5 year old SD110 but so far the pictures themselves have not turned out markedly better.

รูปที่ 4.1 ตัวอย่างการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ amazon.com

1. ชุดข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu ข้อมูลชุดนี้มีทั้งหมด 591 ประโยค ซึ่งประกอบด้วยประโยคที่เป็นประโยคความเห็น 377 ประโยค และประโยคที่ไม่เป็นประโยคความเห็น 214 ประโยค และมีคู่ของคุณลักษณะสินค้าและความเห็นจำนวน 531คู่

2. ชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu ข้อมูลชุดนี้มีทั้งหมด 913 ประโยค ซึ่งประกอบด้วยประโยคที่เป็นประโยคความเห็น 334 ประโยค และประโยคที่ไม่เป็นประโยคความเห็น 579 ประโยค และมีคู่ของคุณลักษณะสินค้าและความเห็นจำนวน 527 คู่

3. ชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon ข้อมูลชุดนี้มีทั้งหมด 1,092 ประโยค ซึ่งประกอบด้วยประโยคที่เป็นประโยคความเห็น 669 ประโยค และประโยคที่ไม่เป็นประโยคความเห็น 423 ประโยค และมีคู่ของคุณลักษณะสินค้าและความเห็นจำนวน 980 คู่

4. ชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon ข้อมูลชุดนี้มีทั้งหมด 2,005 ประโยค ซึ่งประกอบด้วยประโยคที่เป็นประโยคความเห็น 1,003 ประโยค และประโยคที่ไม่เป็นประโยคความเห็น 1,002 ประโยค และมีคู่ของคุณลักษณะสินค้าและความเห็นจำนวน 1,507 คู่

ข้อมูลชุดที่ 1 และชุดที่ 2 เป็นข้อมูลการวิจารณ์สินค้าที่ใช้ในงานวิจัยของ Hu and Liu [2] ข้อมูลชุดที่ 3 เป็นข้อมูลการวิจารณ์สินค้าจากเว็บไซต์อเมซอน และข้อมูลชุดที่ 4 เป็นข้อมูลการวิจารณ์สินค้ารวมจากชุดที่ 2 และชุดที่ 3 ซึ่งข้อมูลทั้งสี่ชุดจะอยู่ในรูปแบบของแฟ้มข้อความ และประกอบด้วยประโยคที่เป็นประโยคแสดงความเห็นและประโยคไม่แสดงความเห็น โดยข้อมูลทั้งหมดจะถูกใช้ในการทดลองสกัดคุณลักษณะสินค้าและความเห็น ด้วยวิธีการที่นำเสนอในงานวิจัยนี้ที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา และวิธีการของงานวิจัยที่ผ่านมา ได้แก่ วิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎ นอกจากนี้ข้อมูลชุดที่ 1

และชุดที่ 2 จะถูกใช้ในการทดลองสกัดประโยคความเห็น ด้วยวิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา เพื่อเปรียบเทียบกับผลการทดลองในการสกัดประโยคความเห็นจากงานวิจัยของ Hu and Liu [2]

4.3 ผลการทดลองชุดข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu

ในการทดลองสกัดคุณลักษณะสินค้าและความเห็นกับชุดข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu ได้แบ่งข้อมูลเพื่อใช้ในการเรียนรู้และใช้ในการทดสอบสำหรับทำการตรวจสอบไขว้ 5 ทบ มีจำนวนประโยคที่ใช้สำหรับการเรียนรู้และการทดสอบในแต่ละชุดการทดลองแสดงดังตารางที่ 4.1 และจำนวนคู่คุณลักษณะสินค้าและความเห็นสำหรับการเรียนรู้และการทดสอบในแต่ละชุดการทดลองแสดงดังตารางที่ 4.2

ตารางที่ 4.1 จำนวนประโยคของข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu ที่ใช้ในการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชุดที่	จำนวนประโยคในการเรียนรู้		จำนวนประโยคในการทดสอบ		รวม
	แสดงความเห็น	ไม่แสดงความเห็น	แสดงความเห็น	ไม่แสดงความเห็น	
1	313	160	64	54	591
2	311	162	66	52	591
3	303	170	74	44	591
4	304	169	73	45	591
5	307	166	70	48	591

ตารางที่ 4.2 จำนวนคู่คุณลักษณะสินค้าและความเห็นของข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu สำหรับการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชุดที่	จำนวนคู่คุณลักษณะสินค้าและความเห็น	
	การเรียนรู้	การทดสอบ
1	432	99
2	426	105
3	419	112
4	428	103
5	423	108

ผลการทดลองสกัดคุณลักษณะสินค้าและความเห็นจากข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu ด้วยวิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา วิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน และวิธีการที่ใช้กฎ แสดงได้ดังตารางที่ 4.3 ถึง ตารางที่ 4.6

ตารางที่ 4.3 ค่าความระลึกในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**57.58**	**56.19**	**58.04**	**59.22**	**66.67**
การใช้กฎ	50.51	42.86	47.32	50.49	51.85
การใช้คำคุณศัพท์ที่ใกล้กัน	49.50	40.95	45.54	48.54	48.15

ตารางที่ 4.4 ค่าความเที่ยงตรงในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**61.29**	**60.21**	**65.00**	**61.62**	**63.72**
การใช้กฎ	37.04	35.43	38.41	42.28	40.58
การใช้คำคุณศัพท์ที่ใกล้กัน	45.37	46.24	49.04	52.63	51.49

ตารางที่ 4.5 ค่าเอฟในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**59.38**	**58.13**	**61.32**	**60.40**	**65.16**
การใช้กฎ	42.74	38.79	42.40	46.02	45.53
การใช้คำคุณศัพท์ที่ใกล้กัน	47.34	43.43	47.22	50.51	49.76

ผลการทดลองจากตารางที่ 4.3 พบว่าวิธีการที่พัฒนาขึ้นมีค่าความระลึกของแต่ละชุดการทดลองสูงสุด รองลงมาคือวิธีการที่ใช้กฎ ส่วนวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน จะมีค่า

ความระลึกของแต่ละชุดการทดลองต่ำสุด แต่เมื่อพิจารณาค่าความเที่ยงตรง จากตารางที่ 4.4 พบว่า วิธีการที่พัฒนาขึ้นมีค่าความเที่ยงตรงของแต่ละชุดการทดลองสูงสุด ส่วนวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันจะมีค่าความเที่ยงตรงของแต่ละชุดการทดลองสูงกว่าวิธีการที่ใช้กฎ และเมื่อพิจารณาค่าเอฟหรือค่าความแม่นยำโดยรวมจากตารางที่ 4.5 พบว่า วิธีการที่พัฒนาขึ้นมีค่าเอฟของแต่ละชุดการทดลองสูงสุด รองลงมาคือวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน และวิธีการที่ใช้กฎจะมีค่าเอฟของแต่ละชุดการทดลองต่ำสุด

ตารางที่ 4.6 ค่าเฉลี่ยการวัดประสิทธิภาพในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีกับข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu

วิธี	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**59.54**	**62.37**	**60.88**
การใช้กฎ	48.60	38.75	43.09
การใช้คำคุณศัพท์ที่ใกล้กัน	46.53	48.95	47.65

จากตารางที่ 4.6 สรุปได้ว่า วิธีการที่ได้พัฒนาขึ้นมีค่าเฉลี่ยความระลึกสูงที่สุด เท่ากับ 59.54 รองลงมาคือ วิธีการที่ใช้กฎมีค่าเฉลี่ยความระลึก เท่ากับ 48.60 และวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันมีค่าเฉลี่ยความระลึกต่ำสุด เท่ากับ 46.53 และเมื่อเปรียบเทียบค่าความเที่ยงตรง พบว่า วิธีการที่พัฒนาขึ้นมีค่าเฉลี่ยความเที่ยงตรงสูงสุด เท่ากับ 62.37 รองลงมาคือวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันมีค่าเฉลี่ยความเที่ยงตรง 48.95 และวิธีการที่ใช้กฎมีค่าเฉลี่ยความเที่ยงตรงต่ำสุด 38.75 นอกจากนี้วิธีการที่พัฒนาขึ้นยังมีค่าเฉลี่ยเอฟหรือค่าวัดความแม่นยำโดยรวมสูงกว่าวิธีการทั้งสอง ซึ่งแสดงให้เห็นว่าวิธีการที่พัฒนาขึ้นโดยการใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพามีประสิทธิภาพในการสกัดคุณลักษณะสินค้าและความเห็นจากชุดข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu สูงกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎ

สำหรับผลการทดลองสกัดประโยคความเห็นจากชุดข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu ด้วยวิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา ในแต่ละชุดการทดลองแสดงดังตารางที่ 4.7 และเมื่อเปรียบเทียบกับผลการทดลองในการสกัดประโยคความเห็นด้วยวิธีการใช้คำคุณศัพท์ที่ใกล้กันในงานวิจัยของ Hu and Liu [2] แสดงดังตารางที่ 4.8 พบว่าวิธีการที่นำเสนอในงานวิจัยนี้มีค่าความระลึกสูงกว่าวิธีการใช้คำคุณศัพท์ที่ใกล้กัน

แต่มีค่าความเที่ยงตรงต่ำกว่า อย่างไรก็ตามเมื่อพิจารณาค่าเอฟ ซึ่งเป็นค่าวัดความแม่นยำโดยรวม พบว่าวิธีการที่นำเสนอมีค่าสูงกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันในงานวิจัยของ Hu and Liu

ตารางที่ 4.7 ค่าประสิทธิภาพในการสกัดประโยคความเห็นของวิธีการที่นำเสนอกับชุดข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu

ชุดที่	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
1	85.94	79.12	82.39
2	80.30	78.82	79.56
3	79.73	74.74	77.15
4	78.08	76.60	77.33
5	82.86	78.65	80.70
เฉลี่ย	81.38	77.59	79.43

ตารางที่ 4.8 ค่าประสิทธิภาพในการสกัดประโยคความเห็นของแต่ละวิธีกับข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu

วิธี	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**81.38**	**77.59**	**79.43**
การใช้คำคุณศัพท์ที่ใกล้กัน	67.50	81.50	73.84

4.4 ผลการทดลองชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu

ในการทดลองสกัดคุณลักษณะสินค้าและความเห็นกับชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu ได้แบ่งข้อมูลเพื่อใช้ในการเรียนรู้และใช้ในการทดสอบสำหรับทำการตรวจสอบไขว้ 5 ทบ มีจำนวนประโยคที่ใช้สำหรับการเรียนรู้และการทดสอบในแต่ละชุดการทดลองดังตารางที่ 4.9 และจำนวนคู่คุณลักษณะสินค้าและความเห็นสำหรับการเรียนรู้และการทดสอบในแต่ละชุดการทดลองดังตารางที่ 4.10

ตารางที่ 4.9 จำนวนประโยคของข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu ที่ใช้ในการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชุดที่	จำนวนประโยคในการเรียนรู้		จำนวนประโยคในการทดสอบ		รวม
	แสดงความเห็น	ไม่แสดงความเห็น	แสดงความเห็น	ไม่แสดงความเห็น	
1	263	468	72	110	913
2	270	461	65	117	913
3	271	460	64	118	913
4	277	454	58	124	913
5	272	459	63	119	913

ตารางที่ 4.10 จำนวนคู่คุณลักษณะสินค้าและความเห็นของข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu สำหรับการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชุดที่	จำนวนคู่คุณลักษณะสินค้าและความเห็น	
	การเรียนรู้	การทดสอบ
1	411	116
2	415	112
3	425	102
4	425	102
5	422	105

ผลการทดลองสกัดคุณลักษณะสินค้าและความเห็นจากข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu แสดงได้ดังตารางที่ 4.11 ถึง ตารางที่ 4.14

ตารางที่ 4.11 ค่าความระลึกในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	49.14	**46.43**	**48.04**	**47.06**	44.76
การใช้กฎ	**50.86**	43.75	41.18	45.10	**51.43**
การใช้คำคุณศัพท์ที่ใกล้กัน	50.00	41.07	40.20	45.10	48.57

ตารางที่ 4.12 ค่าความเที่ยงตรงในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**71.25**	**54.74**	**55.06**	**44.86**	**55.95**
การใช้กฎ	26.82	24.87	20.19	21.50	25.71
การใช้คำคุณศัพท์ที่ใกล้กัน	36.94	33.58	25.31	31.51	32.08

ตารางที่ 4.13 ค่าเอฟในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**58.16**	**50.24**	**51.31**	**45.93**	**49.74**
การใช้กฎ	35.12	31.72	27.10	29.11	34.29
การใช้คำคุณศัพท์ที่ใกล้กัน	42.49	36.95	31.06	37.10	38.64

ตารางที่ 4.14 ค่าเฉลี่ยการวัดประสิทธิภาพในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีกับข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu

วิธี	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**47.09**	**56.37**	**51.08**
การใช้กฎ	46.46	23.82	31.47
การใช้คำคุณศัพท์ที่ใกล้กัน	44.98	31.88	37.25

ผลการทดลองจากตารางที่ 4.11 พบว่าวิธีการที่พัฒนาขึ้นมีค่าความระลึกสูงสุดในชุดการทดลองที่ 2 ถึง 4 แต่ชุดการทดลองที่ 1 และ 5 วิธีการที่ใช้กฎจะมีค่าความระลึกสูงสุด แต่เมื่อพิจารณาค่าความเที่ยงตรง จากตารางที่ 4.12 พบว่า วิธีการที่พัฒนาขึ้นมีค่าความเที่ยงตรงของแต่ละชุดการทดลองสูงสุด ส่วนวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันจะมีค่าความเที่ยงตรงของแต่ละชุดการทดลองสูงกว่าวิธีการที่ใช้กฎ และเมื่อพิจารณาค่าเอฟหรือค่าความแม่นยำโดยรวมจากตารางที่ 4.13 พบว่า วิธีการที่พัฒนาขึ้นมีค่าเอฟของแต่ละชุดการทดลองสูงสุด รองลงมาคือวิธีการที่ใช้

คำคุณศัพท์ที่ใกล้กัน และวิธีการที่ใช้กฎจะมีค่าเอฟของแต่ละชุดการทดลองต่ำสุด และจากค่าเฉลี่ยในตารางที่ 4.14 สรุปได้ว่า วิธีการที่ได้พัฒนาขึ้นมีค่าเฉลี่ยความระลึกสูงที่สุด เท่ากับ 47.09 รองลงมาคือ วิธีการที่ใช้กฎมีค่าเฉลี่ยความระลึก เท่ากับ 46.46 และวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันมีค่าเฉลี่ยความระลึกต่ำสุด เท่ากับ 44.98 และเมื่อเปรียบเทียบค่าความเที่ยงตรง พบว่า วิธีการที่พัฒนาขึ้นมีค่าเฉลี่ยความเที่ยงตรงสูงสุด เท่ากับ 56.37 รองลงมาคือวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันมีค่าเฉลี่ยความเที่ยงตรง 31.88 และวิธีการที่ใช้กฎมีค่าเฉลี่ยความเที่ยงตรงต่ำสุด 23.82 นอกจากนี้วิธีการที่พัฒนาขึ้นยังมีค่าเฉลี่ยเอฟหรือค่าวัดความแม่นยำโดยรวมสูงกว่าวิธีการทั้งสอง ซึ่งแสดงให้เห็นว่าวิธีการที่พัฒนาขึ้นโดยการใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพามีประสิทธิภาพในการสกัดคุณลักษณะสินค้าและความเห็นจากชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu สูงกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎ

สำหรับผลการทดลองสกัดประโยคความเห็นจากชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu ด้วยวิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา ในแต่ละชุดการทดลองแสดงดังตารางที่ 4.15 และเมื่อเปรียบเทียบกับผลการทดลองในการสกัดประโยคความเห็นด้วยวิธีการใช้คำคุณศัพท์ที่ใกล้กันในงานวิจัยของ Hu and Liu [2] แสดงดังตารางที่ 4.16 พบว่าวิธีการที่นำเสนอในงานวิจัยนี้มีค่าความระลึก ค่าความเที่ยงตรง และค่าเอฟสูงกว่าวิธีการใช้คำคุณศัพท์ที่ใกล้กันในงานวิจัยของ Hu and Liu

ตารางที่ 4.15 ค่าประสิทธิภาพในการสกัดประโยคความเห็นของวิธีการที่นำเสนอกับชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu

ชุดที่	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
1	70.42	76.00	73.11
2	78.13	74.02	76.01
3	71.43	76.42	73.84
4	64.91	65.44	65.18
5	67.74	72.31	69.95
เฉลี่ย	70.62	72.84	71.62

ตารางที่ 4.16 ค่าประสิทธิภาพในการสกัดประโยคความเห็นของแต่ละวิธีกับข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu

วิธี	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**70.62**	**72.84**	**71.62**
การใช้คำคุณศัพท์ที่ใกล้กัน	67.65	59.85	63.51

4.5 ผลการทดลองชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon

ในการทดลองสกัดคุณลักษณะสินค้าและความเห็นกับชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon ได้แบ่งข้อมูลเพื่อใช้ในการเรียนรู้และใช้ในการทดสอบสำหรับทำการตรวจสอบไขว้ 5 ทบ มีจำนวนประโยคที่ใช้สำหรับการเรียนรู้และการทดสอบในแต่ละชุดการทดลองดังตารางที่ 4.17 และจำนวนคู่คุณลักษณะสินค้าและความเห็นสำหรับการเรียนรู้และการทดสอบในแต่ละชุดการทดลองดังตารางที่ 4.18

ตารางที่ 4.17 จำนวนประโยคของข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon ที่ใช้ในการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชุดที่	จำนวนประโยคในการเรียนรู้		จำนวนประโยคในการทดสอบ		รวม
	แสดงความเห็น	ไม่แสดงความเห็น	แสดงความเห็น	ไม่แสดงความเห็น	
1	535	339	134	84	1,092
2	536	338	133	85	1,092
3	535	339	134	84	1,092
4	534	340	135	83	1,092
5	537	337	132	86	1,092

ตารางที่ 4.18 จำนวนคู่คุณลักษณะสินค้าและความเห็นของข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon สำหรับการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชุดที่	จำนวนคู่คุณลักษณะสินค้าและความเห็น	
	การเรียนรู้	การทดสอบ
1	785	195
2	779	201
3	776	204
4	772	208
5	800	180

ผลการทดลองสกัดคุณลักษณะสินค้าและความเห็นจากข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon แสดงได้ดังตารางที่ 4.19 ถึงตารางที่ 4.22

ตารางที่ 4.19 ค่าความระลึกในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**67.18**	**70.65**	**68.14**	**65.38**	**69.44**
การใช้กฎ	58.46	58.21	57.35	56.25	58.33
การใช้คำคุณศัพท์ที่ใกล้กัน	56.41	56.72	53.92	54.32	56.67

ตารางที่ 4.20 ค่าความเที่ยงตรงในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**71.58**	**72.82**	**68.81**	**66.01**	**71.84**
การใช้กฎ	48.93	45.53	50.00	46.61	44.12
การใช้คำคุณศัพท์ที่ใกล้กัน	61.45	53.77	61.11	57.07	56.67

ตารางที่ 4.21 ค่าเอฟในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**69.31**	**71.72**	**68.47**	**65.70**	**70.62**
การใช้กฎ	53.27	51.09	53.42	50.98	50.24
การใช้คำคุณศัพท์ที่ใกล้กัน	58.82	55.21	57.29	55.67	56.67

ตารางที่ 4.22 ค่าเฉลี่ยการวัดประสิทธิภาพในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีกับข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon

วิธี	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**68.16**	**70.22**	**69.16**
การใช้กฎ	57.72	47.04	51.80
การใช้คำคุณศัพท์ที่ใกล้กัน	55.61	58.01	56.73

จากตารางที่ 4.19 พบว่าวิธีการที่พัฒนาขึ้นมีค่าความระลึกของแต่ละชุดการทดลองสูงสุด รองลงมาคือวิธีการที่ใช้กฎ แต่เมื่อพิจารณาค่าจากตารางที่ 4.20 พบว่า วิธีการที่พัฒนาขึ้นมีค่าความเที่ยงตรงของแต่ละชุดการทดลองสูงสุด ส่วนวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันจะมีค่าความเที่ยงตรงของแต่ละชุดการทดลองสูงกว่าวิธีการที่ใช้กฎ และเมื่อพิจารณาค่าเอฟหรือค่าความแม่นยำโดยรวมจากตารางที่ 4.21 พบว่า วิธีการที่พัฒนาขึ้นมีค่าเอฟของแต่ละชุดการทดลองสูงสุด รองลงมาคือวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎจะมีค่าเอฟของแต่ละชุดการทดลองต่ำสุด จากตารางที่ 4.22 สรุปได้ว่าวิธีการที่นำเสนอมีค่าเฉลี่ยความระลึกสูงสุดเท่ากับ 68.16 รองลงมาคือวิธีการที่ใช้กฎ เท่ากับ 57.72 และวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันมีค่าเฉลี่ยความระลึกต่ำสุด เท่ากับ 55.61 และเมื่อเปรียบเทียบค่าความเที่ยงตรง พบว่า วิธีการที่พัฒนาขึ้นมีค่าเฉลี่ยความเที่ยงตรงสูงสุดเท่ากับ 70.22 รองลงมาคือวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน มีค่าเฉลี่ยความเที่ยงตรง 58.01 และวิธีการที่ใช้กฎมีค่าเฉลี่ยความเที่ยงตรงต่ำสุด 47.04 นอกจากนี้วิธีการที่พัฒนาขึ้นยังมีค่าเฉลี่ยเอฟหรือค่าวัดความแม่นยำโดยรวมสูงกว่าวิธีการทั้งสอง ซึ่งแสดงให้เห็นว่าวิธีการที่พัฒนาขึ้นโดยการใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา มีประสิทธิภาพในการสกัดคุณลักษณะสินค้าและความเห็นจากชุดข้อมูลการวิจารณ์โทรศัพท์ เคลื่อนที่จากเว็บไซต์ Amazon สูงกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎ

4.6 ผลการทดลองชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon

ในการทดลองสกัดคุณลักษณะสินค้าและความเห็นกับชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon ได้แบ่งข้อมูลเพื่อใช้ในการเรียนรู้และใช้ในการทดสอบสำหรับทำการตรวจสอบไขว้ 5 ทบ มีจำนวนประโยคที่ใช้สำหรับการเรียนรู้และการทดสอบในแต่ละชุดการทดลองดังตารางที่ 4.23 และจำนวนคู่คุณลักษณะสินค้าและความเห็นสำหรับการเรียนรู้และการทดสอบในแต่ละชุดการทดลองดังตารางที่ 4.24

ตารางที่ 4.23 จำนวนประโยคของข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon ที่ใช้ในการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชุดที่	จำนวนประโยคในการเรียนรู้		จำนวนประโยคในการทดสอบ		รวม
	แสดงความเห็น	ไม่แสดงความเห็น	แสดงความเห็น	ไม่แสดงความเห็น	
1	799	805	204	197	2,005
2	787	817	216	185	2,005
3	778	826	225	176	2,005
4	773	831	230	171	2,005
5	767	837	236	165	2,005

ตารางที่ 4.24 จำนวนคู่คุณลักษณะสินค้าและความเห็นของข้อมูลการวิจารณ์กล้องดิจิทัลจางานวิจัยของ Hu and Liu และเว็บไซต์ Amazon สำหรับการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชุดที่	จำนวนคู่คุณลักษณะสินค้าและความเห็น	
	การเรียนรู้	การทดสอบ
1	1,177	330
2	1,133	374
3	1,161	346
4	1,139	368
5	1,146	361

ผลการทดลองสกัดคุณลักษณะสินค้าและความเห็นจากข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon แสดงได้ดังตารางที่ 4.25 ถึง ตารางที่ 4.28

ตารางที่ 4.25 ค่าความระลึกในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**57.27**	**57.49**	**70.52**	**66.58**	**67.87**
การใช้กฎ	46.36	47.06	60.98	54.62	57.34
การใช้คำคุณศัพท์ที่ใกล้กัน	44.85	45.19	58.96	51.90	56.51

ตารางที่ 4.26 ค่าความเที่ยงตรงในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**54.94**	**68.47**	**69.71**	**74.69**	**76.32**
การใช้กฎ	30.97	39.46	46.89	45.48	49.88
การใช้คำคุณศัพท์ที่ใกล้กัน	40.43	48.29	55.74	58.41	51.52

ตารางที่ 4.27 แสดงค่าเอฟในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีในแต่ละชุดการทดลองข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon

วิธี	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**56.08**	**62.50**	**70.11**	**70.40**	**71.85**
การใช้กฎ	37.14	42.93	53.02	49.63	53.35
การใช้คำคุณศัพท์ที่ใกล้กัน	42.53	46.69	57.30	54.96	53.90

ผลการทดลองจากตารางที่ 4.25 พบว่าวิธีการที่พัฒนาขึ้นมีค่าความระลึกของแต่ละชุดการทดลองสูงสุด รองลงมาคือวิธีการที่ใช้กฎ ส่วนวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันจะมีค่าความระลึกของแต่ละชุดการทดลองต่ำสุด แต่เมื่อพิจารณาค่าความเที่ยงตรง จากตารางที่ 4.26 พบว่าวิธีการที่พัฒนาขึ้นมีค่าความเที่ยงตรงของแต่ละชุดการทดลองสูงสุด ส่วนวิธีการที่ใช้คำคุณศัพท์ที่

ใกล้กันจะมีค่าความเที่ยงตรงของแต่ละชุดการทดลองสูงกว่าวิธีการที่ใช้กฎ และเมื่อพิจารณาค่าเอฟหรือค่าความแม่นยำโดยรวมจากตารางที่ 4.27 พบว่า วิธีการที่พัฒนาขึ้นมีค่าเอฟของแต่ละชุดการทดลองสูงสุด รองลงมาคือวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน และวิธีการที่ใช้กฎจะมีค่าเอฟของแต่ละชุดการทดลองต่ำสุด

ตารางที่ 4.28 ค่าเฉลี่ยการวัดประสิทธิภาพในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละวิธีกับข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon

วิธีการ	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	**63.94**	**68.83**	**66.19**
การใช้กฎ	53.27	42.54	47.21
การใช้คำคุณศัพท์ที่ใกล้กัน	51.48	50.87	51.08

จากตารางที่ 4.28 สรุปได้ว่า วิธีการที่ได้พัฒนาขึ้นมีค่าเฉลี่ยความระลึกสูงที่สุด เท่ากับ 63.94 รองลงมาคือ วิธีการที่ใช้กฎมีค่าเฉลี่ยความระลึก เท่ากับ 53.27 และวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันมีค่าเฉลี่ยความระลึกต่ำสุด เท่ากับ 51.48 และเมื่อเปรียบเทียบค่าความเที่ยงตรง พบว่า วิธีการที่พัฒนาขึ้นมีค่าเฉลี่ยความเที่ยงตรงสูงสุด เท่ากับ 68.83 รองลงมาคือวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน มีค่าเฉลี่ยความเที่ยงตรง 50.87 และวิธีการที่ใช้กฎมีค่าเฉลี่ยความเที่ยงตรงต่ำสุด 42.54 นอกจากนี้วิธีการที่พัฒนาขึ้นยังมีค่าเฉลี่ยเอฟหรือค่าวัดความแม่นยำโดยรวมสูงกว่าวิธีการทั้งสอง ซึ่งแสดงให้เห็นว่าวิธีการที่พัฒนาขึ้น โดยการใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา มีประสิทธิภาพในการสกัดคุณลักษณะสินค้าและความเห็นจากชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon สูงกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎ

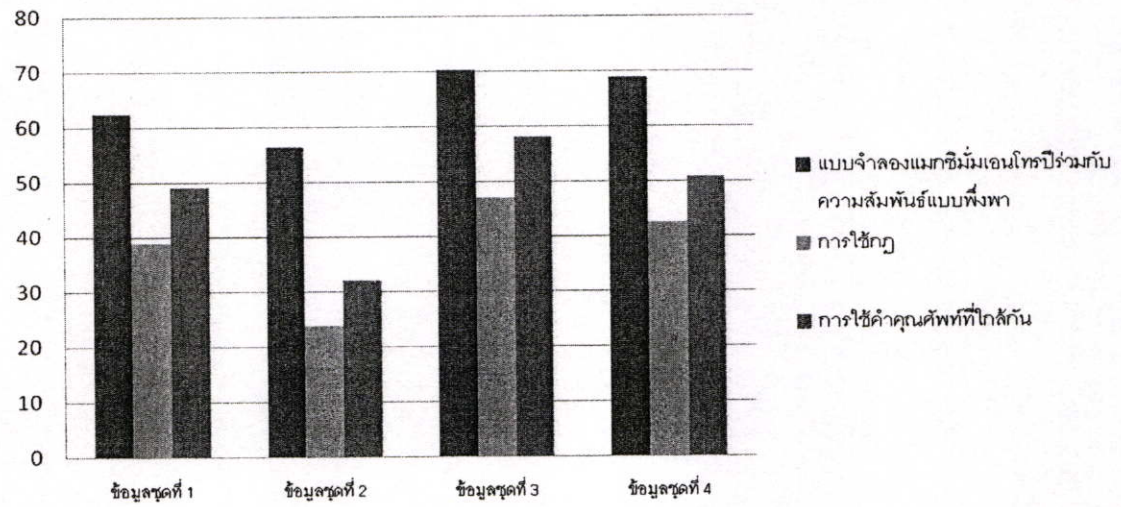
4.7 ผลสรุปเปรียบเทียบผลการทดลองทุกชุดข้อมูล

จากผลการทดลองสกัดคุณลักษณะสินค้าและความเห็นในชุดข้อมูลทุกชุด แสดงเป็นแผนภูมิแท่งในการเปรียบเทียบได้ดังรูปที่ 4.2 ถึง 4.4 และเวลาที่ใช้ในการสกัดคุณลักษณะสินค้าและความเห็นของแต่ละวิธีในชุดข้อมูลทั้ง 4 ชุด แสดงดังตารางที่ 4.29 พบว่า วิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพาใช้เวลาการประมวลผลมากกว่า

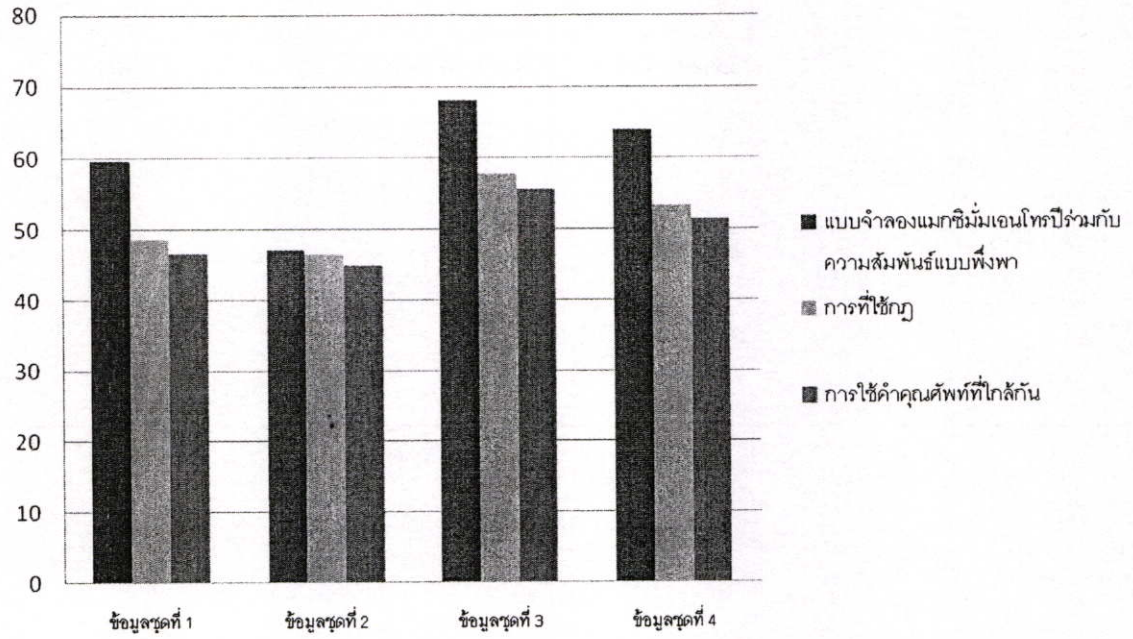
วิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎ เนื่องจากวิธีการที่นำเสนอมีการสกัดคุณสมบัติของแต่ละคู่ของคุณลักษณะสินค้าและความเห็นเพื่อใช้ในการสอนแบบจำลองแมกซิมั่มเอนโทรปี จึงใช้เวลามากกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการที่ใช้กฎที่เป็นวิธีแบบไม่มีการสอนหรือไม่มีการเรียนรู้ แต่อย่างไรก็ตามเมื่อเปรียบเทียบผลการทดลองในชุดข้อมูลทั้ง 4 ชุด พบว่าวิธีการสกัดคุณลักษณะสินค้าและความเห็นที่นำเสนอในงานวิจัยนี้มีประสิทธิภาพมากกว่าวิธีของงานวิจัยที่ผ่านมาที่ใช้คำคุณศัพท์ที่ใกล้กัน และวิธีการใช้กฎ เมื่อพิจารณา จากค่าความที่ยงตรง ค่าความระลึก และค่าเอฟที่มีค่าสูงสุด

ตารางที่ 4.29 แสดงค่าเฉลี่ยเวลาที่ใช้ในการประมวลผลของแต่ละวิธีในทุกชุดข้อมูล

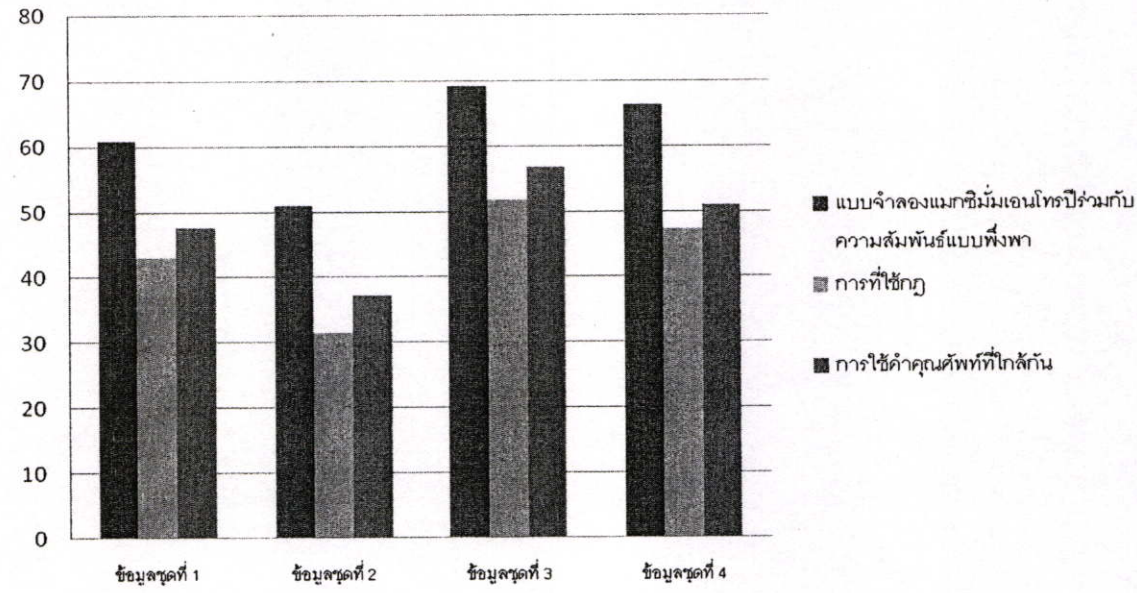
วิธีการ	เวลา (วินาที)			
	ข้อมูล ชุดที่ 1	ข้อมูล ชุดที่ 2	ข้อมูล ชุดที่ 3	ข้อมูล ชุดที่ 4
การใช้แบบจำลองแมกซิมั่มเอนโทรปีและความสัมพันธ์แบบพึ่งพา	642.48	908.58	1015.10	1883.70
การใช้กฎ	47.07	60.27	100.31	130.04
การใช้คำคุณศัพท์ที่ใกล้กัน	40.91	54.33	59.89	101.69



รูปที่ 4.2 แผนภูมิแท่งเปรียบเทียบค่าความเที่ยงตรงในการสกัดคุณลักษณะสินค้าและความเห็น

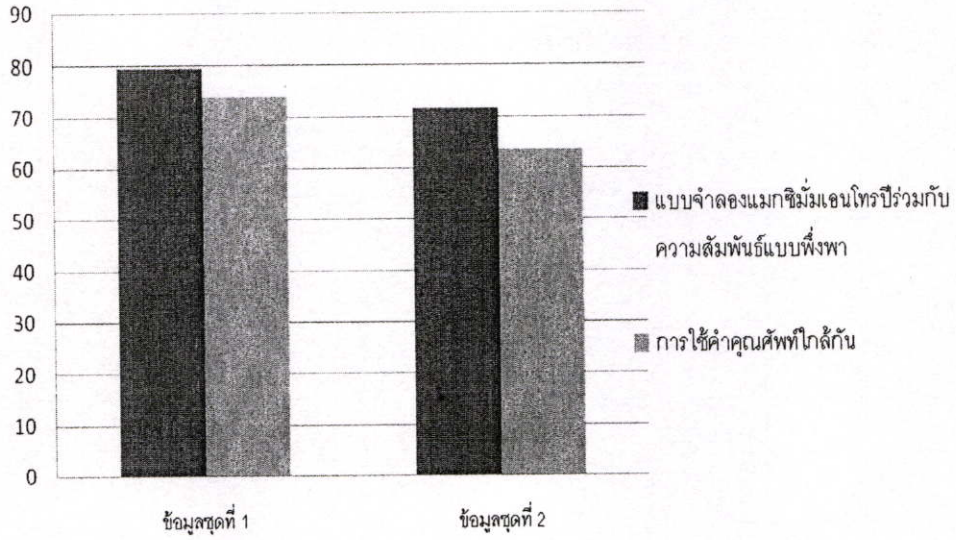


รูปที่ 4.3 แผนภูมิแท่งเปรียบเทียบค่าความระลึกในการสกัดคุณลักษณะสินค้าและความเห็น



รูปที่ 4.4 แผนภูมิแท่งเปรียบเทียบค่าเอฟในการสกัดคุณลักษณะสินค้าและความเห็น

นอกจากนี้จากผลการทดลองในการสกัดประโยคความเห็นในชุดข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu และชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu ซึ่งแสดงดังรูปที่ 4.5 แผนภูมิแท่งเปรียบเทียบค่าเอฟ พบว่าวิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพามีความสามารถในการสกัดประโยคความเห็นที่ดีกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน



รูปที่ 4.5 แผนภูมิแท่งเปรียบเทียบค่าเอฟในการสกัดประโยคความเห็น

4.8 การทดลองเปรียบเทียบกับตัวจำแนกประเภทแบบอื่น

วิธีการที่นำเสนอนี้ใช้การวิเคราะห์ความสัมพันธ์แบบพึ่งพาผสมกับการเรียนรู้ด้วยเครื่อง โดยใช้แบบจำลองแมกซิมั่มเอนโทรปีเป็นตัวจำแนกประเภท วิธีการจำแนกประเภทนอกจากแบบจำลองแมกซิมั่มเอนโทรปีแล้วยังมีตัวจำแนกเบส์อย่างง่าย และโครงข่ายประสาทเทียม ที่เป็นวิธีที่นิยมใช้สำหรับการจำแนกประเภท หัวข้อนี้จึงนำเสนอผลการทดลองในการใช้ตัวจำแนกเบส์อย่างง่าย และโครงข่ายประสาทเทียมแบบหลายชั้นในการสกัดคู่คุณลักษณะสินค้าและความเห็น ผลการทดลองแสดงดังตารางที่ 4.30 ถึง 4.33

ตารางที่ 4.30 ค่าเฉลี่ยการวัดประสิทธิภาพในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละตัวจำแนกในชุดข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu

วิธีการ	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้ตัวจำแนกเบส์อย่างง่าย	56.73	53.56	54.99
การใช้โครงข่ายประสาทเทียม	51.25	57.35	53.07
การใช้แบบจำลองแมกซิมั่มเอนโทรปี	**59.54**	**62.37**	**60.88**

ตารางที่ 4.31 ค่าเฉลี่ยการวัดประสิทธิภาพในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละตัวจำแนกในชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu

วิธีการ	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้ตัวจำแนกเบส์อย่างง่าย	42.97	38.48	40.42
การใช้โครงข่ายประสาทเทียม	45.55	52.25	47.49
การใช้แบบจำลองแมกซิมั่มเอนโทรปี	**47.09**	**56.37**	**51.08**

ตารางที่ 4.32 ค่าเฉลี่ยการวัดประสิทธิภาพในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละตัวจำแนกในชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon

วิธีการ	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้ตัวจำแนกเบส์อย่างง่าย	69.98	63.81	66.72
การใช้โครงข่ายประสาทเทียม	66.71	74.00	69.96
การใช้แบบจำลองแมกซิมั่มเอนโทรปี	68.16	70.22	69.16

ตารางที่ 4.33 ค่าเฉลี่ยการวัดประสิทธิภาพในการสกัดคู่คุณลักษณะสินค้าและความเห็นของแต่ละตัวจำแนกในชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon

วิธีการ	ค่าความระลึก (Recall)	ค่าความเที่ยงตรง (Precision)	ค่าเอฟ (F-measure)
การใช้ตัวจำแนกเบส์อย่างง่าย	62.64	60.28	61.30
การใช้โครงข่ายประสาทเทียม	59.92	64.34	61.92
การใช้แบบจำลองแมกซิมั่มเอนโทรปี	**63.94**	**68.83**	**66.19**

จากตารางที่ 4.30 ถึง 4.33 แสดงค่าเฉลี่อการวัดประสิทธิภาพในการสกัดคุณลักษณะสินค้าและความเห็นในข้อมูลทั้ง 4 ชุด โดยใช้ตัวจำแนกเบส์อย่างง่าย และโครงข่ายประสาทเทียมร่วมกับความสัมพันธ์แบบพึ่งพา เปรียบเทียบกับการใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา พบว่าการสกัดคู่คุณลักษณะสินค้าและความเห็นในชุดข้อมูล 3 ชุด ได้แก่ ข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu ข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และ

เว็บไซต์ Amazon โดยใช้แบบจำลองแมกซิมั่มเอนโทรปีเป็นตัวจำแนกมีค่าความระลึก ค่าความเที่ยงตรงและค่าเอฟสูงสุด ยกเว้นการสกัดคู่คุณลักษณะสินค้าและความเห็นในชุดข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon ที่การใช้โครงข่ายประสาทเทียมเป็นตัวจำแนกจะมีค่าความเที่ยงตรงสูงสุด แต่การใช้แบบจำลองแมกซิมั่มเอนโทรปีมีค่าความระลึกสูงสุด และทั้งการใช้โครงข่ายประสาทเทียมและแบบจำลองแมกซิมั่มเอนโทรปีมีค่าเอฟที่ไม่แตกต่างกันมากคือ 69.96 และ 69.16 ตามลำดับ นอกจากนี้พบว่าในการใช้ตัวจำแนกเบส์อย่างง่ายผลการทดลองส่วนใหญ่มีค่าความระลึกสูงกว่าการใช้โครงข่ายประสาทเทียม แต่โครงข่ายประสาทเทียมมีค่าความเที่ยงตรงสูงกว่าการใช้ตัวจำแนกเบส์อย่างง่าย จากผลการทดลองสรุปได้ว่า แบบจำลองแมกซิมั่มเอนโทรปีเป็นวิธีที่เหมาะสมวิธีหนึ่งในการแก้ปัญหาการสกัดคุณลักษณะสินค้าและความเห็น

นอกจากนี้เมื่อวิเคราะห์ความไวในการจำแนก (Sensitivity) ของตัวจำแนกแต่ละตัวในข้อมูลแต่ละชุดซึ่งมีจำนวนประโยคในการเรียนรู้และประโยคในการทดสอบที่แตกต่างกัน โดยที่ชุดข้อมูลที่ 1 ข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu มีจำนวนประโยคในการเรียนรู้ 473 ประโยคและประโยคในการทดสอบ 118 ประโยค ชุดข้อมูลที่ 2 ข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu มีจำนวนประโยคในการเรียนรู้ 730 ประโยคและประโยคในการทดสอบ 183 ประโยค ชุดข้อมูลที่ 3 ข้อมูลการวิจารณ์กล้องดิจิทัลจากเว็บไซต์ Amazon มีจำนวนประโยคในการเรียนรู้ 874 ประโยคและประโยคในการทดสอบ 218 ประโยค และชุดข้อมูลที่ 4 ข้อมูลการวิจารณ์กล้องดิจิทัลจากงานวิจัยของ Hu and Liu และเว็บไซต์ Amazon มีจำนวนประโยคในการเรียนรู้ 1,604 ประโยคและประโยคในการทดสอบ 401 ประโยค พบว่าความไวในการจำแนกซึ่งวัดจากค่าความระลึก การใช้แบบจำลองแมกซิมั่มเอนโทรปีมีค่าความไวในการจำแนก คือ 59.54, 47.09, 68.16 และ 63.94 ตามลำดับ แม้ว่าขนาดของข้อมูลชุดที่ 1 จะมีจำนวนน้อยที่สุดแต่ความไวในการจำแนกมีค่าสูงกว่าข้อมูลชุดที่ 2 แสดงว่า ถึงแม้ขนาดข้อมูลที่ใช้ในการการเรียนรู้และข้อมูลที่ใช้ในการทดสอบจะแตกต่างกันแบบจำลองแมกซิมั่มเอนโทรปีก็ยังมีความสามารถในการจำแนกคู่ที่เป็นคุณลักษณะสินค้าและความเห็นได้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ปัจจุบันข้อมูลการวิจารณ์สินค้าที่ลูกค้าแสดงความเห็นต่อสินค้าผ่านเว็บไซต์ต่าง ๆ ถือได้ว่าเป็นข้อมูลที่มีประโยชน์ต่อทั้งผู้ผลิตสินค้าและผู้ที่สนใจจะซื้อสินค้า แต่เนื่องจากข้อมูลมีจำนวนมาก และเพิ่มขึ้นอย่างรวดเร็ว และลักษณะของข้อมูลการวิจารณ์สินค้าส่วนใหญ่เป็นประโยคแบบบรรยาย ซึ่งทำให้การวิเคราะห์ข้อมูลด้วยมนุษย์ต้องใช้เวลามาก และมีความสิ้นเปลืองสูง ดังนั้น จึงมีความต้องการในการจัดการกับข้อมูลการวิจารณ์สินค้าเหล่านี้ เพื่อให้ได้ข้อมูลในลักษณะที่สามารถนำไปใช้งานได้อย่างมีประสิทธิภาพ การสรุปความเห็นเป็นวิธีการจัดการกับข้อมูลการวิจารณ์สินค้าวิธีหนึ่ง ซึ่งจะจำแนกความเห็นตามคุณลักษณะของสินค้า และด้านของความเห็นที่มีต่อคุณลักษณะสินค้านั้นๆ ปัญหาหลักที่สำคัญของการสรุปความคิดเห็น คือ การสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้านั้นๆ ปัญหาการสกัดคุณลักษณะสินค้าและความเห็นที่มีต่อคุณลักษณะสินค้า จะถูกแบ่งแยกออกเป็น 2 ส่วน คือ การสกัดคุณลักษณะสินค้าและการสกัดความเห็นที่มีต่อคุณลักษณะสินค้านั้น โดยกระบวนการจะเริ่มจากการสกัดคุณลักษณะสินค้าก่อนแล้วจึงนำคุณลักษณะสินค้าที่สกัดได้ไปสกัดคำที่แสดงความเห็นที่มีต่อคุณลักษณะสินค้านั้นในประโยค วิธีการสกัดความเห็นที่มีต่อคุณลักษณะสินค้านั้น วิธีการที่ผ่านมาจะใช้คำคุณศัพท์ที่ใกล้กันและใช้กฎ ซึ่งมีข้อจำกัดในการสกัดความเห็นสำหรับประโยคแบบยาวที่มีโครงสร้างทางไวยากรณ์ที่ซับซ้อน

งานวิจัยนี้จึงได้นำเสนอวิธีการสกัดคุณลักษณะสินค้าและความเห็นโดยใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพา และจากผลการทดลองสกัดคุณลักษณะสินค้าและความเห็น พบว่าวิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพามีค่าความเที่ยงตรง ค่าความระลึกและค่าเอฟสูงกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันและวิธีการใช้กฎ นอกจากนี้ การทดลองสกัดประโยคความเห็น ถึงแม้ว่าวิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพาจะมีค่าความเที่ยงตรงต่ำกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันในชุดข้อมูลการวิจารณ์โทรศัพท์เคลื่อนที่จากงานวิจัยของ Hu and Liu ทั้งนี้เนื่องมาจากประโยคในชุดข้อมูลการวิจารณ์นี้ส่วนใหญ่เป็นประโยคแบบสั้นที่มีโครงสร้างประโยคไม่ซับซ้อน (Simple Sentence) ซึ่งวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กันจะให้มีความถูกต้องถ้าประโยคการวิจารณ์มีรูปแบบตรงกับรูปแบบที่กำหนด แต่เมื่อพิจารณาค่าเอฟหรือค่าความแม่นยำโดยรวม พบว่าวิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพามีความสามารถในการสกัดประโยคความเห็นที่ดีกว่าวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน

สรุปได้ว่า วิธีการที่ใช้แบบจำลองแมกซิมั่มเอนโทรปีร่วมกับความสัมพันธ์แบบพึ่งพาช่วยให้การสกัดคุณลักษณะสินค้าและความเห็นมีประสิทธิภาพดียิ่งขึ้น ทั้งในแง่ของความถูกต้อง และในแง่ของความสามารถในการสกัด เมื่อเปรียบเทียบกับวิธีการที่ใช้คำคุณศัพท์ที่ใกล้กัน และวิธีการที่ใช้กฎ เนื่องจากข้อสนเทศที่ได้จากการวิเคราะห์ความสัมพันธ์แบบพึ่งพาเป็นข้อสนเทศจากบริบทในระยะไกล ไม่ถูกจำกัดขอบเขตของคำเหมือนวิธีการที่ผ่านมา ทำให้สามารถสกัดคุณลักษณะสินค้าและความเห็นที่อยู่ในรูปแบบของการเกิดร่วมกันในระยะไกลได้ นอกจากนี้การใช้แบบจำลองแมกซิมั่มเอนโทรปีซึ่งเป็นเป็นวิธีการเรียนรู้ด้วยเครื่องแบบมีการสอน และเป็นแบบจำลองที่พยายามสร้างให้มีลักษณะของการกระจายสม่ำเสมอตามข้อเท็จจริงทั้งหมดที่มีอยู่และไม่ตั้งข้อสันนิษฐานใดๆ เกี่ยวกับข้อเท็จจริงที่ไม่รู้ ช่วยให้ระบบสามารถตัดสินใจได้อย่างถูกต้องเหมาะสม ถึงแม้ว่าลักษณะของประโยคและคำที่ใช้ในการวิจารณ์สินค้าของผู้วิจารณ์จะมีความหลากหลายและมีการกระจายของข้อมูลมาก อย่างไรก็ตาม ประสิทธิภาพของวิธีการที่นำเสนอในการสกัดคุณลักษณะสินค้าและความเห็นจะขึ้นอยู่กับจำนวนและลักษณะของชุดตัวอย่างที่ใช้ในการเรียนรู้ด้วย นอกจากนี้ ยังอาจจะมีข้อด้อยในเรื่องของเวลาที่ใช้ในการประมวลผลมากกว่าวิธีการที่ไม่ต้องมีการเรียนรู้

5.2 ข้อเสนอแนะ

ถึงแม้ว่าผลการทดลองจะแสดงให้เห็นถึงประสิทธิภาพที่ดีขึ้นของวิธีการที่นำเสนอในงานวิจัยนี้ แต่ก็ยังมีข้อจำกัดบางประการ สรุปข้อเสนอแนะได้ดังนี้

1. แนวทางที่ใช้ในงานวิจัยนี้เป็นวิธีแบบมีการสอน ในการสกัดคุณลักษณะสินค้าและความเห็นจะต้องทำการเรียนรู้จากคลังประโยคการวิจารณ์สินค้าในแต่ละโดเมน หากไม่มีการนำชุดข้อมูลมาเรียนรู้ ความถูกต้องของระบบจะลดลง ดังนั้น เมื่อต้องการนำไปใช้งานกับโดเมนอื่นจำเป็นต้องนำชุดข้อมูลในโดเมนนั้นมาทำการเรียนรู้ก่อน

2. วิธีการทางด้านการประมวลผลภาษาธรรมชาติที่ใช้ในงานวิจัยนี้เป็นการวิเคราะห์ทางวากยสัมพันธ์ ไม่ได้วิเคราะห์ทางความหมาย จึงมีข้อจำกัดเรื่องของการใช้คำอ้างอิงหรือคำสรรพนามในประโยค นอกจากนั้นแล้วคำบางคำในประโยคหนึ่งจะเข้าใจความหมายได้ถูกต้องดูประโยคก่อนหน้าหรือประโยคตามด้วย เช่นประโยค *"I purchased this phone knowing it was slightly old as compared to the newer models of Nokia phones available. But not good enough for me."* ดังนั้น หากมีการนำวิธีการอื่นทางด้านการประมวลผลธรรมชาติมาใช้ร่วมด้วย เช่น การหาคำอ้างอิง (Anaphora Resolution) และการบูรณาการทางวจนิพนธ์ (Discourse Integration) ซึ่งเป็นการพิจารณาความหมายของประโยค โดยดูจากประโยคข้างเคียง จะช่วยให้ประสิทธิภาพของการสกัดคุณลักษณะสินค้าและความเห็นเพิ่มสูงขึ้น

เอกสารอ้างอิง

[1] Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T. "Mining Product Reputations on the Web." **Proceedings of Conference on Knowledge Discovery and Data Mining,** Edmonton, 2002. pp. 341-349.

[2] Hu, M. and Liu, B. "Mining and Summarization Customer Reviews." **Proceedings of Conference on Knowledge Discovery and Data Mining,** Seattle, WA, August 22-25, 2004. pp. 168-177.

[3] Liu, B., Hu, M. and Cheng, J. "Opinion Observer: Analyzing and Comparing Opinions on the Web." **Proceedings of International World Wide Web Conference,** Chiba, May 10-14, 2005. pp. 342-351.

[4] Popescu, A.M. "**Information Extraction from Unstructured Web Text**." Ph.D. Thesis of University of Washington. 2007.

[5] Yi, J. and Niblack, W. "Sentiment Mining in WebFountain." **Proceedings of International Conference on Data Engineering,** Washington, DC, April 05-08, 2005. pp. 1073-1083.

[6] Gamon, M., Aue, A., Oliver, S. and Ringger, E. "Pulse: Mining Customer Opinions from Free Text." **Proceedings of International Symposium on Intelligent Data Analysis,** 2005. pp. 121-132.

[7] Zhuang, L., Jing, F. and Zhu, X.Y. "Movie Review Mining and Summarization." **Proceedings of Conference on Information and Knowledge Management,** Arlington, VA, November 05-11, 2006.

[8] Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H. and Jin, C. "Red Opal: Product-Feature Scoring from Reviews." **Proceedings of ACM Conference on Electronic Commerce,** San Diego, CA, June 11-15, 2007. pp. 182-191.

[9] Liu, B. **Web Data Mining Exploring Hyperlinks, Contents, and Usage Data.** New York : Springer. 2007.

[10] Baeza-Yates, R. and Ribeiro-Neto, B. **Modern Information Retrieval.** New York : ACM Press. 1999.

[11] Weiss, S.M., Indurkhya, N., Zhang, T. and Damerau, F.J. **Text Mining.** New York : Springer. 2005.

[12] McCallum, A., Freitag, D. and Pereira, F. "Maximum Entropy Markov Models for Information Extraction and Segmentation." **Proceedings of International Conference on Machine Learning,** Stanford, CA, 2000. pp. 591-598.

[13] บุญเสริม กิจศิริกุล. **ปัญญาประดิษฐ์**. ภาควิชาวิศวกรรมคอมพิวเตอร์, คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย. 2546.

[14] Jaynes, E.T. "Information Theory and Statistical Mechanics." **Physics Reviews**, Vol. 106, 1957. pp. 620-630.

[15] Berger, A.L., Stephen, D.P. and Vincent, D.P. "A Maximum Entropy Approach to Natural Language Processing." **Computational Linguistics**, Vol. 22, No. 1, 1996. pp. 39-71.

[16] Reynar, J.C. and Ratnaparkhi, A. "A Maximum Entropy Approach to Identifying Sentence Boundaries." **Proceedings of Applied Natural Language Processing Conference,** Washington, DC, 1997. pp. 16-19.

[17] Chieu, H.L. and Ng, H.T. "A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text." **Proceedings of National Conference on Artificial Intelligence,** Edmonton, 2002. pp. 786-791.

[18] Ratnaparkhi, A. "A Maximum Entropy Model for Part-of-Speech Tagging." **Proceedings of Conference on Empirical Methods in Natural Language Processing,** 1996. pp. 133-141.

[19] Ratnaparkhi, A. "A Simple Introduction to Maximum Entropy Models for Natural Language Processing." **Technical Report IRCS 97-08**, Institute for Research in Cognitive Science, University of Pennsylvania, 1997.

[20] Malouf, R. "A Comparison of Algorithms for Maximum Entropy Parameter Estimation." **Proceedings of Conference on Computational Natural Language Learning,** 2002. pp. 49-55.

[21] Darroch, J. and Ratcliff, D. "Generalized Iterative Scaling for Log-Linear Models." **The Annals of Mathematical Statistics**, Vol. 43, No. 5, 1972. pp. 1470-1480.

[22] Stanford University. "**Stanford Lexicalized Parser - a probabilistic lexicalized NL CFG parser.**" [Online]. Available : http://nlp.stanford.edu/downloads/lex-parser.shtml. 2006.

[23] Allen, J. **Natural Language Understanding.** California : Benjamin Cummings Publishing Company Inc. 1995.

[24] Melcük, I. **Dependency Syntax: Theory and Practice.** New York : State University of New York Press. 1987.

[25] Khan, R.L. **Ontology-Based Information Selection.** Ph.D.dissertation, University of Southern California. 2000.

[26] Wu, W.C. and Liu, L.C. "Ontology-Based Text Summarization for Business News Articles." **Proceedings of International Conference Computers and Their Applications,** Honolulu, Hawaii, March 26-28, 2003. pp. 389-394.

[27] Boufaden, N. "An Ontology-Based Semantic Tagger for IE System." **Proceedings of Annual Meeting of the Association for Computational Linguistics**, Sapporo, 2003.

[28] Gao, M., Liu, C. and Chen, F. "An Ontology Search Engine Based on Semantic Analysis." **Proceedings of International Conference on Information Technology and Applications,** Sydney, 2005.

[29] Cheng, K.C., Pan, S.X. and Kurfess, F. "Ontology-Based Semantic Classification of Unstructured Documents." **Proceedings of International Workshop on Adaptive Multimedia Retrieval,** 2003.

[30] Dave, K., Lawrence, S. and Pennock, D. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews." **Proceedings of International World Wide Web Conference,** Budapest, May 20-24, 2003. pp. 519-528.

[31] Somprasertsri, G. and Lalitrojwong, P. "Automatic Product Feature Extraction from Online Product Reviews Using Maximum Entropy with Lexical and Syntactic Features." **Proceedings of Information Reuse and Integration,** Las Vegas, NA, July 13-15, 2008. pp. 250-255.

[32] Stone, P., Dunphy, D., Smith, M., Ogilvie, D. and et al. **The General Inquirer: A Computer Approach to Content Analysis.** Cambridge, MA: The MIT Press. 1966.

[33] Jannach, D., Shchekotykhin, K. and Friedrich, G. "Ontology Instantiation from Tabular Web Sources - The AllRight System." **Journal of Web Semantics**, Vol. 7, No. 3, 2009. pp.136-153.

ภาคผนวก ก.
ชุดหมวดคำที่ใช้ในงานวิจัย

หมวดคำ	คำอธิบาย	หมวดคำ	คำอธิบาย
CC	Coordinating conjunction	VBD	Verb, past tense
CD	Cardinal number	VBN	Verb, past participle
DT	Determiner	WDT	Wh-determiner
EX	Existential *there*	WP	Wh-pronoun
FW	Foreign word	WP$	Possessive wh-pronoun
IN	Preposition or subordinating conjunction	WRB	Wh-adverb
JJ	Adjective	UH	Interjection
JJR	Adjective, comparative		
JJS	Adjective, superlative		
LS	List item marker		
MD	Modal		
NN	Noun, singular or mass		
NNS	Noun, plural		
NNP	Proper noun, singular		
NNPS	Proper noun, plural		
PDT	Predeterminer		
POS	Possessive ending		
PRP	Personal pronoun		
PRP$	Possessive pronoun		
RB	Adverb		
RBR	Adverb, comparative		
RBS	Adverb, superlative		
RP	Particle		
SYM	Symbol		
TO	*to*		
VB	Verb, base form		
VBG	Verb, gerund or present participle		
VBP	Verb, non-3rd person singular present		
VBZ	Verb, 3rd person singular present		

ภาคผนวก ข.
รายการคำหยุดที่ใช้ในงานวิจัย

ลำดับที่	คำหยุด	ลำดับที่	คำหยุด	ลำดับที่	คำหยุด
1	a	30	asked	59	does
2	about	31	asking	60	doesn't
3	after	32	asks	61	doing
4	again	33	at	62	done
5	against	34	away	63	each
6	ago	35	be	64	early
7	all	36	became	65	either
8	almost	37	because	66	else
9	alone	38	become	67	even
10	along	39	becomes	68	evenly
11	already	40	been	69	ever
12	also	41	before	70	every
13	although	42	being	71	everybody
14	always	43	between	72	everyone
15	am	44	both	73	everything
16	among	45	but	74	everywhere
17	an	46	by	75	fact
18	and	47	came	76	facts
19	another	48	can	77	far
20	any	49	cannot	78	felt
21	anybody	50	come	79	for
22	anyhow	51	could	80	four
23	anyone	52	day	81	from
24	anything	53	did	82	further
25	anyway	54	differ	83	furthered
26	anywhere	55	different	84	furthering
27	are	56	differently	85	furthers
28	as	57	do	86	gave
29	ask	58	doe	87	general

ลำดับที่	คำหยุด	ลำดับที่	คำหยุด	ลำดับที่	คำหยุด
88	generally	117	in	146	me
89	get	118	into	147	mine
90	gets	119	is	148	men
91	getting	120	isn	149	might
92	give	121	isn't	150	month
93	given	122	it	151	more
94	gives	123	its	152	most
95	go	124	itself	153	mostly
96	goes	125	just	154	mr
97	going	126	k	155	mrs
98	gone	127	keep	156	much
99	got	128	keeps	157	must
100	gotten	129	just	158	my
101	had	130	knew	159	myself
102	has	131	know	160	nt
103	have	132	known	161	never
104	having	133	knows	162	no
105	he	134	last	163	nobody
106	her	135	least	164	non
107	here	136	left	165	noone
108	herself	137	less	166	not
109	him	138	let	167	nothing
110	himself	139	lets	168	now
111	his	140	like	169	nowhere
112	hour	141	likely	170	of
113	how	142	make	171	often
114	however	143	many	172	old
115	i	144	may	173	older
116	if	145	maybe	174	oldest

ลำดับที่	คำหยุด	ลำดับที่	คำหยุด	ลำดับที่	คำหยุด
175	once	204	several	233	think
176	one	205	shall	234	thinks
177	only	206	she	235	this
178	onto	207	should	236	those
179	or	208	since	237	three
180	our	209	so	238	through
181	ourselves	210	some	239	thus
182	out	211	somebody	240	till
183	over	212	someone	241	to
184	per	213	something	242	too
185	perhaps	214	somewhere	243	today
186	put	215	stand	244	two
187	putting	216	still	245	unless
188	rather	217	such	246	until
189	really	218	sure	247	upon
190	said	219	take	248	us
191	same	220	taken	249	ve
192	saw	221	than	250	very
193	say	222	that	251	w
194	says	223	the	252	was
195	second	224	their	253	we
196	seconds	225	them	254	week
197	see	226	then	255	went
198	seem	227	there	256	were
199	seemed	228	therefore	257	what
200	seeming	229	these	258	what"s
201	seems	230	they	259	whatever
202	sees	231	thing	260	when
203	seen	232	things	261	where

ลำดับที่	คำหยุด	ลำดับที่	คำหยุด	ลำดับที่	คำหยุด
262	whether	270	will	278	years
263	which	271	with	279	yet
264	while	272	within	280	you
265	who	273	without	281	young
266	whoever	274	won't	282	younger
267	whom	275	would	283	youngest
268	whose	276	wouldn't	284	your
269	why	277	year	285	yours

ภาคผนวก ค.
ความสัมพันธ์แบบพึ่งพาที่ใช้ในงานวิจัย

ความสัมพันธ์	คำอธิบาย	ความสัมพันธ์	คำอธิบาย
abbrev	abbreviation modifier	number	element of compound number
acomp	adjectival complement	parataxis	parataxis
advcl	adverbial clause modifier	partmod	participial modifier
advmod	adverbial modifier	pcomp	prepositional complement
agent	agent	pobj	object of a preposition
amod	adjectival modifier	poss	possession modifier
appos	appositional modifier	possessive	possessive modifier
attr	attributive	preconj	preconjunct
aux	auxiliary	predet	predeterminer
auxpass	passive auxiliary	prep	prepositional modifier
cc	coordination	prepc	prepositional clausal modifier
ccomp	clausal complement	prt	phrasal verb particle
complm	complementizer	punct	punctuation
conj	conjunct	purpcl	purpose clause modifier
cop	copula	quantmod	quantifier phrase modifier
csubj	clausal subject	rcmod	relative clause modifier
csubjpass	clausal passive subject	ref	referent
det	determiner	rel	relative
dobj	direct object	mod	temporal modifier
expl	expletive	xcomp	open clausal complement
infmod	infinitival modifier	xsubj	controlling subject
iobj	indirect object		
mark	marker		
measure	measure-phrase modifier		
neg	negation modifier		
nn	noun compound modifier		
nsubj	nominal subject		
nsubjpass	passive nominal subject		
num	numeric modifier		

ภาคผนวก ง.
ผลงานวิจัยที่ได้รับการตีพิมพ์

งานวิจัยที่ได้รับการตีพิมพ์

[1] Somprasertsri, G. and Lalitrojwong, P. "Automatic Product Feature Extraction from Online Product Reviews Using Maximum Entropy with Lexical and Syntactic Features." **Proceedings of the 6th IEEE International Conference on Information Reuse and Integration,** Las Vegas, NA, July 13-15, 2008. pp. 250-255.

[2] Somprasertsri, G. and Lalitrojwong, P. "A Maximum Entropy for Product Feature Extraction in Online Customer Reviews." **Proceedings of the 3rd IEEE International Conference on Cybernetics and Intelligent System**, Chengdu, September 21-24, 2008. pp. 575-580.

[3] Somprasertsri, G. and Lalitrojwong, P. "Extracting Product Features and Opinions from Product Reviews Using Dependency Analysis" **Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery**, Yantai, August 10-12, 2010. pp. 2358-2362.

[4] Somprasertsri, G. and Lalitrojwong, P. "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization." **Journal of Universal Computer Science,** Vol. 16, No. 6, 2010. pp. 938-995.

Automatic Product Feature Extraction from Online Product Reviews Using Maximum Entropy with Lexical and Syntactic Features

Gamgarn Somprasertsri[1,2] and Pattarachai Lalitrojwong[1]
[1]*Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang Bangkok, Thailand, s7066001@kmitl.ac.th, pattarachai@it.kmitl.ac.th*
[2]*Faculty of Informatics, Mahasarakham University, Mahasarakham, Thailand, gamgarn.s@msu.ac.th*

Abstract

The task of product feature extraction is to find product features that customers refer to their topic reviews. It would be useful to characterize the opinions about the products. We propose an approach for product feature extraction by combining lexical and syntactic features with a maximum entropy model. For the underlying principle of maximum entropy, it prefers the uniform distributions if there is no external knowledge. Using a maximum entropy approach, firstly we extract the learning features from the annotated corpus, secondly we train the maximum entropy model, thirdly we use trained model to extract product features, and finally we apply a natural language processing technique in postprocessing step to discover the remaining product features. Our experimental results show that this approach is suitable for automatic product feature extraction.

1. Introduction

Recently, the number of online shopping customers has increased due to the rapid growth of e-commerce, and the increase of online merchants. To enhance the customer satisfaction, merchants and product manufacturers allow customers to review or express their opinions on the products they buy from the websites, such as amazon.com, cnet.com, and epinions.com. Online customer reviews become the source of information which is very useful to both potential customers and product manufacturers. People read them for making a decision on whether to purchase the product. For a product manufacturer, knowing the preferences of customers is highly valuable for product development, marketing and consumer relationship management. Mining reviews mostly in free-form text can be extremely expensive. Besides, it is hard to find the useful information from the plenty of customer reviews. Therefore, this trend has raised many interesting and challenging research topics such as subjectivity classification, sentiment classification, and review mining and summarization.

Subjectivity classification is a task for classifying the sentences or the documents which contain opinions from factual, as in [1][2]. It is useful for many natural language processing applications such as question answering, information extraction, and so on. The task of sentiment classification is to judge whether a review expresses a positive or negative opinion. For example, [3][4] developed methods for sentiment classification in document level. The systems assign a positive or negative sentiment for the whole review document. The sentiment of phrases and sentences has also been studied in [5][6]. Even if sentiment classification is useful, it does not find what the reviewer liked and disliked. Review mining and summarization is the task of producing a sentiment summary, which consists of sentences from reviews that capture the author's opinion. Review summarization is interested in features or objects on which customers have opinions. It also determines whether the opinions are positive or negative. This makes it differ from traditional text summarization. Most existing works on review mining and summarization mainly focus on product reviews. For example, [7][8][9] concentrated on mining and summarizing reviews by extracting opinion sentences regarding product features. In another domain, [10] proposed a multi-knowledge based approach for movie review mining and summarization.

In general, mining and summarizing customer reviews involve three tasks (Figure 1): firstly, feature extraction identifies and extracts object features that have been commented in each review; secondly, sentiment assignment determines the polarity of each feature to be positive or negative; and thirdly, summary visualization summarizes the result in order to show this result more effectively.

978-1-4244-2660-7/08/$25.00 ©2008 IEEE

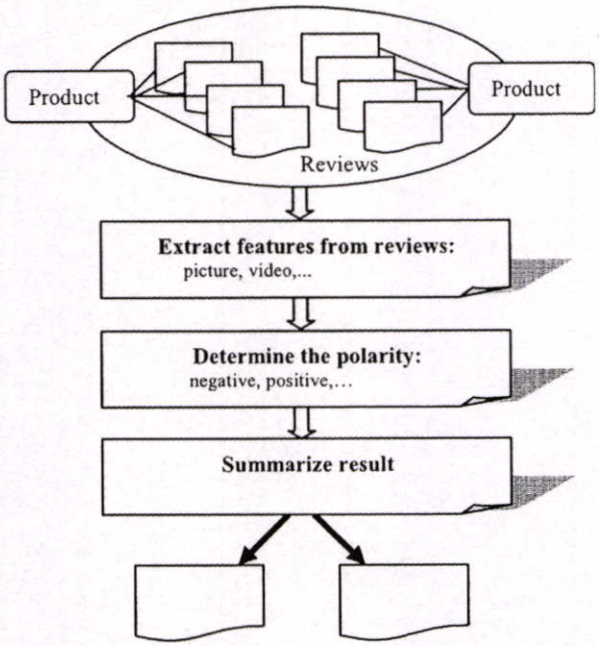


Figure 1. Review mining and summarization process.

The feature extraction is an important task of review mining and summarization. Current studies on feature extraction are mainly carried out from online product reviews. The product feature identification would be useful to characterize the opinions which the customers review or express about the products. The product reviews on the Web are in three formats [11]:

- Format 1 - Pros, cons and the detailed review: The reviewers describe pros and cons in the form of short phrases and also write the detail of reviews separately.
- Format 2 - Pros and cons: The reviewers describe pros and cons in the form of full sentences separately.
- Format 3 - Free format: The reviewers write the reviews in the free form that no separation of pros and cons.

In format 1, pros and cons usually consist of short phrases and incomplete sentences. For example, *"pros: fabulous photo quality, large LCD, great battery life, great features"*. The reviews of format 2 and 3 usually consist of long sentences and complete sentences. For example, *"The larger lens of the g3 gives better picture quality in low light, and the 4-times optical zooms gets you just that much closer"*. However, the product features extraction from reviews of format 2 and 3 is more challenge because the complete sentences are more complex and contain a large amount of irrelevant information.

This paper focuses on the product feature extraction from reviews of format 3. We propose an approach by combining lexical and syntactic features with maximum entropy model for extracting the product features. Our goal is to investigate whether the maximum entropy model is suitable for automatic product feature extraction. Our experimental results show that this approach is effective. The rest of this paper is organized as follows. Section 2 describes related work on the task of product feature extraction. Section 3 introduces the maximum entropy model. Section 4 discusses how to extract product features from online product reviews using maximum entropy. Section 5 presents and discusses the experimental results. Finally, Section 6 concludes our work.

2. Related work

Hu and Liu's work in [12] can be considered as the pioneer work on feature extraction from reviews. Their feature extraction algorithm is based on heuristics that depend on feature terms' respective occurrence counts. They use association rule mining based on the Apriori algorithm to extract frequent itemsets as explicit product features (only in the form of noun phrases). In association rule mining, the algorithm does not consider the position of the words in a sentence. In order to remove incorrect frequent features, they use feature pruning that consists of compactness pruning and redundancy pruning. To improve the work over [12], Liu, Hu, and Cheng [13] propose a technique based on language pattern mining to identify product features from pros and cons in reviews in the form of short sentences. They also make an effort to extract implicit features.

Popescu and Etzioni [8] developed an unsupervised information extraction system called OPINE extracting product features and opinions from reviews. OPINE first extracts noun phrases from reviews and retains those with frequency greater than an experimentally set threshold and then assesses those by OPINE's feature assessor for extracting explicit features. The assessor evaluates a noun phrase by computing a Point-wise Mutual Information score between the phrase and meronymy discriminators associated with the product class.

Carenini, Ng and Zwart [14] proposed feature extraction for capturing knowledge from product reviews. In their method, the output of Hu and Liu's system [12] was used as the input to their system, and map the input to the user-defined taxonomy features hierarchy thereby eliminating redundancy and providing conceptual organization.

Finally, Yi and Niblack [15] developed a set of feature term extraction heuristics and selection algorithms for extracting a feature term from product reviews. The feature term is a part of relationship with the given topic, an attribute of relationship with the given topic, and an attribute of relationship with a known feature of the given

topic. In the first step, they extract a noun phrase with the Beginning define Base Noun Phrase (bBNP) heuristics. Then, they select a feature term from the noun phrase using the likelihood score.

Our motivation for the task of product feature extraction is that the syntactic dependency and context information may be useful for determining whether the word is a product feature or non-product feature.

3. The maximum entropy model

The Maximum entropy (ME) was first described by Jaynes in [16] and more recently in a draft manuscript available on the Web [17]. The maximum entropy model is a framework for integrating information from many heterogeneous information sources for classification [18]. This model implements the intuition that the best model is the one consistent with the set of constraints imposed by the evidence but otherwise is as uniform as possible [19]. The maximum entropy approach has in recent years been used for a wide variety of classification problems in natural language processing, such as sentence boundary detection [20][21], information extraction [22], and part-of-speech tagging [23]. In study of Nigam, Lafferty and McCallum [24] found that maximum entropy worked better than Naïve Baye's classification for theirs classification. Unlike Naïve Baye's machine learning, maximum entropy makes no independence assumptions about the occurrence of features.

For our work, we use maximum entropy model to extract product features from online product reviews. This task can be re-formulated as a classification problem, in which the task is to observe some syntactic dependency and context information $x \in X$ and predict the class $y \in Y$. We can design classes such as product feature and non-product feature.

We can implement classifier cl: $X \rightarrow Y$ with a conditional probability model by simply choosing the class y with the highest conditional probability p in the context x:

$$cl(x) = \arg\max_{y} p(y \mid x) \tag{1}$$

The conditional probability $p(y|x)$ is defined as follows [21]:

$$p(y \mid x) = \frac{1}{Z(x)} \prod_{i=1}^{k} \alpha_i^{f_i(x,y)} \tag{2}$$

$$Z(x) = \sum_{y} \prod_{i} \alpha_i^{f_i(x,y)} \tag{3}$$

where y refers to the outcome, x is the history (or context), k is the number of features and $Z(x)$ is a normalization factor to ensure that $\sum p(y|x)=1$. Each parameter α_i corresponds to one feature f_i and can be interpreted as a weight for that feature. We use Generalized Iterative Scaling (GIS) algorithm [25] to estimate parameters α_i or weights of the selected features.

Under the maximum entropy framework, the probability for a class y and object x depends solely on the features that are active for the pair (x, y), where a feature is defined here as a function f: $X \times Y \rightarrow \{0, 1\}$ that maps a pair (x, y) to either 0 or 1. The feature is defined as follows:

$$f_{y'}(x, y) = \begin{cases} 1 & if\ y' = y\ and\ cp(x) = true \\ 0 & otherwise \end{cases} \tag{4}$$

where $cp(x)$ is contextual predication that returns true or false, corresponding to the presence or absence of useful information in some context, or history $x \in X$.

4. ME model for extracting product features

In this paper, we aim to extract explicit product features commented by customers. The product feature can be a brand name, a model name of a commodity, a property, a part, a feature of a product, a related concept, or a part of related concept [8]. In general, such product feature is often the product itself or its specific features, such as quality (e.g. "The picture quality is amazing"). We define the product feature extraction problem as a classification task: given a sequence of words (w_1, w_2,..., w_n) in a sentence, we generate a sequence of labels (y_1, y_2,..., y_n) indicating whether the word is a product feature or non-product feature.

4.1 Features for product feature extraction

To use the maximum entropy to extract product features, we define features or important information in order to constrain the model. We denote the features employed for learning as learning features, discriminative from the product features we discussed above. For each word from the training data, we compute several features automatically. The features are as follow.

Word: The target words and the part of speech tags of the target words.

Rare: Rare word information may be useful for identifying product features. It is common that a customer review contains many things that are not directly related to product features. Different customers usually have different stories. Those frequent noun/noun phrases (non-rare words) are likely to be product features and infrequent noun/noun phrases (rare words) are likely to be non-product features. A rare word in our work denotes a word which occurs less than five times in the training set. The count of five was chosen by subjective inspection of words in the training data.

Alphanumeric: The words contain letters and numerals. This information may be useful for determining whether the word is a product feature or non-product feature.

Dependency: The words and the part of speech tag of the words which are dependent on the target words in the dependency tree. The dependency tree derived from the syntactic parse tree. We compute the syntactic parse tree by using the Stanford lexicalized parser [26].

For example, we determining the sentence as "*it takes excellent picture*". The corresponding syntactic parse tree is shown in Figure 2 and the dependency tree is shown in Figure 3. The target word is *picture* and it occurs three times in training data. These features are shown following.

- **Word**: $picture_{target}$ (the target word), $NN_{target\text{-}tag}$ (POS of the target word)
- **Rare**: the target word is the rare word
- **Alphanumeric**: the target word contain only letters
- **Dependency**: $excellent_{target\text{-}dep1}$ (word on which *target* is dependent), $JJ_{dep1\text{-}tag}$ (POS of *target-dep1*), $takes_{target\text{-}dep2}$ (word on which *target* is dependent), $VBZ_{dep2\text{-}tag}$ (POS of *target-dep2*)

We trained maximum entropy model using features derived from the feature streams described above.

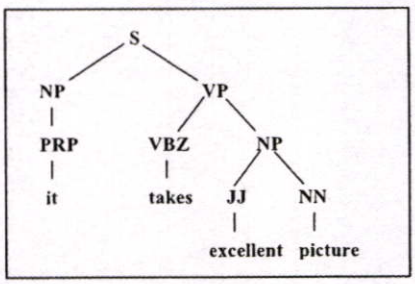


Figure 2. The syntactic parse tree for the sentence "it takes excellent picture"

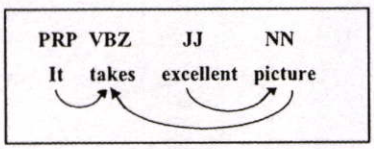


Figure 3. The dependency tree for the sentence "it takes excellent picture"

4.2 System overview

The system is divided into two modules: training module and product feature extraction module.

1) Training module

The training ME model consists of three steps. Firstly, prepare training data which includes parsing and manually product feature annotation. Secondly, extract learning features of each word in training data. Thirdly, train the model by using maximum entropy model. The result of training model is weight of each feature function.

2) Product feature extraction module

The processes of product feature extraction are shown in Figure 4.

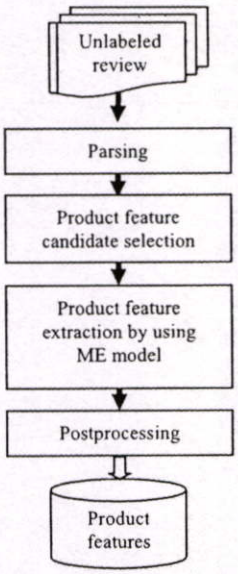


Figure 4. Process of product feature extraction

- Product feature candidate selection

After parsing the sentence, the next step is the identification of product feature candidates. This step selects words from a tagged sentence such as nouns and adjectives, which are indicated by NN and JJ respectively. Nouns and adjectives are reasonably used because most product features are nouns. However, some adjectives may appear as product features.

- Product feature extraction by using ME model

The trained model is used to predict product feature candidates from unlabeled reviews after parsing. We will simply choose the class with the highest conditional probability p according to Equation 1.

- Postprocessing

This step aims to discover the remaining product features in the reviews by matching the list of extracted product features against each word in the reviews. We applied a natural language processing technique to deal with the compound product features which are not extracted by ME model. The compound product features will be extracted if they or their head noun match the list of extracted product features. We take the sentence "*The scroll wheel was a nice idea to keep less clutter*" as an example. The word "*wheel*" is the head of "*scroll wheel*". It is extracted by ME model, thus the "*scroll wheel*" will

be extracted as a product feature. The pseudo code of this process is provided in Figure 5.

```
1. Procedure RefinementExtraction(feature_list, review)
2. begin
3. for each sentence in review
4.   for each candidate w_i in sentence
5.   begin
6.     if (w_i match f_j in feature_list) and
         (w_i's ME score < threshold) then
7.       w_i's class = product feature
8.     else if (w_{i+1}'s POS tag is noun) and
           (w_{i+1}'s class is product feature) then
9.         w_i's class = product feature
10.  endfor;
11. endfor;
12. end
```

Figure 5. Pseudo code of postprocessing step

5. Experimental results

For our experiments we used reviews on electronic products including one digital camera and one MP3 player from [27]. All the reviews are from the Amazon web site. We use 1,500 sentences for product feature extraction. This set of data was split into a training set of 80% and a testing set of 20%. We employed the Maxent version 2.4.0 as our classification tool. The parameters of the maximum entropy model can be trained with 100 iterations of the Generalized Iterative Scaling algorithm. More iterations would not affect to the increase of accuracy of the parameters. Evaluation for this task uses precision, recall and F-measure. They are defines as follows:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall}$$

where *TP* means true positive (actual product feature and predicted as product feature), *FP* means false positive (actual non-product feature and predicted as product feature), *FN* means false negative (actual product feature and predicted as non-product feature), and *F* is the inverse harmonic mean of precision and recall, a summary statistic to account for the inherent trade-off between them.

We conducted this experiment with two goals: firstly, to investigate how well the models with different feature combinations perform on the online product reviews; and secondly, to investigate how well our approach performs the product feature extraction.

Table 1. The Precision, Recall, and F-measure on the system output using combination of features

Features	Precision (%)	Recall (%)	F-measure (%)
Word + Rare	68.91	59.64	63.94
Word + Dependency	70.79	56.50	62.84
Word + Alphanumeric	68.91	59.64	63.94
Word + Dependency + Rare	73.18	56.28	63.62
Word + Alphanumeric +Rare	68.91	59.64	63.94
Word + Dependency + Alphanumeric	72.54	56.28	63.38
Word + Alphanumerric + Rare + Dependency	74.18	56.05	63.86

Table 2. The performance of system for product feature extraction

	Precision (%)	Recall (%)	F-measure (%)
Baseline	74.18	56.05	63.86
Our approach	71.63	69.06	70.32

Table 3. The Precision comparison of Hu and Liu's approach and our approach

Data set	Hu and Liu's approach	Our approach
MP3 player	69.20	72.25
Digital camera	71.00	72.22

We built several models to compare the relative utility of the features described in the previous section. From Table 1, the first column indicates which combination of features was used in models. It is very interesting to see that the system achieves a very high score of the precision when it used all features whereas it achieves a very high score of the recall when it used the combination of word with rare, word with alphanumeric or word with alphanumeric and rare. This result shows that it will improve the precision of the system by incorporate more contextual information. This phenomenon supports that context and part-of-speech information is useful for identifying product features, frequent noun/noun phrases (non-rare words) are likely to be product features and some product features contain the letters and numerals such as SD100 and 7-megapixel.

We include the postprocessing to this model for extracting the remaining product features in the reviews. To examine a postprocessing step, we also compare our approach with baseline. The baseline used only the maximum entropy with the word, rare, alphanumeric and dependency features. Our approach used a maximum entropy classifier extracting product features and applied

a natural language processing technique to deal with compound product feature candidates by head word consistency. The performance of the system is shown in Table 2. For our approach, the precision is decreased slightly. However, the recall is increased by 19.06%. Besides, our approach performs better than baseline by 7.19% F-measure. Additional, we compare our approach with Hu and Liu's approach [12]. They use association rule mining to extract product features. Table 3 shows the precision of system on the 2 datasets. We observe that both the precision of our approach are higher than those of Hu and Liu's approach. There is important difference between our approach and Hu and Liu's approach: they no use of both the context information and syntactic structure but we use the syntactic dependency and context information for determining whether the word is a product feature or non-product feature.

6. Conclusion

Product feature extraction is an important task of review mining and summarization. In this paper, we propose the approach of product feature extraction by using maximum entropy model with lexical and syntactic features. Our approach used a maximum entropy classifier extracting product features and includes the postprocessing step to discover the remaining product features in the reviews by matching the list of extracted product features against each word in the reviews. The results show that this approach is effective and suitable to be used for automatic product feature extraction. Furthermore, we have examined the extraction results. It has been found the errors are caused mostly by long sentences or complex sentences. In the future, we would like to add the sentence boundary detection to our model and increase the size of the training corpus in order to obtain more reasonable feature distributions and parameters.

7. References

[1] E. Riloff, W. Janyce, and W. Theresa, "Learning subjective nouns using extraction pattern bootstrapping," In *Proceedings of ACL SIGNLL Conference*, 2003, pp. 25-32.

[2] V. Hatzivassiloglou, and W. Janyce, "Effects of adjective orientation and gradability on sentence subjectivity," In *Proceedings of COLING Conference*, 2000.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," In *Proceedings of EMNLP Conference*, 2002.

[4] P.D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," In *Proceedings of ACL Conference*, 2002, pp. 417-424.

[5] S.M. Kim, and E. Hovy, "Determining the sentiment of opinion," In *Proceedings of COLING Conference*, 2004, pp. 1367–1373.

[6] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," In *Proceedings of HLT/EMNLP Conference*, 2005.

[7] M. Hu, and B. Liu, "Mining and summarizing customer reviews," In *the Proceedings of the ACM SIGKDD Conference*, 2004, pp. 168–177.

[8] A. M. Popescu, and O. Etzioni, "Extracting product features and opinions from reviews," In *the Proceedings of the EMNLP Conference*, 2005, pp. 339–346.

[9] B. Shi, and K. Chang, "Mining chinese reviews," In *the Proceedings of the ICDMW Workshops*, 2006.

[10] L. Zhuang, F. Jing, and X. Y. Zhu, "Movie review mining and summarization," In *the Proceedings of the CIKM Conference*, 2006.

[11] B. Liu, *Web Data Mining Exploring Hyperlinks, Contenrs, and Usage Data*. Springer, New York, 2007.

[12] M. Hu, and B. Liu, "Mining opinion features in customer reviews," In *the Proceedings of the AAAI Conference*, 2004.

[13] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," In *the Proceedings of the WWW Conference*, 2005.

[14] G. Carenini, R. T. Ng, and E. Zwart, "Extracting knowledge from evaluative text," In *the Proceedings of the K-CAP Conference*, 2005, pp. 11–18.

[15] J. Yi, and W. Niblack, "Sentiment mining in WebFountain," In *the Proceedings of the ICDE Conference*, 2005.

[16] E.T. Jaynes, "Information theory and statistical mechanics," *Physics Reviews*, 1957, vol 106, pp.620-630.

[17] E.T. Jaynes, *Probability theory: The logic of sciece*. Manuscript for book, 1998.
http://bayes.wustl.edu/etj/prob.html.

[18] C. Manning, and H. Schutze, *Foundations of statistical natural language processing*. MIT Press, MA: Cambridge, 1999.

[19] A.L. Berger, D.P. Stephen, and D.P. Vincent, "A maximum entropy approach to natural language processing," *Computational Linguistics*, 1996, vol. 22, no. 1, pp. 39–71.

[20] J.C. Reynar, and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," In *the Proceedings of the ANLP Conference*, 1997, pp. 16-19.

[21] A. Ratnaparkhi, *Maximum entropy models for natural language ambiguity resolution*, PhD thesis, University of Pennsylvania, 1998.

[22] H.L. Chieu, and H.T. Ng, "A maximum entropy approach to information extraction from semi-structured and free text," In *the Proceedings of the AAAI Conference*, 2002, pp. 786-791.

[23] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," In *the Proceedings of the EMNLP Conference*, 1996, pp. 133–141.

[24] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," In *the Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filteringe*, 1996.

[25] J. Darroch, and D. Ratcliff, "Generalized interative scaling for log-linear model," *The Annals of Mathematical Statistics* , 1972, vol. 43, no. 5, pp. 1470–1480.

[26] Stanford Lexicalized Parser - a probabilistic lexicalized NL CFG parser, 2006. http://nlp.stanford.edu/downloads/lex-parser.shtml.

[27] M. Hu, and B. Liu, Feature based summary of customer reviews dataset. http://www.cs.uic.edu/liub/FBS/FBS.html, 2004.

A Maximum Entropy Model for Product Feature Extraction in Online Customer Reviews

Gamgarn Somprasertsri
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand
s7066001@kmitl.ac.th

Pattarachai Lalitrojwong
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand
pattarachai@it.kmitl.ac.th

Abstract—**Product feature extraction is an important task of review mining and summarization. The task of product feature extraction is to find product features that customers refer to in their topic reviews. It would be useful to characterize the opinions which they review or express about the products. In this paper, we propose an approach to product feature extraction using a maximum entropy model. Maximum entropy is a probability distribution estimation technique. It is widely used for classification problems in natural language processing, such as question answering, information extraction, and part-of-speech tagging. The underlying principle of maximum entropy is that without external knowledge, one should prefer distributions that are uniform. Using a maximum entropy approach, at first we extract features from the corpus, train maximum entropy model with an annotated corpus, and then use it with additional product feature discovery to extract product features from customer reviews. Our experimental results show that this approach can work effectively for product feature extraction with 71.88% precision and 75.23% recall.**

Keywords—**product feature extraction, maximum entropy model, text mining, review mining and summarization**

I. INTRODUCTION

Recently, the number of online shopping customers has been increased due to the rapid growth of e-commerce, an increasing number of online merchants, and more and more people becoming comfortable with the internet. To enhance customer satisfaction, merchants and product manufacturers allow customers to review or express opinions on the products they buy from their websites, for instance, amazon.com, cnet.com, and epinions.com. Online customer reviews become the source of information which is very useful to both potential customers and product manufacturers. People read them for making a decision on whether to purchase the product. For a product manufacturer, knowing the preferences of customers is highly valuable for product development, marketing and consumer relationship management. Mining reviews mostly in free-form text can be extremely expensive. Besides, it is hard to find useful information from plenty of customer reviews. This trend has raised many interesting and challenging research topics such as subjectivity classification, sentiment classification, and review mining and summarization.

Subjectivity classification is distinguishing sentences or documents that present opinions from factual information, as in [1][2]. Many natural language processing applications could benefit from being able to distinguish between factual and subjective information such as question answering, information extraction, and so on. The task of sentiment classification is to judge whether a review expresses a positive or negative opinion. For example, [3][4] developed methods for document level sentiment classification. The systems assign a positive or negative sentiment for the whole review document. Sentiment of phrases and sentences has also been studied in [5][6]. Even if sentiment classification is useful, it does not find what the reviewer liked and disliked. Review mining and summarization is the task of producing a sentiment summary, which consists of sentences from reviews that capture the author's opinion. Review summarization is only interested in features or objects on which customers have opinions. It also determines whether the opinions are positive or negative. This makes it differ from traditional text summarization. Most existing works on review mining and summarization mainly focus on product reviews. For example, [7][8][9] concentrated on mining and summarizing reviews by extracting opinion sentences regarding product features. In another domain, [10] proposed a multi-knowledge based approach for movie review mining and summarization.

In general, mining and summarizing customer reviews involve three tasks (Figure 1): firstly, feature extraction identifies and extracts object features that have been commented in each review; secondly, sentiment assignment determines the polarity of each feature to be positive or negative; and thirdly, summary visualization summarizes the result in order to show this result more effectively. Product feature extraction is an important task of review mining and summarization. Identifying product features would be useful to characterize the opinions the customers review or express about the products. In this paper, we only focus on the first task of review mining and summarization. We propose a maximum entropy model for extracting product features. Our goal is to investigate whether the maximum entropy model is suitable for automatic product feature extraction. Our experimental results show that this approach is effective.

978-1-4244-1674-5/08 /$25.00 ©2008 IEEE

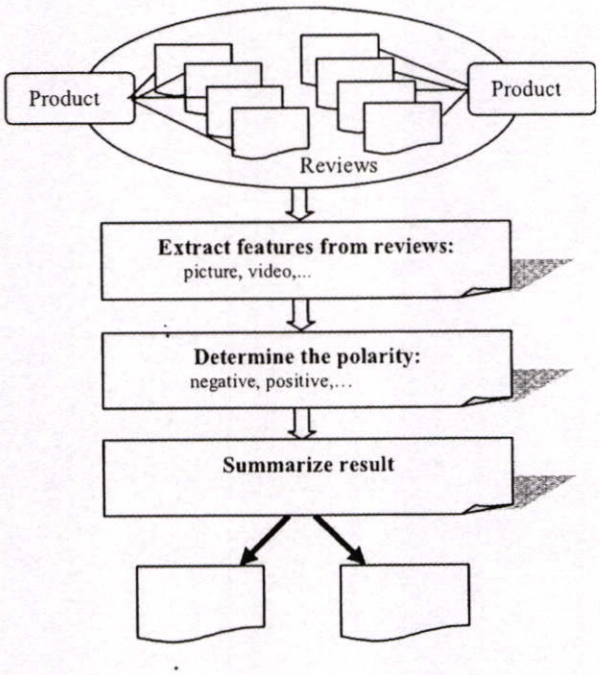


Figure 1. Review mining and summarization process

The rest of this paper is organized as follows. Section II describes related work on the task of product feature extraction. Section III introduces the maximum entropy model. Section IV discusses how to extract product features from online customer reviews using maximum entropy. Section V presents and discusses experimental results. Finally, Section VI concludes our work.

II. Related Work

Hu and Liu's work in [11] can be considered as the pioneer work on feature extraction from reviews. Their feature extraction algorithm is based on heuristics that depend on feature terms' respective occurrence counts. They use association rule mining based on the Apriori algorithm to extract frequent itemsets as explicit product features (only in the form of noun phrases). In association rule mining, the algorithm does not consider the position of the words in a sentence. In order to remove incorrect frequent features, they use feature pruning that consists of compactness pruning and redundancy pruning. To improve the work over [11], Liu, Hu, and Cheng [12] propose a technique based on language pattern mining to identify product features from pros and cons in reviews in the form of short sentences. They also make an effort to extract implicit features.

Popescu and Etzioni [8] developed an unsupervised information extraction system called OPINE extracting product features and opinions from reviews. OPINE first extracts noun phrases from reviews and retains those with frequency greater than an experimentally set threshold and then assesses those by OPINE's feature assessor for extracting explicit features. The assessor evaluates a noun phrase by computing a Point-wise Mutual Information score between the phrase and meronymy discriminators associated with the product class.

Carenini, Ng, and Zwart [13] proposed feature extraction for capturing knowledge from product reviews. In their method, the output of Hu and Liu's system [11] was used as the input to their system, and map the input to the user-defined taxonomy features hierarchy thereby eliminating redundancy and providing conceptual organization.

Finally, Yi and Niblack [14] developed a set of feature term extraction heuristics and selection algorithms for extracting a feature term from product reviews. The feature term is a part of relationship with the given topic, an attribute of relationship with the given topic, and an attribute of relationship with a known feature of the given topic. In the first step, they extract a noun phrase with the Beginning define Base Noun Phrase (bBNP) heuristics. Then, they select a feature term from the noun phrase using the likelihood score.

Our motivation for building the product feature extraction system described in this paper is that the grammars and word information may be useful for determining whether the word is a product feature or non-product feature. For this motivation, we will employ the maximum entropy model which can indeed combine various features of grammars or words into a probability model for product feature extraction in online customer reviews.

III. Maximum Entropy Model

The maximum entropy (ME) model was first described by Jaynes in [15] and more recently in a draft manuscript available on the Web [16]. The maximum entropy model is a framework for integrating information from many heterogeneous information sources for classification [17]. This model implements the intuition that the best model is the one consistent with the set of constraints imposed by the evidence but otherwise is as uniform as possible [18]. The maximum entropy approach has in recent years been used for a wide variety of classification problems in natural language processing, such as sentence boundary detection [19][20], information extraction [21], and part-of-speech tagging [22]. The underlying principle of maximum entropy is that without external knowledge, one should prefer distributions that are uniform. For our work, we use maximum entropy model to extract product features from online customer reviews. This task can be re-formulated as a classification problem, in which the task is to observe some linguistic context $x \in X$ and predict the correct linguistic class $y \in Y$.

We can implement classifier cl: $X \rightarrow Y$ with a conditional probability model by simply choosing the class y with the highest conditional probability in the context x:

$$cl(x) = \arg\max_{y} p(y \mid x) \quad (1)$$

The conditional probability $p(y|x)$ is defined as follows [20]:

$$p(y \mid x)=\frac{1}{Z(x)}\prod_{i=1}^{k}\alpha_i^{f_i(x,y)} \quad (2)$$

$$Z(x)=\sum_{y}\prod_{i}\alpha_i^{f_i(x,y)} \quad (3)$$

where y refers to the outcome, x is the history (or context), k is the number of features and $Z(x)$ is a normalization factor to ensure that $\sum_y p(y|x)=1$. Each parameter α_i, corresponds to one feature f_i and can be interpreted as a weight for that feature. The parameters α_i are estimated by a procedure called Generalized Iterative Scaling (GIS) [23]. This is an iterative method that improves the estimation of the parameters at each iteration.

Under the maximum entropy framework, the probability for a class y and object x depends solely on the features that are active for the pair (x, y), where a feature is defined here as a function f: $X \times Y \rightarrow \{0, 1\}$ that maps a pair (x, y) to either 0 or 1. The feature is defined as follows:

$$f_{y'}(x,y)=\begin{cases}1 & if\ y'=y\ and\ cp(x)=true\\ 0 & otherwise\end{cases} \quad (4)$$

where $cp(x)$ is contextual predication that returns true or false, corresponding to the presence or absence of useful information in some context, or history $x \in X$. For example, to predict which the class of target word belongs (as shown in Table I). The classifier considers surrounding context of the target word. If the target word is product feature and the previous word is "the", a feature function can be set as Equation 5. On the other hand, if the target word is non-product feature and it contains small letters, a feature function can be set as Equation 6.

$$f_i(x_j,y_j)=\begin{cases}1 & if\ y_j=YES\ and\ previousword="the"\\ 0 & otherwise\end{cases} \quad (5)$$

$$f_i(x_j,y_j)=\begin{cases}1 & if\ y_j=NO\ and\ capital(word_j)=False\\ 0 & otherwise\end{cases} \quad (6)$$

TABLE I. CLASSES DEFINED FOR THE CLASSIFICATION TASK

Class	Description
YES	Word claimed to be product feature
NO	Word claimed to be non-product feature

IV. PRODUCT FEATURE EXTRACTION

The main objective of this research is to identify product features from reviews, namely, product feature extraction. The product feature can be a brand name, a model name of a commodity, a property, a part, a feature of a product, a related concept, and a part of related concept [8]. In general, such product features are often the product itself or its specific features, such as *quality* (e.g. "*The picture quality is amazing*"). The product features mentioned in a review can be classified into two types: explicit and implicit features. For example, the sentence from a review of a DVD player, *"the sound is good"* shows the explicit product feature. It means that the customer is satisfied with the *sound* of the DVD player, and the *sound* is explicitly mentioned in the sentence. On the contrary, the sentence *"it does tend to run quite hot"* shows that the customer is talking about the *heat* of the DVD player, but the *heat* is not explicitly mentioned in the sentence. Therefore, *heat* is an implicit feature in this sentence. Extracting features from free-form text of a review is a challenging task because of the use of natural language.

A. The Overview of System

In this research work, we aim to extract explicit product features commented by customers. We leave finding implicit features to our future work. We define the product feature extraction problem as a classification task: given a sequence of words $(x_1, x_2, \ldots, x_n)$ in a sentence, we generate a sequence of labels $(y_1, y_2, \ldots, y_n)$ indicating whether the word is a product feature or non-product feature. We apply the maximum entropy model to extract the product features. The processes involve two phases: firstly to train the ME model; and secondly to test the model extracting product features from unlabeled reviews. We first prepare a training data set by manually labeling product features of reviews. The process of product feature extraction is shown in Figure 2 which can be described as follows:

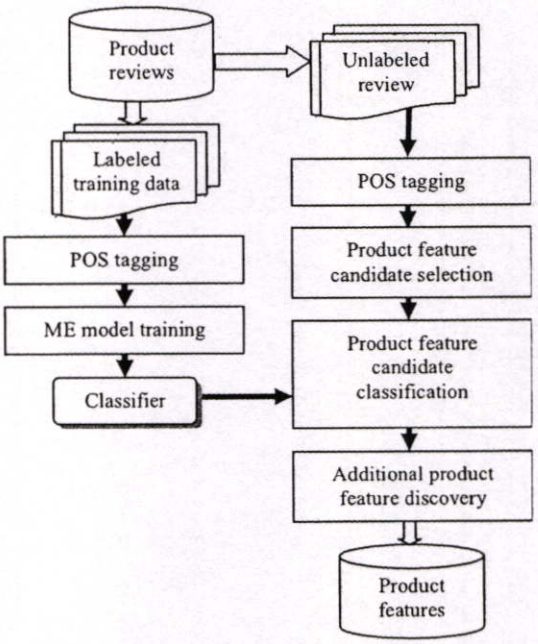


Figure 2. Process of product feature extraction

1) Part-of-speech tagging

We utilize the Stanford lexicalized parser [24] to analyze each review and tag the part of speech of each word. The following is an example of the output such as nouns, verbs, adjectives and so on. *The/DT macro/NN mode/NN is/VBZ*

exceptional/JJ ,/, the/DT pictures/NNS are/VBP very/RB clear/JJ and/CC you/PRP can/MD take/VB the/DT pictures/NNS with/IN the/DT lens/NN unbelievably/NN close/RB the/DT subject/NN.

2) ME model training

Training the ME model involves two steps. The first step of the maximum entropy approach is to extract features or important information in order to constrain the model accordingly. For product feature extraction, the features or important information can be the context information, the part of speech information and so on which will be explored in the next section. The second step is to train a model by using these features according to Equation 2.

3) Product feature candidate selection

For product feature extraction, the first step is POS tagging. After tagging the sentence, the next step is identifying product feature candidates. This step selects words from a tagged sentence such as nouns and adjectives, which are indicated by NN and JJ respectively. Using only nouns and adjectives is reasonable because most product features are nouns; however, some adjectives may appear as product features.

4) Product feature candidate classification

The trained model is employed to classify product feature candidates from unlabeled reviews after POS tagging. We will simply choose the class with the highest conditional probability p according to Equation 1.

5) Additional product feature discovery

This step aims to improve the performance of product feature extraction. After extracting product features by the ME based classifier, we applied a natural language processing technique to deal with compound product feature candidates which are not extracted by the classifier. The product feature candidates will be extracted if they or their head noun matches the product features extracted by the classifier. We take the sentence "*The scroll wheel was a nice idea to keep less clutter*" as an example. The word "*wheel*" is the head noun of "*scroll wheel*" extracted by the classifier, thus the "*scroll*" will be extracted as a product feature. The pseudo code of this process is provided in Figure 3.

```
1. Procedure ExtractRefinement(feature_list, review)
2. begin
3. for each sentence in review
4.   for each candidate w_i in sentence
5.   begin
6.     if (w_i match f_j in feature_list) and
         (w_i's ME score < threshold) then
7.       w_i's class = product feature
8.     else if (w_{i+1}'s POS tag is noun) and
           (w_{i+1}'s class is product feature) then
9.         w_i's class = product feature
10.  endfor;
11. endfor;
12. end
```

Figure 3. Pseudo code of additional product feature discovery

B. Features forProduct Feature Classification

To use the maximum entropy to extract product features, we define features or important information in order to constrain the model. We denote the features employed for learning as learning features, discriminative from the product features we discussed above. We compute several features automatically. Table II summarizes the features we used for our model and the symbols we will use in the rest of this paper. The features can be described as follows.

TABLE II. FEATURES IN OUR MODEL

Symbol	Feature name	Description
F1	Context	words in a [-4, +4] window centered on w_i
F2	Part-of-Speech Tag	POS tags in a [-4, +4] window centered on t_i
F3	Rare Word	words which occur less than five times in the training set
F4	Alphanumeric	words containing letters and numerals
F5	Capitalized	words starting with a capital letter

1) Context

Words preceding or following the target word may be useful for determining its category. For example, "*the sound is wonderful*". If the target word is "*sound*" and the following words are "*is*" and "*wonderful*", then this will help the model to classify "*sound*" as a product feature. The more context words analyzed, the better and more precise the results. However, widening the context window quickly leads to an explosion of the number of possibilities to calculate. In our experiment, a suitable window size is +/-4 words.

2) Part-of-Speech Tag

Part of speech tag is quite useful for identifying product features. Verbs and prepositions usually indicate the product feature boundaries whereas nouns and adjectives are usually good candidates for product features.

3) Rare Word

Rare word information may be useful for identifying product features. It is common that a customer review contains many things that are not directly related to product features. Different customers usually have different stories. Those frequent noun-noun phrases (non-rare words) are likely to be product features and infrequent noun-noun phrases (rare words) are likely to be non-product features. A rare word in our work denotes a word which occurs less than five times in the training set. The count of five was chosen by subjective inspection of words in the training data.

4) Orthography

Orthography is a characteristic of words such as words containing letters and numerals, words starting with a capital letter, and so on. This information may be useful for determining whether the word is a product feature or non-product feature. Two types of orthography considered in our model are alphanumeric and capitalized. Alphanumeric means words containing letters and numerals. Capitalized means

words starting with a capital letter. The regular expression of the alphanumeric and the capitalized are presented in Table III.

TABLE III. ORTHOGRAPHY FOR OUR MODEL

Feature name	Regular Expression
Alphanumeric	.*[A-Za-z].*[0-9].*|.*[0-9].*[A-Za-z].*
Capitalized	[A-Z][a-z]+

V. EXPERIMENTS

For our experiments, we used reviews on electronic products such as digital cameras and MP3 players from the Amazon web site. We annotated a randomly selected sample of 1,555 sentences for product feature extraction. Each word of each sentence was classified as a product feature or a non-product feature. This set of data was split into a training set of 1,255 sentences and a testing set of 300 sentences. Words in sentences are represented as vectors of binary features. The training originally yielded 13,066 features. We have used Maxent toolkit version 2.4.0 from [25]. The model was trained with features in the form of words in sentence contexts, POS tags, Orthography and rare word. The parameters of the maximum entropy model can be trained with 100 iterations of the Generalized Iterative Scaling algorithm which must be calculated repeatedly. Normally, it is a good "rule of thumb" to carry out 100 iterations [20]. More iteration would not increase the accuracy of the parameters.

The evaluation methods are precision (P), recall (R), and F-score (F). They are defined as follow:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F = \frac{2PR}{P + R}$$

where *TP* means the number of product features extracted correctly; *FP* means the number of words mistakenly claimed to be product features; and *FN* means the number of product features not extracted.

We conducted these experiments with three goals: firstly, to investigate how well our product feature extraction model with different window sizes of the contexts and POS tags; secondly, to investigate how well the model with different feature combinations performs on the customer reviews; and thirdly, to see how well our approach (ME model with additional product feature discovery) performs the product feature extraction compare to baseline (ME model without additional product feature discovery).

TABLE IV. PERFORMANCE OF THE MODEL USING CONTEXT WITH DIFFERENT WINDOW SIZES

Window Size	Precision (%)	Recall (%)	F-Score (%)
+/-1 word	**72.18**	61.21	66.25
+/-2 words	71.58	61.21	65.99
+/-3 words	69.66	61.68	65.43
+/-4 words	71.20	**62.38**	**66.50**

TABLE V. PERFORMANCE OF THE MODEL USING PART-OF-SPEECH TAG WITH DIFFERENT WINDOW SIZES

Window Size	Precision (%)	Recall (%)	F-Score (%)
+/-1 word	69.92	58.64	63.79
+/-2 words	72.38	58.18	64.51
+/-3 words	72.14	57.48	63.98
+/-4 words	**73.73**	**60.98**	**66.75**

TABLE VI. PERFORMANCE OF THE MODEL USING THE COMBINATION OF FEATURES

. Features	Precision (%)	Recall (%)	F-Score (%)
F1F2	76.80	57.24	65.60
F1F2F3	77.74	**57.94**	**66.40**
F1F2F4	77.53	57.24	65.86
F1F2F5	77.19	57.71	66.04
F1F2F3F4	**78.10**	57.48	66.22
F1F2F3F5	77.85	57.48	66.13
F1F2F4F5	76.88	57.48	65.78
F1F2F3F4F5	77.64	56.78	65.59

TABLE VII. THE PERFORMANCE OF SYSTEM FOR PRODUCT FEATURE EXTRACTION

	Precision (%)	Recall (%)	F-Score (%)
Baseline	**78.10**	57.48	66.22
Our approach	71.88	**75.23**	**73.52**

Tables IV and V show the results of using the maximum entropy model with different window sizes of the contexts and POS tags. The model performed well when it used contexts or part-of-speech tags in window size +/-4 words.

In Table VI, the first column indicates which combination of features was used in our model. It is very interesting to see that the system achieves a very low score of the recall when it used all features whereas it achieves a very high score of the precision when it used the combination of contexts, part-of-speech tags, rare word and alphanumeric features and achieves a high score of the recall when it used the combination of contexts, part-of-speech tags, and rare word features. This

phenomenon supports that context and part-of-speech information is useful for identifying product features, frequent noun-noun phrases (non-rare words) are likely to be product features and some product features contain letters and numerals such as SD100 and 7-megapixel.

We also compared our approach with baseline. The baseline used only the maximum entropy with the contexts, part-of-speech tags, rare word and alphanumeric features. Our approach used a maximum entropy classifier extracting product features and then applied a natural language processing technique to deal with compound product feature candidates by head word consistency. The performance of the system is shown in Table VII. The baseline achieves 57.48% recall and 78.10% precision. In our approach, the precision is decreased slightly. However, the recall is increased by 17.75% (over 75%). Besides, our approach performs better than using only the maximum entropy model by 7.3% F-score. This result shows that by using an appropriate additional product feature discovery method, our approach can achieve high recall in customer reviews.

We have examined the extraction results manually. It has been found the errors are caused mostly by long and complex sentences. Sometimes people write several sentences without clearly pausing between them. In such cases, the model can not detect sentence boundaries. So the words in the first sentences will be considered as words in the second sentences. This causes the analysis errors in the product feature extraction process. In the future, we can improve our approach by using syntactic dependencies, instead of context windows and adding sentence boundary detection to our model. We also need to increase the size of the training corpus in order to obtain more reasonable feature distributions and parameters. We expect that these improvements will yield an improved product feature extraction task.

VI. Conclusion

Review mining and summarization is the task of producing sentiment summary, which consists of sentences from reviews that capture the author's opinion. Product feature extraction is an important task of review mining and summarization. This task is to find out product features that customers refer to in their topic reviews. The goal of our work is to use a machine learning technique to perform automatic product feature extraction. Maximum entropy, a new approach to the task, is applied in order to estimate a function that performs classification of product features and non-product features. The performance of the system with additional product feature discovery can be measured by 71.88% precision, 75.23% recall, and 73.52% F-score. This result shows that the maximum entropy model is effective and suitable to be used for automatic product feature extraction.

References

[1] E. Riloff, W. Janyce, and W. Theresa, "Learning subjective nouns using extraction pattern bootstrapping," In Proceedings of ACL SIGNLL Conference, pp. 25-32, 2003.

[2] V. Hatzivassiloglou, and W. Janyce, "Effects of adjective orientation and gradability on sentence subjectivity," In Proceedings of COLING Conference, 2000.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," In Proceedings of EMNLP Conference, 2002.

[4] P.D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," In Proceedings of ACL Conference, pp. 417-424, 2002.

[5] S.M. Kim, and E. Hovy, "Determining the sentiment of opinion," In Proceedings of COLING Conference, pp. 1367–1373, 2004.

[6] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," In Proceedings of HLT/EMNLP Conference, 2005.

[7] M. Hu, and B. Liu, "Mining and summarizing customer reviews," In the Proceedings of the ACM SIGKDD Conference, pp. 168–177, 2004.

[8] A. M. Popescu, and O. Etzioni, "Extracting product features and opinions from reviews," In the Proceedings of the EMNLP Conference, pp. 339–346, 2005.

[9] B. Shi, and K. Chang, "Mining chinese reviews," In the Proceedings of the ICDMW Workshops, 2006.

[10] L. Zhuang, F. Jing, and X. Y. Zhu, "Movie review mining and summarization," In the Proceedings of the CIKM Conference, 2006.

[11] M. Hu, and B. Liu, "Mining opinion features in customer reviews," In the Proceedings of the AAAI Conference, 2004.

[12] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the Web," In the Proceedings of the WWW Conference, 2005.

[13] G. Carenini, R. T. Ng, and E. Zwart, "Extracting knowledge from evaluative text," In the Proceedings of the K-CAP Conference, pp. 11–18, 2005.

[14] J. Yi, and W. Niblack, "Sentiment mining in WebFountain," In the Proceedings of the ICDE Conference, 2005.

[15] E.T. Jaynes, "Information theory and statistical mechanics," Physics Reviews, vol 106, pp.620-630, 1957.

[16] E.T. Jaynes, Probability theory: The logic of science. Manuscript for book, 1998. http://bayes.wustl.edu/etj/prob.html.

[17] C. Manning, and H. Schutze, Foundations of statistical natural language processing. MIT Press, MA: Cambridge, 1999.

[18] A.L. Berger, D.P. Stephen, and D.P. Vincent, "A maximum entropy approach to natural language processing," Computational Linguistics, vol. 22, no. 1, pp. 39–71, 1996.

[19] J.C. Reynar, and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," In the Proceedings of the ANLP Conference, pp. 16-19, 1997.

[20] A. Ratnaparkhi, Maximum entropy models for natural language ambiguity resolution, PhD thesis, University of Pennsylvania, 1998.

[21] H.L. Chieu, and H.T. Ng, "A maximum entropy approach to information extraction from semi-structured and free text," In the Proceedings of the AAAI Conference, pp. 786-791, 2002.

[22] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," In the Proceedings of the EMNLP Conference, pp. 133–141, 1996.

[23] J. Darroch, and D. Ratcliff, "Generalized interative scaling for log-linear model," The Annals of Mathematical Statistics , vol. 43, no. 5, pp. 1470–1480, 1972.

[24] Stanford Lexicalized Parser - A probabilistic lexicalized NL CFG parser, 2006. http://nlp.stanford.edu/downloads/lex-parser.shtml.

[25] J. Baldridge, T. Morton, and G. Bierner, Java-based opennlp maximum entropy package: Maxent. http://maxent.sourceforge.net.

Extracting Product Features and Opinions from Product Reviews Using Dependency Analysis

Gamgarn Somprasertsri*
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand
s7066001@kmitl.ac.th

Pattarachai Lalitrojwong
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand
pattarachai@it.kmitl.ac.th

***Abstract*—In web pages, the reviews are written in natural language and are unstructured-free-texts scheme. Online product reviews is considered as a significant informative resource which is useful for both potential customers and product manufacturers. The task of manually scanning through large amounts of review one by one is computational burden and is not practically implemented with respect to businesses and customer perspectives. Therefore it is more efficient to automatically process the various reviews and provide the necessary information in a suitable form. The task of product feature and opinion is to find product features that customers refer to their topic reviews. It would be useful to characterize the opinions about product. In this paper, we propose an approach to extract product features and to identify the opinions associated with these features from reviews through syntactic information based on dependency analysis.**

Keywords-customer review; opinion extraction; opinion mining; dependency analysis

I. INTRODUCTION

Online customer reviews become a cognitive source of information which is very useful for both potential customers and product manufacturers. Customers have utilized this piece of this information to support their decision on whether to purchase the product. For product manufacturer perspective, understanding the preferences of customers is highly valuable for product development, marketing and consumer relationship management. In a general web page, the reviews are written in natural language scheme and are free of texts with unstructured paradigm. With the great and rapid growth of web contents, customer reviews become available where a customer is able to express opinions on products and services. This trend has seen increasingly attention in sentiment analysis or opinion mining. In the opinion mining community, there are many challenging research topics such as subjectivity classification, sentiment classification, and opinion summarization.

Subjectivity classification is a task for classifying the sentences or the documents which contain opinions from factual, as in [1][2]. It is useful for many natural language processing applications such as question answering, information extraction, and so on. The task of sentiment classification is to judge whether a review expresses a positive or negative opinion. For example, [3][4] developed methods for sentiment classification in document level. The systems assign a positive or negative sentiment for the whole review document. The sentiment of phrases and sentences has also been studied in [5][6]. Even if sentiment classification is useful, it does not find what the reviewer liked and disliked. Review mining and summarization is the task of producing a sentiment summary, which consists of sentences from reviews that capture the author's opinion. Review summarization is interested in features or objects on which customers have opinions. It also determines whether the opinions are positive or negative. This makes it differ from traditional text summarization. Most existing works on review mining and summarization mainly focus on product reviews. For example, [7][8][9] concentrated on mining and summarizing reviews by extracting opinion sentences regarding product features. In another domain, [10] proposed a multi-knowledge based approach for movie review mining and summarization.

In general, mining and summarizing customer reviews involve three tasks: firstly, feature and opinion extraction identifies object features that have been commented in each review; secondly, sentiment assignment determines the polarity of each feature to be positive or negative; and thirdly, summary visualization summarizes the result in order to show this result more effectively.

The high-level problem of opinion summarization addresses how to determine the opinion that an author expresses in natural language text with respect to a certain feature. Let us consider an example of a customer review of a digital camera.

"This camera is very easy to use. The viewing screen is easy to see and very clear. The pictures are clear and good color. To compare other digital cameras we have used, this one if definitely superior and we would highly recommend."

In this example, we can extract several phrases such as *"very easy to use"*, *"viewing screen is easy to see and very clear"*, and *"pictures are clear and good color"*. The phrases represent the customer's opinion rather than facts. Particularly, opinion words such as *"very easy to use"*, *"easy to see"*, *"very clear"*, *"clear"*, and *"good color"* are used to express customer's positive sentiment regarding the product features which are referred by *"to use"*, *"viewing screen"*, and *"picture"*. The task of manually scanning through large amounts of review one by one requires a lot of time and cost for both businesses and customers.

* Corresponding author. Faculty of Informatics, Mahasarakham University, Mahasarakham, Thailand, gamgarns@msu.ac.th

978-1-4244-5933-9/10/$26.00 ©2010 IEEE

In this study, we address how to associate descriptions of different product features with opinion expressions found in a review. Our goal is to develop ways to establish a correct relationship between the product feature (the topic of the sentiment) and the opinion word (the subjective expression of the product feature). We propose an approach to extract product features and opinions from product reviews through incorporating the syntactic information based on dependency analysis. Our work is mainly focused on product reviews but the methodology in general works for a boarder range of opinions.

The rest of this paper is constructed as follows: Section II presents related work. Section III briefly describes the syntactic information based on dependency analysis, and in Section IV, we illustrate our approach. Experimental results and discussion are given in Section V. Finally Section VI concludes this work.

II. Related Work

There are many methods developed for the solution to opinion summarization problems. Most researchers work on product reviews. Other researchers have studied in another domain [10] proposed a multi-knowledge based approach for movie review mining and summarization.

Hu and Liu's work in [7] can be considered as the pioneer work on feature-based opinion summarization. Their feature extraction algorithm is based on heuristics that depend on feature terms' respective occurrence counts. They use association rule mining based on the Apriori algorithm to extract frequent itemsets as explicit product features (only in the form of noun phrases). In association rule mining, the algorithm does not consider the position of the words in a sentence. In order to remove incorrect frequent features, they use feature pruning that consists of compactness pruning and redundancy pruning. To improve the work over [7], Liu, Hu, and Cheng [11] propose a technique based on language pattern mining to identify product features from pros and cons in reviews in the form of short sentences. They also make an effort to extract implicit features. Moreover, Carenini, Ng and Zwart [12] proposed feature extraction for capturing knowledge from product reviews. In their method, the output of Hu and Liu's system [7] was used as the input to their system, and the input was mapped to the user-defined taxonomy features hierarchy thereby eliminating redundancy and providing conceptual organization.

Popescu and Etzioni [8] developed an unsupervised information extraction system called OPINE, which extracted product features and opinions from reviews. OPINE first extracts noun phrases from reviews and retains those with frequency greater than an experimentally set threshold and then assesses those by OPINE's feature assessor for extracting explicit features. The assessor evaluates a noun phrase by computing a Point-wise Mutual Information score between the phrase and meronymy discriminators associated with the product class.

The work of Yi and Niblack [13] is based on a set of feature term extraction heuristics and selection algorithms for extracting a feature term from product reviews. The feature term is part of a relationship with the given topic, an attribute of a relationship with the given topic, and an attribute of a relationship with a known feature of the given topic. In the first step, they extract a noun phrase with the beginning define Base Noun Phrase (bBNP) heuristics. Then, they select a feature term from the noun phrase using the likelihood score.

Corresponding to these issues, we have carried out some studies on product feature extraction as reported in [14]. In our previous work, we have used combining lexical and syntactic features with the maximum entropy model for extracting the product features. There is an important difference between our approach and Hu and Liu's approach: they do not use both the context information and syntactic structure but we use the syntactic dependency and context information for determining whether the word is a product feature or non-product feature.

To identify the expressions of opinions associated with features. Some researchers considered that a product feature and its opinion words usually co-occur within a certain distance in the sentence. Hu and Liu [7] focused on adjacent adjectives that modify feature nouns or noun phrases. They use adjacent adjectives as opinion words that associated with features. Kim and Hovy [5] explored the following four sizes of regions which may contain both of product features and their opinions. The four regions are: (1) full sentences; (2) words between the opinion holder and the topic; (3) region 2 +/- two words; and (4) from the first word behind the holder to the end of sentences. In other research, Popescu and Etzioni [8] apply manual extraction rules in order to find the opinion words. This idea is similar to that of [7] and [5], but instead of using a window of size or adjacent adjectives they define extraction rules to find the expressions of opinions.

In conclusion, the above methods simply analyze co-occurences of expressions within a short distance or patterns. Some important links between product feature and opinion may be missed. In view of these limitations of the existing approaches, we proposed a method to exploit syntactic information to deal with the semantic relationship between the product feature and the opinion words. Our motivation is that the dependency relation may be useful for extracting the product features and identifying opinions that associate with product features in each sentence.

III. Syntactic Information Basaed on Dependency Analysis

Dependency grammars represent sentence structures as a set of dependency relationships. A dependency relationship is an asymmetric binary relationship between a word called head, and another word called modifier. Fig. 1 shows the dependency tree for a sentence *"The movie mode is also working great."*.

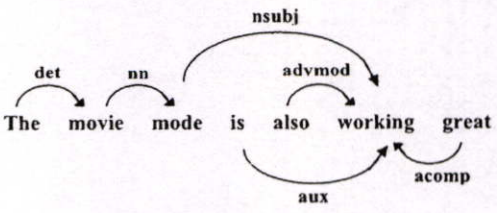


Figure 1. Example of dependency tree

The dependency tree is derived from the syntactic parse tree. We compute the syntactic parse tree by using the Stanford lexicalized parser [15].

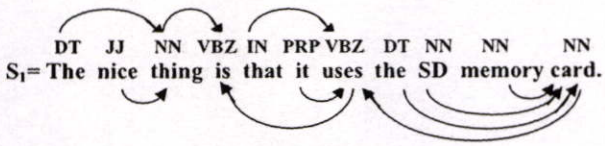
In the relationship of feature and opinion words, the terms have much more varied semantics. There is a large variety of linguistic constructions that express the relation between them. To reduce the variation of linguistic constructions, we assume that the shortest dependency path tracing from a product feature through the dependency tree to an opinion word gives a concrete syntactic structure expressing a relation between the pair. The dependency path and syntactic relationship are used together to find relationships between the product feature and opinion.

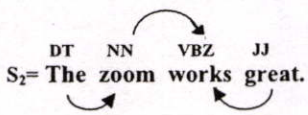
We adopted a syntactic relationship consisting of six different relationships as shown in Table 1.

TABLE I. SYNTACTIC RELATIONSHIPS

Relationship	Explanations
Child	Product feature depends on the opinion word
GrandChild	Product feature depends on the word which depends on the opinion word
Sibling	Both opinion word and product feature depend on the same word
Parent	Opinion word depends on the product feature
GrandParent	Opinion word depends on the word which depends on the product feature
Indirect	None of the above relationships

Fig. 2 we show two examples of sentences as dependency trees and the dependency paths linking product features and opinion words in the sentences.





S_1: Feature= "SD memory card" Opinion word = "nice"	NP → VBZ → VBZ ← NN ← JJ
S_2: Feature = "zoom" Opinion word = "great"	NN → VBZ ← JJ

Figure 2. Example of sentences as dependency tree and dependency paths of relation

IV. SYSTEM ARCHITECTURE

The architectural of the proposed approach consists of three main modules: firstly to perform such as parsing sentences, analyzing noun phrases, and analyzing dependency; secondly to train the ME model; and thirdly to extract product feature-opinion pairs from unlabeled reviews. We first prepare a training data set by manually labeling product feature-opinion pairs of reviews. Detailed steps are given as follows.

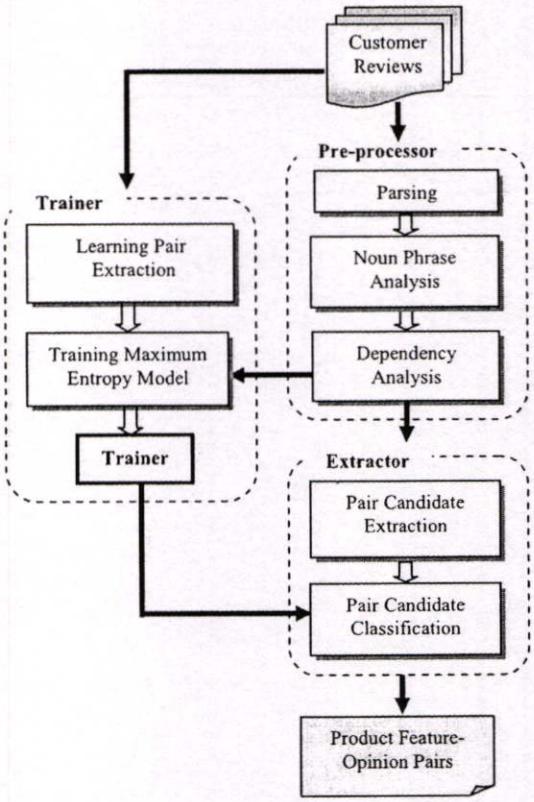


Figure 3. Architecture of System

1) Pre-processing module

To start the pre-processor, reviews are submitted to a pipeline including parsing, noun phrase analysis, and dependency analysis. First, we parse the review sentences with a Stanford lexicalized parser. The output syntactic parse trees are automatically converted into their dependency representations.

In general, most product features indicating words are nouns or noun phrases. Therefore, after parsing the sentence, the next step is to identify a noun phrase as a product feature candidate. We adopted linguistic filtering pattern (e.g. NN, NN NN, JJ NN, NN NN NN, JJ NN NN, JJ JJ NN, NN IN NN, and NN IN DT NN) to extracting noun phrases. Where NN, JJ, DT, and IN are the part-of-speech (POS) tags for noun, adjective, determiner, and preposition respectively defined by the Penn Treebank [16]. Next, for each noun phrase in every dependency

parse tree, we exhaustively generate potential syntactic information of noun phrase-adjective word pair.

2) Training module

The training ME model consists of three steps. Firstly, prepare training data which includes pre-processing and product feature-opinion pair annotation. Secondly, extract learning features of each pair in the training data. Thirdly, train the model by using a maximum entropy model. The result of the training model is the weight of each feature function.

To use the maximum entropy to extract product feature-opinion pairs, we define features or important information in order to constrain the model. We denote the features employed for learning as learning features, discriminative from the product features we discussed above. For each pair from the training data, we compute several features automatically. The features are as follow.

Product feature word: potential product feature as a noun or noun phrase.

Opinion word: potential opinion word as an adjective.

Dependency path: The shortest path between feature word and opinion word in a dependency graph.

Syntactic relationship: The classes of syntactic relationship between feature word and opinion word.

3) Extraction module

In order to extract product features and to identify of opinions associated with these features (product feature-opinion pairs), we rely on the observation that there are characteristic words used to describe the product feature and the opinion word. We found that most opinion expressions indicating words are adjectives whereas the nouns build the product features. Therefore, this module extracts pairs that are noun-adjective word pairs. Each such pair becomes a pair candidate. So, there may be more than one pair candidate on a sentence. Next, the trained model is used to predict product feature-opinion pair candidates from unlabeled reviews after parsing. We will simply choose the class with the highest conditional probability.

V. EXPERIMENT AND DISCUSSION

For our experiments, we used reviews on digital cameras from the Amazon web site. We used 1,250 sentences and conducted 5-fold cross validation on that dataset. This set of data was split into a training set of 80% and a testing set of 20%. As pre-processing we parsed this corpus using the Stanford lexicalized parser. We employed the OpenNLP Maxent [17] as our classification tool. The parameters of the maximum entropy model can be trained with 100 iterations of the Generalized Iterative Scaling algorithm. More than 100 iterations would not affect to the increase of accuracy of the parameters. To evaluate the method, we use precision, recall, and F-score to measure the effectiveness of our approach. When dealing with multiple datasets, we adopted the macro average to assess the overall performance across all datasets. The macro average is calculated by simply taking the average performance obtained for each dataset.

We compare the product features-opinion pairs extracted by our approach with co-occurrence approach. We conducted the experiments to compare with adjacent based method [7]. Beside, the pattern based method used by [8] is adopted to compare with our method. The result is compared with two approaches because they are the opinion summarization most relevant to our work and they have evaluated their performance on product review datasets.

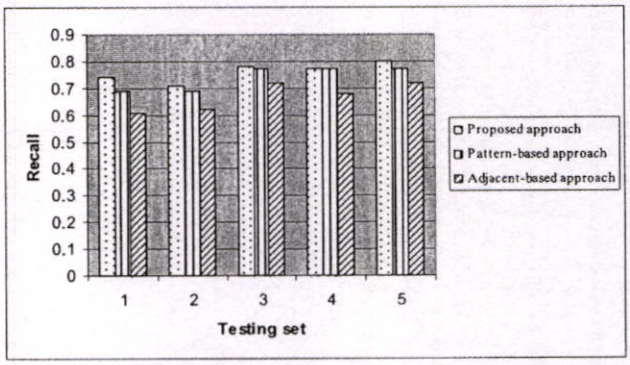


Figure 4. The recall of different approaches on test data

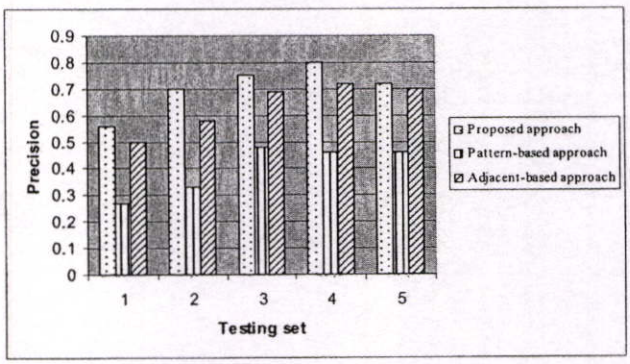


Figure 5. The precision of different approaches on test data

TABLE II. AVERAGE RESULTS ON TEST DATA OF DIFFERENT APPROACHES

Approaches	Precision	Recall	F-score
Proposed approach	**0.71**	**0.76**	**0.73**
Adjacent approach	0.64	0.67	0.65
Pattern approach	0.40	0.74	0.52

Fig. 4 and Fig. 5 show recall and precision of the different approaches on the test sets, respectively. These figures present the results obtained with the proposed approach, which outperforms adjacent-based approach and pattern-based approach. Table 2 demonstrates the average results calculated from five test sets. The macro-averaged F-score of the proposed approach is 0.73, whereas macro-averaged F-score of adjacent-based and pattern-based are 0.65 and 0.52, respectively. Their intuition is that a opinion expression

associated with a product feature will occur in its vicinity, whereas our approach takes advantage of the dependency relationship and machine learning. There are two reasons behind the satisfactory performance of the proposed approach. Firstly, the dependency path and syntactic relationship between product feature and opinion expression are useful for identifying the relation between them. Secondly, a maximum entropy classifier may be doing a good job on separating the opinion-relevant product feature pairs from the opinion-irrelevant product feature ones.

For further improvement, we have examined the extraction results manually. It has been found the errors are caused mostly by complex sentences. For example, a complex sentence such as *"It's difficult to take a good picture with this camera." confuses our method, because the sentence that describes negative expression and it's not relevant to extract "good picture"*. However, our method can solve the problem which is more than one product feature in a sentence.

VI. Conclusion

In this paper, we tried to solve the problem of extracting the product feature and identifying opinions that associate with product features in each review sentence. We have proposed a novel way to recognize product feature-opinion pairs which uses a probabilistic model with syntactic information based on dependency relationship. The experiments of extracting product features and identifying opinions associated with these features show encouraging results. As part of our future work, we would like to understand the reasons behind the unsatisfactory performance on the complex sentence. The possible improvements could consist of using more natural language processing techniques.

Acknowledgment

The authors would like to thank Mahasarakham University for overseas conference funding.

References

[1] Riloff E, Janyce W, Theresa W. Learning subjective nouns using extraction pattern bootstrapping. In Proc. of ACL SIGNLL Conference, 2003, pp. 25-32.

[2] Hatzivassiloglou V, Janyce W. Effects of adjective orientation and gradability on sentence subjectivity. In Proc. of COLING Conference, 2000.

[3] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques. In Proc. of EMNLP Conference, 2002.

[4] Turney P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proc. of ACL Conference, 2002, pp. 417-424.

[5] Kim S M, Hovy E. Determining the sentiment of opinion. In Proc. of COLING Conference, 2004, pp. 1367–1373.

[6] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis. In Proc. of HLT/EMNLP Conference, 2005.

[7] Hu M, Liu B. Mining and summarizing customer reviews. In the Proc. of the ACM SIGKDD Conference, 2004, pp. 168–177.

[8] Popescu A M, Etzioni O. Extracting product features and opinions from reviews. In the Proc. of the EMNLP Conference, 2005, pp. 339–346.

[9] Shi B, Chang K. Mining chinese reviews. In the Proc. of the ICDMW Workshops, 2006.

[10] Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization. In the Proc. of the CIKM Conference, 2006.

[11] Liu B, Hu M, Cheng J. Opinion observer: analyzing and comparing opinions on the web. In Proc. 14th Int. Conf. World Wide Web, Chiba, Japan, 2005, pp.342-351.

[12] Carenini G, Ng R T, Zwart E. Extracting knowledge from evaluative text. In Proc. 3rd Int. Conf. Knowledge Capture, 2005, pp.11–18.

[13] Yi J, Niblack W. Sentiment mining in WebFountain. In Proc. 21st Int. Conf. Data Engineering, 2005, pp.1073-1083.

[14] Somprasertsri G, Lalitrojwong P. Automatic Product Feature Extraction from Online Product Reviews Using Maximum Entropy with Lexical and Syntactic Features. In Proc. IEEE Int. Conf. Information Reuse and Integration, 2008, pp.250–255.

[15] Stanford Lexicalized Parser - a probabilistic lexicalized NL CFG parser, 2006. http://nlp.stanford.edu/downloads/lex-parser

[16] Marcus M P, Santorini B, Marcinkiewicz M A. Building a large annotated corpus of English: the penn treebank. Computational Linguistics, 1993, 19.

[17] Baldridge J, Morton T, Bierner G. Java-based opennlp maximum entropy package: Maxent. http://maxent.sourceforge.net

Journal of Universal Computer Science, vol. 16, no. 6 (2010), 938-955
submitted: 15/9/09, accepted: 4/3/10, appeared: 28/3/10 © *J.UCS*

Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization

Gamgarn Somprasertsri
(King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
s7066001@kmitl.ac.th)

Pattarachai Lalitrojwong
(King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
pattarachai@it.kmitl.ac.th)

Abstract: Online customer reviews is considered as a significant informative resource which is useful for both potential customers and product manufacturers. In web pages, the reviews are written in natural language and are unstructured-free-texts scheme. The task of manually scanning through large amounts of review one by one is computational burden and is not practically implemented with respect to businesses and customer perspectives. Therefore it is more efficient to automatically process the various reviews and provide the necessary information in a suitable form. The high-level problem of opinion summarization addresses how to determine the sentiment, attitude or opinion that an author expressed in natural language text with respect to a certain feature. In this paper, we dedicate our work to the main subtask of opinion summarization. The task of product feature and opinion extraction is critical to opinion summarization, because its effectiveness significantly affects the performance of opinion orientation identification. It is important to properly identify the semantic relationships between product features and opinions. We proposed an approach for mining product feature and opinion based on the consideration of syntactic information and semantic information. By applying dependency relations and ontological knowledge with probabilistic based model, the result of our experiments shows that our approach is more flexible and effective.

Keywords: Opinion Mining, Opinion Summarization, Text Mining, Customer Feedback, Dependency Grammars, Maximum Entropy
Categories: I.2.7, H.2.8, H.3.1, H.3.5

1 Introduction

Recently, a number of online shopping customers have dramatically increased due to the rapid growth of e-commerce, and the increase of online merchants. To enhance the customer satisfaction, merchants and product manufacturers allow customers to review or express their opinions on the products or services. The customers can now post a review of products at merchant sites, e.g., amazon.com, cnet.com, and epinions.com. These online customer reviews, thereafter, become a cognitive source of information which is very useful for both potential customers and product manufacturers. Customers have utilized this piece of this information to support their decision on whether to purchase the product. For product manufacturer perspective, understanding the preferences of customers is highly valuable for product development, marketing and consumer relationship management.

In a general web page, the reviews are written in natural language scheme and are free of texts with unstructured paradigm. In comparison, numerical and categorical data are well structured, which make them relatively easy to handle. On the contrary, customer reviews are unstructured data. To be handled, these data demand knowledge from different areas, e.g., database, information retrieval, information extraction, machine learning, and natural language processing. With the great and rapid growth of web contents, customer reviews become available where a customer is able to express opinions on products and services. This trend has seen increasingly attention in sentiment analysis or opinion mining. In the opinion mining community, there are many challenging research topics such as subjectivity classification, sentiment classification, and opinion summarization.

Subjectivity classification is the task of classifying the sentences or the documents which contain opinions from factual, for instance in [Riloff, 03]. It is useful for many natural language processing applications such as question answering, information extraction, and so on. The task of sentiment classification is to judge whether a review expresses a positive or negative opinion. For example, Pang et al [Pang, 02] developed methods for sentiment classification in document level. The systems assign a positive or negative sentiment for the whole review document. The sentiment of phrases and sentences has also been studied in [Wilson, 05]. Even if sentiment classification is useful, it does not imply the underlining information, i.e. what the reviewer liked and disliked. Opinion summarization [Hu, 04, Popescu, 05, Gamon, 05, Yi, 05, and Carenini, 06] is the task of producing a sentiment summary, which consists of sentences from reviews that capture the author's opinion. The summarization task is interested in features or objects on which customers have opinions. This is different from traditional text summarization that involves reducing a larger corpus of multiple documents into a short of paragraph that conveys the meaning of text. The product reviews on the Web are in three formats [Liu, 07]:

- Format 1 - Pros, cons and the detailed review: The reviewers describe pros and cons in the form of short phrases and also write the detail of reviews separately.
- Format 2 - Pros and cons: The reviewers describe pros and cons in the form of full sentences separately.
- Format 3 - Free format: The reviewers write the reviews in the free form that no separation of pros and cons.

In format 1, pros and cons usually consist of short phrases and incomplete sentences, for example "*pros: fabulous photo quality, large LCD, great battery life, great features*". The reviews of format 2 and 3 usually consist of long sentences and complete sentences, for example "*I have taken hundreds of photos with it and i continue to be amazed by their quality*". However, the product features and opinions extraction from reviews of format 2 and 3 is more challenge because the complete sentences are more complex and contain a large amount of irrelevant information. The task of manually scanning through large amounts of review one by one requires a lot of time and cost for both businesses and customers. Therefore, a good summarization system can help them in getting the required and relevant information without going through all the reviews present on the site.

The high-level problem of opinion summarization addresses how to determine the opinion that an author expresses in natural language text with respect to a certain feature. Let us consider an example of a customer review of a digital camera.

"This camera is very easy to use. The viewing screen is easy to see and very clear. The pictures are clear and good color. To compare other digital cameras we have used, this one if definitely superior and we would highly recommend."

In this example, we can extract several phrases such as "*very easy to use*", "*viewing screen is easy to see and very clear*", and "*pictures are clear and good color*". The phrases represent the customer's opinion rather than facts. Particularly, opinion words such as "*very easy to use*", "*easy to see*", "*very clear*", "*clear*", and "*good color*" are used to express customer's positive sentiment regarding the product features which are referred by "*to use*", "*viewing screen*", and "*picture*".

This study, we address the specific problem that is how to associate descriptions of different product features with opinions found in reviews of format 3. The task is not only technically challenging – applying natural language processing, but also very useful in practice. In feature-opinion mining, most of the existing researches usually depend on the co-occurrence of product features and opinion words. The methods acquire relations based on fixed position of words. However, the approaches are not effective for many cases. Look at the following review sentences.

(1) It has movie mode that works *good* for a digital camera.

(2) It is *great* having the LCD display.

(3) I bought my canon g3 about a month ago and i have to say i am very *satisfied*.

(4) The *nice* thing is that it uses the SD memory card.

In these samples, the words in underline are product feature and the words in italic are opinion. The approach of co-occurrence of words is not the way to deal with this kind of problem. In this paper, our goal is to develop ways to establish a correct relationship between the product feature (the topic of the sentiment) and the opinion word (the subjective expression of the product feature). The basic purpose of our approach is to mine the product features and opinion words that associate with product features in each sentence.

The remainder of the paper is organized as follows. Section 2 describes previous work on the task of product feature and opinion extraction. Section 3 introduces dependency relations for product feature-opinion mining. Section 4 discusses how to mine product feature-opinion pairs from online customer reviews. Section 5 presents and discusses the experimental results. Finally, Section 6 concludes our work.

2 Previous Work

Opinion summarization essentially consists of three main tasks. The first task of opinion summarization is to extract the features of a product and to identify opinions that associate with product features in each sentence and then identify the opinion orientation. Finally produce a structured sentence list according to the feature-opinion pairs as the summary. The task of product feature and opinion extraction is critical to opinion summarization, because its effectiveness significantly affects the performance of opinion orientation identification. Many previous works [Hu, 04, Popescu, 05, Liu, 05, Yi, 05, and Zhuang, 06] usually depend on the co-occurrence of words.

Hu's work in [Hu, 04] can be considered as the pioneer work on feature-based opinion summarization. Their feature extraction algorithm is based on heuristics that depend on feature terms' respective occurrence counts. They use association rule mining based on the Apriori algorithm to extract frequent itemsets as explicit product features (only in the form of noun phrases). Association rule is an implication of the form X=>Y, where X and Y are database itemsets. Two measures have been developed to evaluate association rules, which are support and confidence. Itemsets that have support at least equal to minimum support are called frequent itemsets [Daly, 04]. In Hu's work, each resulting frequent itemset is a possible feature. They define an itemset as frequent if it appears in more than 1% minimum support of the review sentences. In this approach, the algorithm does not consider the position of the words in a sentence. In order to remove incorrect frequent features, they use feature pruning that consists of compactness pruning and redundancy pruning. To improve the work over Hu et al, Liu et al [Liu, 05] proposed a technique based on language pattern mining to identify product features from pros and cons in reviews in the form of short sentences. They also make an effort to extract implicit features. Moreover, Carenini et al [Carenini, 05] proposed feature extraction for capturing knowledge from product reviews. In their method, the output of Hu's system was used as the input to their system, and the input was mapped to the user-defined taxonomy features hierarchy thereby eliminating redundancy and providing conceptual organization. To identify the expressions of opinions associated with features. Hu et al focused on adjacent adjectives that modify feature nouns or noun phrases. They use adjacent adjectives as opinion words that associated with features. For each sentence in reviews, if it contains any frequent feature, extract the nearby adjective. It is considered an opinion.

Popescu et al [Popescu, 05] developed an unsupervised information extraction system called OPINE, which extracted product features and opinions from reviews. OPINE first extracts noun phrases from reviews and retains those with frequency greater than an experimentally set threshold and then assesses those by OPINE's feature assessor for extracting explicit features. The assessor evaluates a noun phrase by computing a Point-wise Mutual Information score between the phrase and meronymy discriminators associated with the product class. Popescu et al apply manual extraction rules in order to find the opinion words. This idea is similar to that of Hu et al [Hu, 04], but instead of using adjacent adjectives they define extraction rules to find the expressions of opinions. For example,

If $\exists(M, NP = \text{f}) \rightarrow \text{po} = M$: (expensive) scanner

If $\exists(S = \text{f}, P, \text{O}) \rightarrow \text{po} = O$: Lamp has (problems)

If $\exists(S, P, O = \text{f}) \rightarrow \text{po} = P$: I (hate) this scanner

If $\exists(S = \text{f}, P) \rightarrow \text{po} = P$: Program (crashed)

M = modifier, NP = noun phrase, S = subject, P = predicate, O = object, f = feature and po = potential opinion

Yi et al [Yi, 05] developed a set of feature term extraction heuristics and selection algorithms for extracting a feature term from product reviews. The feature term is a part of relationship with the given topic, an attribute of relationship with the given topic, and an attribute of relationship with a known feature of the given topic. In the first step, they extract a noun phrase with the Beginning define Base Noun Phrase (bBNP) heuristics. Then, they select a feature term from the noun phrase using the

likelihood score. As a processing step to opinion extraction, they utilized some patterns based on sentiment extraction pattern such as

<"impress" + PP(by;with)>: the target or feature is subject phrase (SP) and the opinion is "impress"

<"take" + OP SP>: the target or feature is subject phrase (SP) and the opinion is object phrase (OP)

Zhuang et al [Zhuang, 06] studied in movie review domain. They proposed a multi-knowledge based approach for movie review mining and summarization. They used the keyword list and dependency relation templates together to mine explicit feature-opinion pairs. For example,

NN – amod – JJ

NN – nsubj – JJ

NN – nsubject – VB – dobj - NN

In conclusion, the above methods acquire relations based on explicit adjacency. They simply analyze co-occurences of expressions within a short distance or patterns. Some important links between product feature and opinion may be missed. In view of these limitations of the existing approaches, we proposed a method to exploit syntactic information and semantic information to deal with the semantic relationship between the product feature and the opinion words. Our motivation is that the dependency relation may be useful for extracting the product features and identifying opinions that associate with product features in each sentence. In addition, the idea behind this method is to use machine learning to automatically replace manual extraction of rules to identify the expressions of opinions associated with features.

3 Dependency Relations for Feature-Opinion Mining

Dependency grammars represent sentence structures as a set of dependency relationships. A dependency relationship [Melcük, 87] is an asymmetric binary relationship between a word called head or governor, and another word called modifier or dependent. The dependency of words will form a dependency tree. The syntactic structure of a sentence consists of dependencies shown in Figure 1.

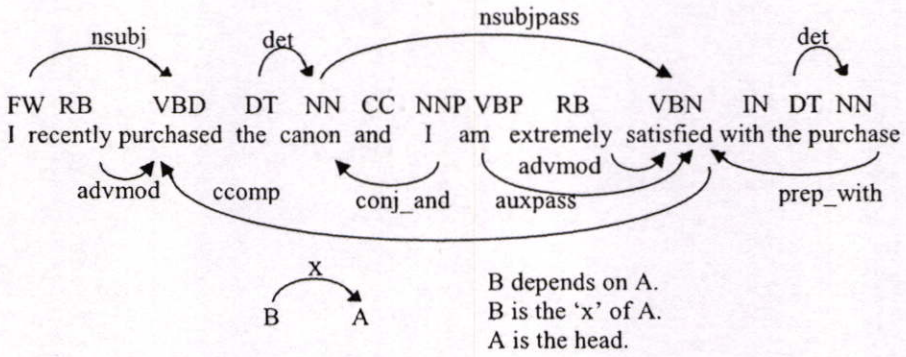


Figure 1: The syntactic structure of a sentence consists of dependencies

Each relationship has a word as head. The other is the dependent. A word has one head at most. However, a word may have several dependents.

With these relations defined by the dependency tree, we find there are five relations for mining product feature and opinion as following. PF refers to product features, O refers to opinion words and A refers to ancestors.

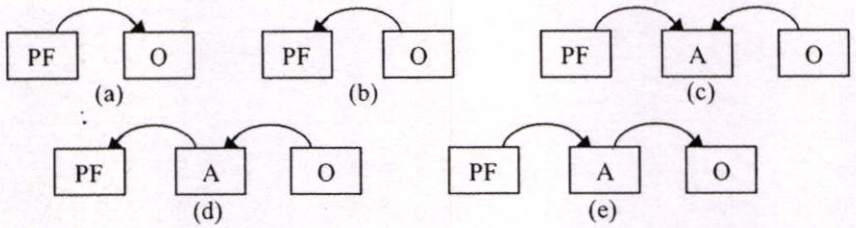


Figure 2: The relations of dependency sub-trees for product feature-opinion mining

1) **Child**: The product features are in the children as (a) in Figure 2. In such relation, the product feature is the subject or object of the verbs and the opinion word is a verb or a complement of a copular verb, for example

(1) "*I like this camera.*"

Dependency relation:

{nsubj(like-2, I-1), det(camera-4, this-3), dobj(like-2, camera-4)}

The dependencies are written abbreviated_relation_name(head, dependent) where the head and the dependent are words in the sentence to which the word number in the sentence is append. In the brackets, the first word is the parent and the second word is the child. In (1), the word "*camera*" is the product feature. The word "*like*" is the opinion. The "*camera*" is the direct object of the verb "*like*".

(2) "*The battery life is good.*"

Dependency relation:

{det(life-3, The-1), nn(life-3, battery-2), nsubj(good-5, life-3), cop(good-5, is-4)}

The phrase "*battery life*" is the product feature. The word "*good*" is the opinion. The "*battery life*" is a noun phrase which is the subject. The "*good*" is the complement of the copular verb.

2) **Parent**: The product features are in the parents as (b) in Figure 2. In such relation, the opinion words are in the modifiers of product features, which include adjectival modifier, relative clause modifier, etc., for example

(3) "*I have found that this camera take incredible pictures.*"

Dependency relation:

{nsubj(found-3, I-1), aux(found-3, have-2), complm(take-7, that-4), det(camera-6, this-5), nsubj(take-7, camera-6), ccomp(found-3, take-7), amod(pictures-9, incredible-8), dobj(take-7, pictures-9)}

The word "*picture*" is the product feature. The word "*incredible*" is the opinion which is the adjectival modifier of a word "*picture*".

3) **Sibling**: The product features and the opinion words are in the children of the same ancestor as (c) in Figure 2. In such relation, the opinion word may also be in an adverbial modifier, a complement of the verb, or a predicative, for example

(4) "*The pictures some time turn out blurry.*"
Dependency relation:
{det(picture-2, The-1), nsubj(turns-5, picture-2), det(time-4, some-3), dep(turns-5, time-4), dep(blurry-7, out-6), acomp(turns-5, blurry-7)}

The word "*picture*" is the product feature which is the subject. The word "*blurry*" is the opinion which is the adverbial modifier of a verb.

4) **Grand Parent**: The product features are in the parents of the words that are in the parents of the opinion words as (d) in Figure 2. In such relation, the opinion words are adjectival complement of modifiers of product features, for example

(5) "*It has movie mode that works good for a digital camera.*"
Dependency relation:
{nsubj(has-2, It-1), nn(mode-4, movie-3), dobj(has-2, mode-4), rel(works-6, that-5), rcmod(mode-4, works-6), acomp(works-6, good-7), det(camera-11, a-9), amod(camera-11, digital-10), prep_for(good-7,camera-11)}

The phrase "*movie mode*" is the product feature. The word "*good*" is the opinion which is the adverbial complement of relative clause modifier of noun phrase "*movie mode*".

5) **Grand Child**: The product features are in the children of the words that are in the children of the opinion words as (e) in Figure 2. In such relation, the product feature is the subject or object of the complements and the opinion word is a verb or a complement of a copular verb, for example

(6) "*It's great having the LCD display.*"
Dependency relation:
{nsubj(great-3, It-1), cop(great-3, 's-2), xcomp(great-3, having-4), det(display-7, the-5), nn(display-7, LCD-6), dobj(having-4, display-7)}

The phrase "*LCD display*" is the product feature. The word "*great*" is the opinion which is the complement of a copular verb. The "*LCD display*" is the object of a clausal complement.

4 Mining Product Feature-Opinion

In this section, we described our methods to mine product feature-opinion from online customer reviews. The product feature can be a brand name, a model name of a commodity, a property, a part, a feature of a product, a related concept, or a part of a related concept [Popescu, 05]. Section 4.1 explains some pre-processing steps. The core methods are described in Section 4.2 and Section 4.3. Figure 3 gives the architecture overview for our approach.

4.1 Pre-Processing

To start the pre-processing, reviews are submitted to a pipeline including parsing and dependency analysis. Firstly, we parse the review sentences by using the Stanford Parser. After that we exhaustively generate a dependency tree as shown in Figure 1.

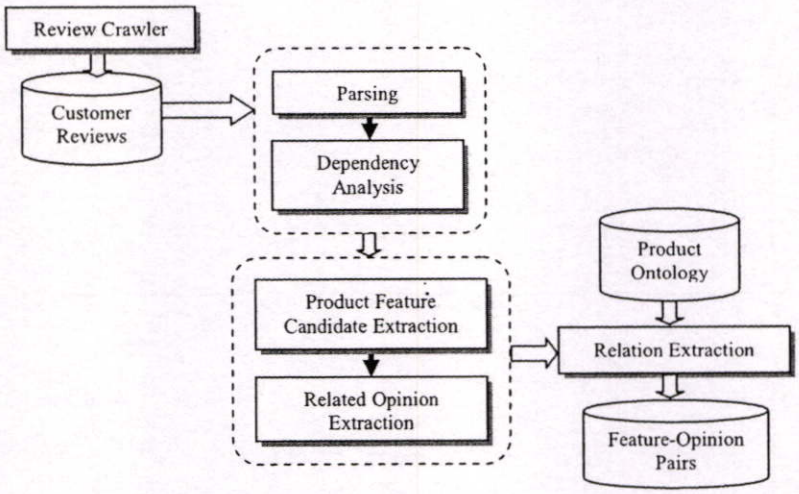


Figure 3: The architecture of our approach

4.2 Product Feature-opinion Candidate Extraction

When mining product feature-opinion, we first identifies product feature on which many customers have expressed their opinions. If a product feature appears, we will search for the related opinions and product features.

4.2.1 Product Feature Candidate Extraction

In general, most product features indicating words are nouns or noun phrases. Therefore, after parsing the sentence, the next step is to identify a noun phrase as a product feature candidate. We adopted linguistic filtering pattern and General Inquirer Dictionary [Stone, 66] to extracting product feature candidate. We also discard stop words to reduce noise. A definite linguistic filtering pattern is a noun phrase as the following patterns:

-NN,
-NNNN, JJ NN
-NN NN NN, JJ NN NN, JJ JJ NN, NN IN NN
-NN IN DT NN

where NN, JJ, DT, and IN are the POS tags for noun, adjective, determiner, and preposition respectively defined by the Penn Treebank [Marcus 93]. Algorithm 1 demonstrates the process to extract all the product feature candidates in reviews.

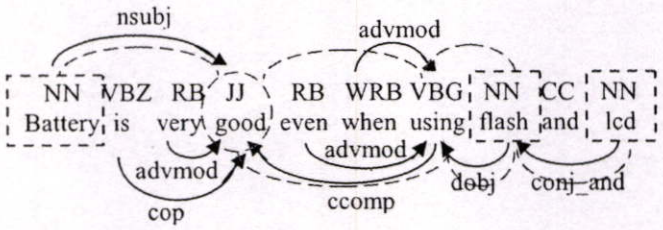
4.2.2 Related Opinion Extraction

This step is to identify product feature-opinion candidates. For each product feature candidate in every dependency parse tree, we search for the related opinion words. Some adjectives and verbs may be used for both favorable and unfavorable predictes. Thus, this paper uses adjectives and verbs as opinion words. The procedure of extracing opinions as following manner (Algoritm 2).

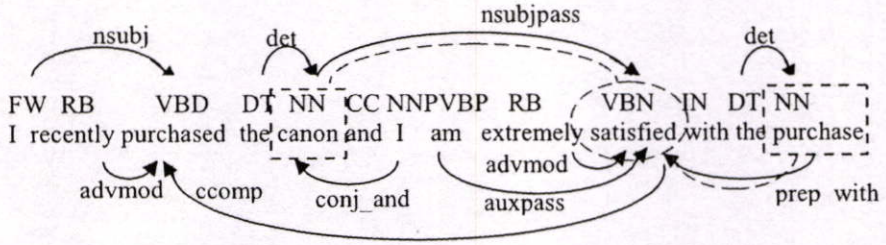
Algorithm 1. Pseudo-Code for extracting product feature candidates

```
//Input:   S – Set of tagged sentences; s = s1,s2,…,sm
           P – Set of noun phrase patterns
          GI – Set of word in GI dictionary
//Output: PS – Set of product feature candidates
PS = ø
For each tagged sentence sn ∈ S
   PC = ø
       For i=1 to end of sentence sn
          If  i < Length(sn) – 2  Then  x = 3
          Else If  i = Length(sn) – 2  Then  x = 2
               Else If  i = Length(sn) – 1  Then  x = 1
                    Else x = 0
                    End
               End
          End
          For j = x to 0
               GT = Ti to Ti+j        /* POS Tag of wordi to wordi+j of sn */
               GW = wordi to wordi+j
               If  GT ∈ P and GW ∉ GI then
                  i = i+j
                    PC = PC ∪ GW
                    Break
               End
          End
    PS = PS ∪ PC
End
```

Consider the following examples: "*Battery is very good even when using flash and LCD.*" and "*I recently purchased the Canon and I am extremely satisfied with the purchase.*" Figure 4 shows the procedure of product feature-opinion candidate extraction. Firstly we find the product features, and then find the opinions through the dependency tree (in the manner as Algorithm 2). In these samples, the words in circle shape are the effective opinions of the product feature candidates in square shape. We can extract several pairs such as *(battery, good)*, *(flash, good)*, *(lcd, good)*, *(cannon, satisfied)*, and *(purchase, satisfied)*. Each of such pairs becomes a product feature-opinion candidate. After product feature-opinion candidate extraction, we predict the opinion-relevant product feature relation using the probabilistic based model.



(a) "*Battery is very good even when using flash and LCD.*"



(b) "*I recently purchased the Canon and I am extremely satisfied with the purchase.*"

Figure 4: Example product feature-opinion candidate extraction

Algorithm 2. Pseudo-Code for extracting the product feature-opinion pairs

```
//Input     DT – Set of dependency trees
            PS – Set of product feature candidates in each sentence
            GI – Set of word in GI dictionary
//Output    FOS – Set of product feature-opinion pairs
PairExtract(dt_i, ps_i) /* Return set product feature-opinion pairs of each dependency tree */
FO = ø
  For m=1 to end of product feature candidate ps_i
     Rnode = ps_i(m)                    /* Initial product feature candidate to root node */
     f = FirstVisit(dt_i,Rnode)         /* Find first neighbor, return -1 if no neighbor */
     While (f ~= -1)
        If (neighbor is adjective) or (neighbor is adveb and ∈ GI) then
           pair = [Rnode,neigbor]       /* product feature-opinion pair */
           FO = FO ∪ pair
           f = -1
        Else
           f = NextVisit(dt_i,Rnode)    /* Find next neighbor, return -1 if no neighbor */
        End
     End
  End
PairExtract(DT, PS, GI)                 /* Return set of product feature-opinion pairs */
  For each dependency tree dt_i ∈ DT
      FO = PairExtract(dt_i, ps_i)
  End
```

4.3 Predicting a Relation by Maximum Entropy Model

Maximum entropy model was first described by Jaynes [Jaynes, 57]. The maximum entropy model chooses the least biased distribution, which maximizes uncertainty in the distribution subject to given constraints [Tan, 07]. For our work, we use maximum entropy model to predict the opinion-relevant product feature relation. This task can be re-formulated as a classification problem, in which the task is to observe some linguistic context $x \in X$ and predict the correct linguistic class $y \in Y$. We can design classes such as opinion-relevant product feature and opinion-irrelevant product feature. We can implement classifier cl: $X \rightarrow Y$ with a conditional probability model by simply choosing the class y with the highest conditional probability p in the context x:

$$cl(x) = \arg\max_{y} p(y \mid x) \tag{1}$$

The conditional probability $p(y|x)$ is defined as follows [Ratnaparkhi, 98]:

$$p(y \mid x) = \frac{1}{Z(x)} \prod_{i=1}^{k} \alpha_i^{f_i(x,y)} \tag{2}$$

$$Z(x) = \sum_{y} \prod_{i} \alpha_i^{f_i(x,y)} \tag{3}$$

where y refers to the outcome, x is the history (or context), k is the number of features and $Z(x)$ is a normalization factor to ensure that $\sum_y p(y|x)=1$. Each parameter α_i corresponds to one feature f_i and can be interpreted as a weight for that feature.

We use the Generalized Iterative Scaling (GIS) algorithm [Darroch, 72] to estimate parameters or weights of the selected features. Under the maximum entropy framework, the probability for a class y and object x depends solely on the features that are active for the pair (x, y), where a feature is defined here as a function f: $X \times Y \rightarrow \{0, 1\}$ that maps a pair (x, y) to either 0 or 1. The feature is defined as follows:

$$f_{cp,y'}(x,y) = \begin{cases} 1 & \text{if } y = y' \text{ and } cp(x) = true \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $cp(x)$ is contextual predication that returns true or false, corresponding to the presence or absence of useful information in some context, or history $x \in X$. For example, to predict which the class of product feature-opinion candidate belongs (as shown in Table 1). The classifier considers dependency relation of the target product feature-opinion candidate. Supposing the opinion word depends on the product feature. The relation of the target product feature-opinion pair is parent, a feature function can be set as follows:

$$f_i(x_j, y_j) = \begin{cases} 1 & \text{if } y_j = YES \text{ and } \mathrm{Rel}(PARENT) = true \\ 0 & \text{otherwise} \end{cases}$$

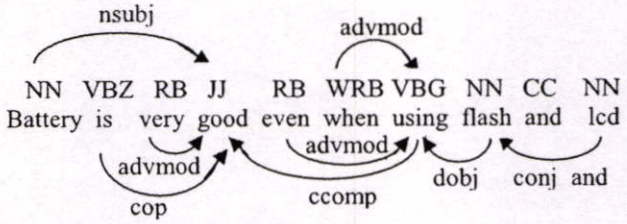
Class	Description
YES	Feature-opinion pair claimed to be opinion-relevant product feature
NO	Feature-opinion pair claimed to be opinion-irrelevant product feature

Table 1: Classes defined for the classification task

In order to use the maximum entropy to classify product feature-opinion candidates, we define important information in order to constrain the model. We use syntactic information to classifying product feature-opinion pair. One of the chalenges for this problem is dut to the wide variation of surface text. To reduce the variation of linguistic constructions, we assume that the shortest dependency path tracing from a product feature through the dependency tree to an opinion word gives a concrete syntactic structure expressing a relation between the pair. Our idea is to learn such patterns from the dependency paths for each relationship. Furthermore, we attempt to capture relating product feature and opinion using dependency relations between them. For our work, we adopted a dependency relation consisting of six different relations as presented in Table 2.

Relation	Description
Parent	Opinion depends on the product feature.
Child	Product feature depends on the opinion.
Sibling	Both opinion and product feature depend on the same word.
Grandparent	Opinion depends on the word which depends on the product feature.
Grandchild	Product feature depends on the word which depends on the opinion.
Indirect	None of the above relations

Table 2: Dependency relations for product feature-opinion mining



Pair 1 (battery, good), path: NN→JJ, relation: Child
Pair 2 (flash, good), path: NN→VB→JJ, relation: GrandChild
Pair 3 (lcd, good), path: NN→NN→VB→JJ, relation: Indirect

Figure 5: Example of dependency paths and dependency relations

Let us consider the dependency tree of example "*Battery is very good even when using flash and LCD*" as shown in Figure 5. We can extract several product feature-opinion candidates such as "*battery, good*", "*flash-good*", and "*lcd-good*". Each such pair becomes a pair candidate. For effective relation extraction, we group product features by using product ontology that we will describe in next section. The maximum entropy model is used to predict opinion-relevant product feature. Firstly, for each pair, we compute several features automatically. We denote the features employed for learning as learning features, discriminative from the product features we discussed above. The features are opinions, grouped product features, dependency paths and dependency relations. We will simply choose the class with the highest conditional probability p according to Equation 1.

5 Product Ontology

In an abstract sense, an online customer review is a list of those product features or concepts that a customer likes or dislikes. Different customers will often refer to identical product features using inconsistent or incompatible terminology. Furthermore, customers might refer to a particular feature in different ways. For example, "*memory card*", "*compact flash*", "*compactflash*", "*CF card*", and "*memory stick*" are string for describing "*removable memory*". To solve this issue, we use sematic information encodsed in ontology. Figure 6 illustrates a part of our ontology.

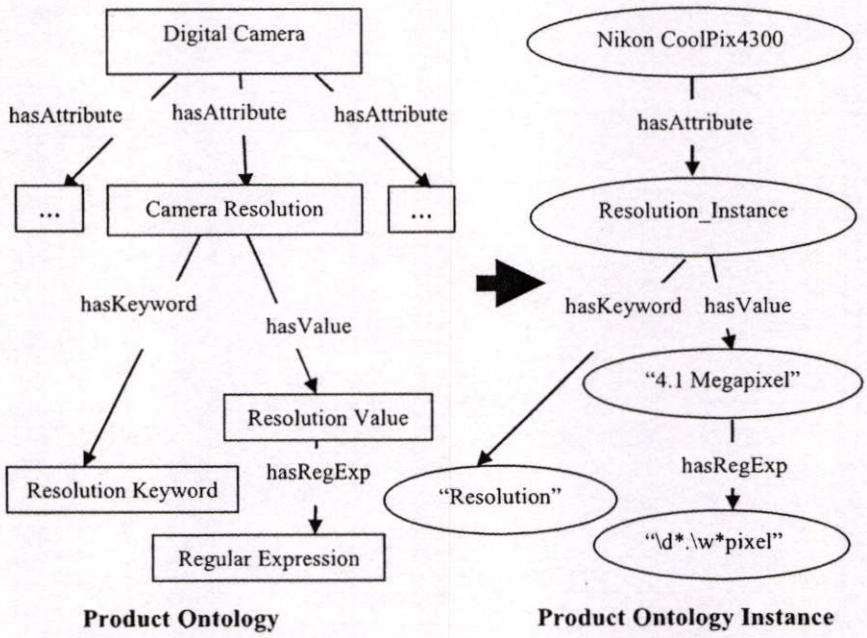


Figure 6: Fragments of product ontology and product ontology instance

Ontology plays a pivotal role here by providing a source of shared and precisely defined terms that can be used in such meta-data [Bhatt, 06]. Ontology can create an agreed-upon vocabulary for sharing knowledge, exchanging information, and eliminating ambiguity [Xue, 09]. In our work, we use ontology to normalize the language for distinguishing between different product features. In this paper, we design ontology by applying core ontology of Jannach et al [Jannach, 09]. Product ontology is expressed in a tree-hierarchy of concepts. We manually construct product ontology by integrating manufacturer product descriptions and terminologies in customer reviews. The root of the tree represents the product. Subsequent sub-trees represent attributes of the product.

According to the problem of describing a product feature in different ways, it is important to group terminologies with similar meaning together. Our work uses a simple method. The basic idea is to employ product ontology to group terminologies using simple regular expression patterns as showed on Figure 6. If a product feature candidate dose not matches any regular expression, using itself as a grouped product feature.

6 Experimental Settings

6.1 Data and Evaluation

The dataset used in our experiments included two sets on digital cameras from Hu's previous work [Hu, 04] and digital camera reviews from *Amazon.com*. The sentences in the dataset have manually generated tags indicating product features and opinions. We conducted 5-fold cross validation on that dataset. We employed the OpenNLP maximum entropy package as our classification tool.

To evaluate the method, we use precision, recall, and F-score to measure the effectiveness of our approach. When dealing with multiple datasets, we adopted the macro average to assess the overall performance across all datasets. The macro average is calculated by simply taking the average performance obtained for each dataset. Therefore, the definitions of precision, recall and F-score are as following.

$$\text{Precision} = \frac{PC}{PM} \qquad \text{Recall} = \frac{PC}{PT} \qquad \text{F - score} = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}}$$

PC = number of correctly mined product feature-opinion pairs;
PM = number of all mined product feature-opinion pairs;
PT = number of all correct product feature-opinion pairs.

6.2 Experimental Results and Discussion

In order to evaluate our method on the task of mining product features and opinions, we used a dataset of 1250 sentences described in Section 6.1. We randomly divided the dataset into five equal-sized folds. We used four folds as the training data and one fold as the testing data. We conducted the experiments to compare with Hu's approach (adjacent based method). Beside, the patterns used by Popesecu's approach (pattern based method) are adopted to compare with our method. The result is

compared with Hu's approach and Popescu's approach because they are the opinion summarization most relevant to our work and they have evaluated their performance on product review datasets. The product feature candidates are extracted by the method described in Section 4.2.1. Baseline is our approach without using product ontology.

We use precision, recall and F-score to evaluate performances. Five-fold cross-validation results of extracting product features and opinions are shown in Figure 7, Figure 8, and Figure 9. Table 3 shows the average results of different methods. The results show that our method outperforms others in the precision, the recall and the F-score. It shows that our method is superior to adjacent based method and some pattern based method with two main reasons. One reason, in adjacent based method, for each product feature, its nearest opinion word is used to construct the product feature-opinion pair. It produces many invalid pairs due to the complexity of sentences in product reviews. A second important reason, the pattern based method could not discover the relations between product features and opinions from the complex sentences. Beside, our approach performs a little better than non-ontology in recall and F-score because the right words dose not include in the ontology at design time.

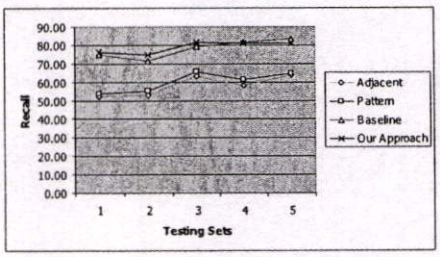


Figure 7: Recall of different methods

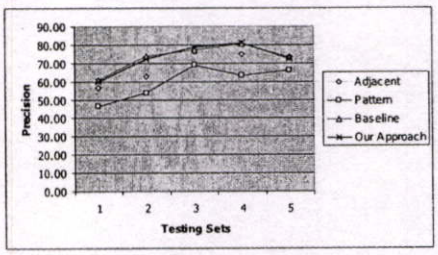


Figure 8: Precision of different methods

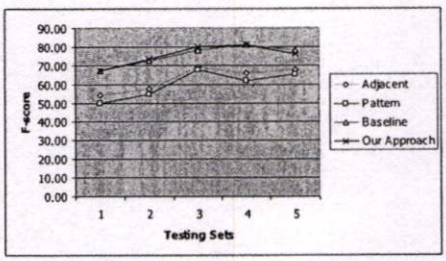


Figure 9: F-score of different methods

Methods	Precision (%)	Recall (%)	F-score (%)
Adjacent Based	68.65	57.93	62.69
Pattern Based	59.65	59.95	59.72
Baseline	73.12	77.98	75.34
Our Approach	72.65	78.77	75.45

Table 3: Average results for total performance with 5-fold cross validation on dataset

Table 4 shows the comparison of extracting product features and opinions of each method on simple and complex sentences. It is notable that the adjacent based and pattern based method can extract product features and opinions only from simple sentences. Interestingly, our method can extract product features and opinions form both type of sentences. Our method focus on opinion as adjective and verb exclude noun as in (3) because most of opinions as an adjective or a verb. In summary, we conclude that the approach is more flexible and effective than the adjacent based approach and opinion pattern based approach.

Structure	Example	Methods		
		Adjacent Based	Pattern Based	Our Approach
Simple	(1) There is a *great* camera.	Yes	Yes	Yes
	(2) The optical zoom works *great*.	Yes	Yes	Yes
	(3) Lens has *problems*.	No	Yes	No
	(4) I *like* this camera.	No	Yes	Yes
	(5) It is *great* having the LCD display.	No	No	Yes
Complex	(6) It has movie mode that works *good* for a digital camera.	No	No	Yes
	(7) I bought my canon g3 about a month ago and i have to say i am very *satisfied*.	No	No	Yes
	(8) The *nice* thing is that it uses the SD memory card.	No	No	Yes

Table 4: The comparison of extracting product features and opinions of difference methods

7 Conclusion and Future Work

In this paper, a dependency and semantic based approach is proposed for mining opinions from online customer reviews. We focused on extracting relations between product features and opinions. We have proposed a novel way to capture the actual relations of product features in sentences regardless the distance from them to opinions. Experimental results show the effectiveness of the proposed approaches.

As part of our future work, we would like to understand the reasons behind the unsatisfactory performance on the complex sentence. For example, a complex

sentence such as "*With the automatic settings, i really haven't taken a bad picture yet.*" confuses our method, because the sentence that describes positive expression and it's not relevant to extract "*bad picture*". The possible improvements could consist of using more natural language processing techniques. Finally, we would also investigate self-learning methods for classification that may provide a mechanism for further reducing the amount of labeled data required to produce highly accurate results.

Acknowledgements

The authors would like to thank the reviewers for their insightful comments and the valuable suggestions.

References

[Bhatt, 06] Bhatt, M., Flahive, A., Wouters, C., Rahayu, W., Taniar, D.: MOVE: A Distributed Framework for Materialized Ontology View Extraction, Algorithmica, Vol. 45, No. 3 (2006), 457-481.

[Carenini, 05] Carenini, G., Ng, R. T., Zwart, E.: Extracting Knowledge from Evaluative Text, In Proc. 3rd Int. Conf. Knowledge Capture, 2005, 11–18.

[Carenini, 06] Carenini, G., Ng, R., Pauls, A.: Multi-document Summarization of Evaluative Text, In Proc. 11th Conf. European Chapter of the ACL, 2006.

[Daly, 04] Daly, O., Taniar, D.: Exception Rules Mining Based on Negative Association Rules, In Proc. Int. Conf. ICCSA, Lecture Notes in Computer Science, Vol. 3046 (2004), 543-555.

[Darroch, 72] Darroch, J., Ratcliff, D.: Generalized Iterative Scaling for Log-linear Model, The Annals of Mathematical Statistics, Vol. 43, No.5 (1972), 1470–1480.

[Gamon, 05] Gamon, M., Aue, A., Corston-Oliver, S., Ringger, E.: Pulse: Mining Customer Opinions from Free Text, In Proc. 6th Int. Symp. Advances in intelligent data analysis, 2005, 121–132.

[Hu, 04] Hu, M., Liu, B.: Mining and Summarizing Customer Reviews, In Proc. 10th Int. Conf. Knowledge Discovery and Data Mining, Seattle, WA, 2004, 168–177.

[Jannach, 09] Jannach, D., Shchekotykhin, K., Friedrich, G.: Automated Ontology Instantiation from Tabular Web Sources-The ALLRIGHT System, Web Semantics: Sci.Serv.Agents World Wide Web, 2009.

[Jaynes, 57] Jaynes, E T.: Information Theory and Statistical Mechanics, Physical Reviews, Vol. 106 (1957), 620-630.

[Liu, 05] Liu, B., Hu, M., Cheng, J.: Opinion Observer: Analyzing and Comparing Opinions on the Web, In Proc. 14th Int. Conf. World Wide Web, Chiba, Japan, 2005, 342-351.

[Liu, 07] Liu, B.: Web Data Mining Exploring Hyperlinks, Contenrs, and Usage Data, Springer, New York, 2007.

[Melcük, 87] Melcük, I.: Dependency Syntax: Theory and Practice, State University of New York Press, 1987.

[Marcus, 93] Marcus, M. P., Santorini, B., Marcinkiewicz, M. A.: Building a Large Annotated Corpus of English: The Penn Treebank, Computational Linguistics, 1993.

[Pang, 02] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques, In Proc. Conf. Empirical Methods in Natural Language Processing, 2002, 79-86.

[Popescu, 05] Popescu, A. M., Etzioni, O.: Extracting Product Features and Opinions from Reviews, In Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, 2005, 339–346.

[Ratnaparkhi, 98] Ratnaparkhi, A.: Maximum Entropy Models for Natural Language Ambiguity Resolution, PhD thesis, University of Pennsylvania, 1998.

[Riloff, 03] Riloff, E., Janyce, W., Theresa, W.: Learning Subjective Nouns Using Extraction Pattern Bootstrapping, In Proc. 7th Conf. Natural Language Learning, 2003, 25-32.

[Stone, 66] Stone, P., Dunphy, D., Smith, M., Ogilvie, D., et al.: The General Inquirer: A Computer Approach to Content Analysis, Cambridge, MA: The MIT Press, 1966.

[Tan, 07] Tan, L., Taniar, D.: Adaptive Estimated Maximum-Entropy Distribution Model, Information Science, Vol. 177, No. 15 (2007), 3110-3128.

[Wilson, 05] Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-level Sentiment Analysis, In Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, 2005, 347-354.

[Xue, 09] Xue, Y., Wang, C., Ghenniwa, H. H., Shen, W.: A Tree Similarity Measuring Method and Its Application to Ontology Comparison, Journal of Universal Computer Science, Vol.15, No.9 (2009), 1766-1781.

[Yi, 05] Yi, J., Niblack, W.: Sentiment Mining in WebFountain, In Proc. 21st Int. Conf. on Data Engineering, 2005, 1073–1083.

[Zhuang, 06] Zhuang, L., Jing, F., Zhu, X. Y.: Movie Review Mining and Summarization, In Proc. 15th ACM Int. Conf. Information and knowledge management, 2006, 43-50.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวแกมกาญจน์ สมประเสริฐศรี
วัน เดือน ปีเกิด	21 พฤศจิกายน 2512
ที่อยู่	70 ถ.ราษฎรอุทิศ ต.ในเมือง อ.เมือง จังหวัดร้อยเอ็ด โทร. 0-4351-3338
ประวัติการศึกษา	2535 วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น 2541 วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ประสบการณ์การทำงาน	
พ.ศ. 2535-2539	โปรแกรมเมอร์ บริษัทมินิแบ (ประเทศไทย) จำกัด
พ.ศ. 2541- ปัจจุบัน	อาจารย์ประจำคณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม