บทที่ 2
แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

การศึกษาทฤษฎีการวัดความคล้ายคลึงของเอกสาร และเอกสารที่เกี่ยวข้อง ผู้ค้นคว้าพบว่ามีแนวคิดและทฤษฎีต่าง ๆ ที่เกี่ยวกับการประมวลผลภาษาธรรมชาติ มโนทัศน์เรื่องของคำในภาษาไทย การตัดคำภาษาไทย แบบจำลองเวกเตอร์สเปซ วิธีการวัดความคล้ายคลึง การวัดประสิทธิภาพของระบบ รวมถึงงานวิจัยที่เกี่ยวข้อง เพื่อนำไปสู่การพัฒนาระบบงานที่เป็นขั้นตอน ผู้ศึกษาจึงรวบรวมแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง ดังรายละเอียดตามลำดับดังนี้

2.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)

วิโรจน์ อรุณมานะกุล (2545) กล่าวไว้ว่า ภาษาคอมพิวเตอร์ คือ ศาสตร์ที่เกี่ยวข้องกับการทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ หรือเรียกอีกอย่างว่าการประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) ศาสตร์ทางด้านนี้เป็นแขนงวิชาหนึ่งที่เกี่ยวข้องกับศาสตร์หลาย ๆ ศาสตร์ ได้แก่ คอมพิวเตอร์ จิตวิทยา วิศวกรไฟฟ้า และสถิติ โดยที่ทางคอมพิวเตอร์จะเน้นที่การศึกษาในเรื่องของการประมวลผลภาษาธรรมชาติ (NLP) เรื่องของการแทนรูปความรู้ (Knowledge Representation) เรื่องของเทคนิคต่าง ๆ ของการแจงส่วนประโยค เป็นต้น ประโยชน์ที่ได้จากการทำให้คอมพิวเตอร์สามารถติดต่อสื่อสารกับมนุษย์ด้วยภาษามนุษย์เองนั้นชัดเจนในตัวเอง เพราะจะส่งผลให้การใช้งานคอมพิวเตอร์เป็นไปอย่างสะดวกมากยิ่งขึ้น คอมพิวเตอร์จะสามารถเข้ามาช่วยในงานด้านต่าง ๆ ที่เกี่ยวข้องกับภาษาได้มากขึ้น เช่น เป็นเครื่องแปลภาษามนุษย์จากภาษาหนึ่งไปเป็นอีกภาษาหนึ่ง (Machine Translation) สามารถนำคอมพิวเตอร์มาช่วยตรวจและวิเคราะห์เอกสารต่าง ๆ ที่มีว่าเกี่ยวข้องกับเรื่องใด ทำให้สามารถใช้คอมพิวเตอร์ช่วยในการค้นคืนข้อมูล (Information Retrieval) ตามความต้องการของผู้ใช้ได้ หรือสามารถให้คอมพิวเตอร์ช่วยสรุปสาระและประเด็นสำคัญ (Information Summarization) ที่ปรากฏในเอกสารนั้น ๆ เป็นต้น

ระบบการประมวลผลภาษาธรรมชาติอาศัยการพัฒนาโมดูลต่าง ๆ ตามแนวคิดทางภาษาศาสตร์ คือ หากอินพุตที่รับเข้ามาเป็นข้อความต่อเนื่อง ระบบประมวลผลก็จะต้องมีโมดูลการแยกแยะคำ โดยทำการวิเคราะห์ระดับวจีวิภาค (Morphological Analysis) เพื่อหารูปคำและวิภักติปัจจัยที่มี จากนั้นมีการหาหมวดคำของคำแต่ละคำ (Part-of-Speech Tagging) เพื่อเป็นข้อมูลพื้นฐานในการวิเคราะห์ทางวากยสัมพันธ์และทางความหมาย (Syntactic and Semantic Analysis) จากนั้นจึงผนวกประโยคที่วิเคราะห์ได้เข้าเป็นส่วนหนึ่งของข้อความทั้งหมดที่ประมวลผลแล้วเก็บไว้ในแบบจำลอง

ปริเฉก (Discourse Model) ซึ่งก็จะต้องมีกระบวนการตรวจสอบการอ้างอิงว่าคำนาม คำสรรพนามที่ใช้ในประโยคนั้นอ้างถึงสิ่งที่เคยอ้างมาก่อนหรือไม่ (Reference Resolution) เมื่อระบบต้องการโต้ตอบกลับเป็นภาษาธรรมชาติ ก็ทำกระบวนการที่ย้อนทิศทาง คือ จากความหมายที่ต้องการสื่อให้หารูปคำที่เหมาะสมสำหรับมโนทัศน์ต่าง ๆ ที่เกี่ยวข้อง จากนั้นจึงทำการหาโครงสร้างที่ถูกต้องทางไวยากรณ์สำหรับสร้างประโยคนั้น ๆ อาจมีการเปลี่ยนรูปคำให้เป็นรูปที่ปรากฏใช้จริงมีวิภักติปัจจัยติดมาด้วย

ระบบภาษาธรรมชาติเป็นระบบที่ได้รับการอ้างอิงหลักการทางด้านวิชาการปัญญาประดิษฐ์และฐานความรู้ โดยนำความรู้ทางภาษาศาสตร์มาจัดเก็บไว้ด้วยระบบคอมพิวเตอร์เป็นฐานความรู้ ระบบจะเรียกใช้ฐานความรู้ตีความหมาย ถ่ายทอดความรู้และโต้ตอบด้วยภาษาธรรมชาติ

ระบบภาษาธรรมชาติเป็นระบบซอฟต์แวร์ที่มีคุณสมบัติที่สำคัญ ดังนี้

1) ส่วนของระบบอินพุตเอาท์พุตที่ใช้ในการสั่งการหรือติดต่อกับคอมพิวเตอร์จะอยู่ในลักษณะที่เป็นภาษาธรรมชาติ

2) การประมวลผลภายในระบบจะใช้หลักพื้นฐานของฐานความรู้ที่เกี่ยวกับไวยากรณ์ความหมายและความเข้าใจภาษาธรรมชาติ

2.1.1 ความรู้ที่ระบบใช้ในการวิเคราะห์และสร้างอินพุตเอาท์พุตภาษาธรรมชาติ

ความรู้ที่ระบบใช้ในการวิเคราะห์แล้วสร้างอินพุตเอาท์พุตภาษาธรรมชาติจะประกอบด้วย

1) ความรู้ทางภาษาศาสตร์ (Conceptual Linguistics) เป็นความรู้ที่เกิดจากการใช้หลักภาษามาวิเคราะห์และสังเคราะห์

2) ความรู้ทางด้านมโนทัศน์ (Conceptual Knowledge) เป็นความรู้ที่ระบบจะต้องเข้าใจในความหมายและสามารถแยกความแตกต่างได้

3) การถ่ายทอดความรู้ (Inferential Knowledge) เป็นขบวนการที่จะใช้วินิจฉัยและถ่ายทอดความรู้มาใช้จากฐานความรู้

4) รูปแบบของผู้ใช้ เป็นการใช้ความรู้ทำความเข้าใจผู้ใช้ เพื่อให้ระบบมีความฉลาดเหมาะกับการใช้งานมากขึ้น

5) ความรู้ทางความหมายเกี่ยวเนื่อง (Discourse Knowledge) เป็นความรู้เกี่ยวกับรูปแบบของผู้ใช้ การโต้ตอบและสนทนา การถ่ายทอดความหมายและการตัดสินปัญหาโดยใช้ความเข้าใจ

6) ความรู้เกี่ยวกับการถ่ายทอดความรู้ เป็นความรู้ที่ใช้ในหลักการของปัญญาประดิษฐ์เกี่ยวกับเทคนิคการสร้างกฎเกณฑ์การถ่ายทอดและกระบวนการวินิจฉัย กฎเกณฑ์การโต้ตอบและการสนทนา รวมถึงการแทนด้วยหลักการทางคณิตศาสตร์ตรรกะและการวินิจฉัยหาคำตอบ

2.1.2 ความแตกต่างของภาษาคอมพิวเตอร์กับภาษามนุษย์

ภาษาคอมพิวเตอร์เป็นระบบที่มีการกำหนดขอบเขตไว้ในกรอบจำกัด มีการใช้คำจำกัด มีไวยากรณ์ที่ใช้จำกัดและการตีความหมายที่ชัดเจนจึงสามารถเรียกภาษาคอมพิวเตอร์อีกอย่างหนึ่งว่าภาษาที่มีรูปแบบ (Formal Language) ส่วนภาษาธรรมชาติเป็นภาษาที่มีขอบเขตกว้างมากจนยากที่จะหารู้แบบที่ตายตัวได้ กฎเกณฑ์ของภาษาธรรมชาติเป็นกฎเกณฑ์ที่เกิดขึ้นในการใช้ภาษาและเป็นที่ยอมรับของกลุ่มชนผู้ใช้นั้น ๆ

2.1.3 การแบ่งแยกหน่วยย่อยของภาษา (Natural Language Entity)

การแบ่งแยกหน่วยย่อยของภาษานั้นจะประกอบด้วย

1) ตัวอักษร (Alphabet) คือ สัญลักษณ์ที่ใช้แทนเสียง เป็นกลุ่มของสัญลักษณ์ที่จำกัดกลุ่มหนึ่ง

2) คำ (Word) คือ กลุ่มของตัวอักษรที่มาเรียงต่อกันเป็นคำ

3) ประโยค (Sentence) คือ กลุ่มของคำที่นำมาเรียงต่อกันเพื่อแทนความหมาย ประโยคจึงเป็นข้อความ

2.1.4 ขั้นตอนการเข้าใจภาษาธรรมชาติ

บุญเสริม กิจศิริกุล (2546) กล่าวไว้ว่า การเข้าใจในภาษาธรรมชาตินั้น (Natural Language Understanding) มีขั้นตอนดังต่อไปนี้

1) การวิเคราะห์ทางองค์ประกอบ (Morphological Analysis) เป็นการวิเคราะห์ในหน่วยคำว่าคำ ๆ หนึ่งสามารถแยกย่อยได้เป็นอะไรบ้างก็สามารถบอกได้ว่าคำนี้มีหน้าที่อะไร

2) การวิเคราะห์ทางวากยสัมพันธ์ (Syntactic Analysis) เป็นการวิเคราะห์ทางไวยากรณ์จุดมุ่งหมายก็เพื่อต้องการดูว่าประโยคที่รับเข้า ซึ่งประกอบด้วยคำหลาย ๆ คำเรียงต่อกันนั้นมีโครงสร้างเชิงวากยสัมพันธ์เป็นอย่างไร คำไหนเป็นประธาน กริยา กรรม หรือส่วนใดเป็นวลี เป็นต้น

3) การวิเคราะห์ทางความหมาย (Semantic Analysis) เป็นการวิเคราะห์ความหมายเมื่อได้โครงสร้างโดยการวิเคราะห์ทางวากยสัมพันธ์มาแล้วก็จะกำหนดค่าของแต่ละคำว่าหมายถึงสิ่งใด

4) บูรณาการทางวจนิพนธ์ (Discourse Integration) เป็นการพิจารณาความหมายของประโยค โดยดูจากประโยคข้างเคียง เนื่องจากคำบางคำในประโยคหนึ่งจะเข้าใจความหมายได้ถูกต้องดูประโยคก่อนหน้าหรือประโยคตามด้วย

5) การวิเคราะห์ทางปฏิบัติ (Pragmatic Analysis) เป็นการแปลความหมายของประโยคใหม่อีกครั้งว่าที่จริงแล้วว่าที่จริงแล้วผู้พูดตั้งใจจะหมายความว่าอย่างไรหรือต้องการสื่อความหมายอะไร

2.1.5 ภาษาไทย (Thai Language)

นววรรณ พันธุเมธา (2527) กล่าวไว้ว่า คำ หมายถึง กลุ่มของหน่วยเสียงหรือกลุ่มของตัวอักษรที่มีความหมายในภาษา คำในภาษาเขียนไทยจะประกอบขึ้นจากพยางค์ตั้งแต่ 1 พยางค์ขึ้นไป และแต่ละพยางค์นั้นก็ประกอบขึ้นจากตัวอักษรตั้งแต่ 2 ตัวขึ้นไป

ตัวอักษรไทย (Thai Alphabet) คือ เครื่องหมายที่ใช้แทนเสียงในภาษาไทย ซึ่งสามารถจำแนกออกมาเป็นรูปได้ดังนี้

(ก) รูปพยัญชนะมี 44 รูป ใช้แทนเสียงพยัญชนะ 21 เสียง รูปพยัญชนะสามารถแบ่งแยกย่อยได้เป็น 3 หมู่ (ไตรยางค์) คือ

1) อักษรสูง มี 11 ตัว ได้แก่ ข ฃ ฉ ฐ ถ ผ ฝ ศ ษ ส และ ห

2) อักษรกลาง มี 9 ตัว ได้แก่ ก จ ฎ ฏ ด ต บ ป และ อ

3) อักษรต่ำ มี 24 ตัว ได้แก่ ค ฅ ฆ ง ช ซ ฌ ญ ฑ ฒ ณ ท ธ น พ ฟ ภ ม ย ร ล ว ฬ และ ฮ

พยัญชนะทั้ง 44 ตัว ถูกนำไปใช้ในการสร้างพยางค์ โดยมีหน้าที่ในพยางค์ได้ ดังนี้

1) เป็นพยัญชนะต้น เช่น *ก*าร *ข*าย และ *บ*าง เป็นต้น

2) เป็นพยัญชนะท้ายพยางค์ (ตัวสะกด) เช่น กา*ร* ขา*ย* และ บา*ง* เป็นต้น

3) เป็นอักษรควบ เช่น *ค*วาม หรือ *แ*ทรก เป็นต้น

4) เป็นสระ (อ ว ย และ ร) เช่น ข*อ*น อ*ว*น เสี*ย* และ *สรรค์* เป็นต้น

5) เป็นตัวการันต์ เช่น จัน*ทร์* (ทร เป็นตัวการันต์) ศิล*ป์* (ป เป็นตัวการันต์) เป็นต้น

(ข) สระมี 32 เสียง ได้แก่

1) อะ	2) อา	3) อิ	4) อี	5) อึ	6) อือ	7) อุ
8) อู	9) เอะ	10) เอ	11) แอะ	12) แอ	13) โอะ	14) โอ
15) เอาะ	16) ออ	17) เออะ	18) เออ	19) อัวะ	20) อัว	21) เอียะ
22) เอีย	23) เอือะ	24) เอือ	25) อำ	26) ใอ	27) ไอ	28) เอา
29) ฤ	30) ฤา	31) ฦ	32) ฦา			

โดยสระทั้ง 32 เสียงข้างต้น ประกอบขึ้นจากรูปสระทั้ง 21 รูป

(ค) วรรณยุกต์มี 4 รูป ใช้แทนเสียงวรรณยุกต์ 5 เสียง วรรณยุกต์ถือเป็นอักษรที่ใช้แทนระดับเสียงสูงต่ำในภาษาไทย และเป็นสิ่งที่ช่วยเปลี่ยนความหมายของคำได้

2.1.6 พยางค์ในภาษาไทย

เรืองเดช ปันเขื่อนขัติย์ (2541) กล่าวไว้ว่า พยางค์ คือ ส่วนหนึ่งของคำที่ประกอบด้วยสระตัวเดียวจะมีความหมายหรือไม่มีก็ได้ พยางค์หนึ่งมีส่วนประสมของอักษรต่าง ๆ กัน คือ

1) พยัญชนะต้น + สระ + วรรณยุกต์ เช่น ตา ดี ไป และนา เป็นต้น พยางค์ชนิดนี้เรียกว่าประสม 3 ส่วน

2) พยัญชนะต้น + สระ + วรรณยุกต์ + ตัวสะกด เช่น คน กิน และข้าว เป็นต้น หรือพยัญชนะต้น + สระ + วรรณยุกต์ + ตัวการันต์ เช่น โลห์ เล่ห์ เป็นต้น พยางค์ทั้ง 2 ชนิดนี้ เรียกว่าประสม 4 ส่วน

3) พยัญชนะต้น + สระ + วรรณยุกต์ + ตัวสะกด + ตัวการันต์ เช่น แพทย์ รักษ์ สิทธิ์ ฤทธิ์ และโรจน์ เป็นต้น พยางค์ชนิดนี้เรียกว่า ประสม 5 ส่วน

ดังนั้นจะเห็นว่า พยางค์แต่ละพยางค์อย่างน้อยต้องมีส่วนประสม 3 ส่วน คือ พยัญชนะ สระ และวรรณยุกต์ อย่างมากจะต้องมีส่วนประสม 5 ส่วน คือ พยัญชนะ สระ วรรณยุกต์ ตัวสะกด ตัวการันต์ จะน้อยกว่า 3 ส่วน หรือมากกว่า 5 ส่วนไม่ได้ และคำหนึ่ง ๆ จะมีพยางค์เดียวหรือหลายพยางค์ก็ได้

2.1.7 มโนทัศน์เรื่องคำและการรู้จักคำในภาษาไทย

Leonard Bloomfield (1933) กล่าวไว้ว่า คำ หมายถึง เสียงพูดหรือสายอักขระที่มีความหมายในภาษา โดยคำจะเป็นหน่วยที่เล็กที่สุดที่สามารถปรากฏตามลำพังได้ (Minimum Free Form)

Robins (1964) กล่าวไว้ว่า สิ่งที่เป็นคำจะมีความคงตัว (Stability) ไม่สามารถจะแยกย่อยให้เล็กลงไปได้อีก และจะจัดลำดับส่วนที่อยู่ในคำเสียใหม่ก็ไม่ได้ ส่วนตำราไวยากรณ์ทั้งหลาย

ปราณี กุลละวณิชย์ และคณะ (2535) กล่าวไว้ว่า มโนทัศน์เรื่องคำเป็นพื้นฐานสำหรับปัญหาการตัดคำในภาษาไทย ในส่วนนี้จะแสดงให้เห็นถึงลักษณะของภาษาไทยที่เป็นสาเหตุของปัญหาการตัดคำ ซึ่งเกิดจากการที่ภาษาไทยไม่มีการเขียนแยกคำที่ชัดเจน และนอกจากนี้ปัญหาการตัดคำส่วนหนึ่งยังเกิดจากการที่ยังไม่มีเกณฑ์ที่ชัดเจนในการตัดสินขอบเขตของคำ หากพิจารณาระบบการเขียนของภาษาไทยจะพบว่ามีเอกลักษณ์เฉพาะตัวคือ คำจะเขียนเรียงติดต่อกันไปเป็นประโยคหรือข้อความได้โดยไม่มีการเว้นช่องว่างระหว่างคำเหมือนในภาษาอื่น ๆ หลายภาษา เช่น ภาษาอังกฤษ และภาษาฝรั่งเศส เป็นต้น ดังนั้นในข้อเขียนภาษาไทยจึงไม่ปรากฏตัวแบ่งขอบเขตของคำ (Word Boundary Delimiter) ที่ชัดเจนแน่นอนว่า คำหนึ่ง ๆ เริ่มต้นที่ใดและสิ้นสุดที่ใด ดังนั้นจึงทำให้เกิดความกำกวมในการตัดคำในภาษาไทย เช่น สายอักขระ (String) ที่ว่า “ตากลม” สามารถตัดคำเป็น “ตา-กลม” (Eye) (Round) หรือ “ตาก-ลม” (to Expose) (Wind) หรือในกรณีของสายอักขระ “แม่น้ำ” อาจจะพิจารณาให้เป็นลำดับของคำ (Word Sequence) ที่ประกอบด้วยคำ 2 คำ

คือ “แม่” (Mother) และ “น้ำ” (Water) หรืออาจพิจารณาให้เป็นคำประสมหนึ่งคำว่า “แม่น้ำ” (River) ก็ได้

เรืองเดช ปันเขื่อนขัติย์ (2541) กล่าวไว้ว่า คำ หมายถึง กลุ่มของหน่วยเสียงหรือกลุ่มของตัวอักษรที่มีความหมายในภาษา คำจำกัดความเหล่านี้มีความแตกต่างกันในด้านมุมมองในการตัดสินคำ

อมรา ประสิทธิ์รัฐสินธุ์ และคณะ (2544) กล่าวไว้ว่า มโนทัศน์เรื่องคำนี้ ตำราไวยากรณ์ต่าง ๆ มักไม่ได้อธิบายหรือให้คำจำกัดความไว้ชัดเจนมากนัก โดยมักถือเอาว่าเจ้าของภาษารู้ดีอยู่แล้วว่าคำคืออะไร ถึงแม้มโนทัศน์เรื่องของคำยังไม่ชัดเจนนักก็ตาม ก็สามารถถือได้ว่าคำเป็นหน่วยที่มีความสำคัญในความคิดของผู้พูดภาษา ดังจะสามารถสังเกตจากการที่เด็กเริ่มเรียนรู้ภาษาก็เรียนเป็นคำ พจนานุกรมในภาษามักบรรจุคำเอาไว้ หรือการยืมระหว่างภาษาก็เป็นการยืมคำเสียเป็นส่วนใหญ่

ลักษณะอีกประการหนึ่งของภาษาไทยที่เป็นปัญหาต่อการประมวลผลภาษา คือ คำในภาษาไทยสามารถจัดได้เป็นหลายประเภทตามกลวิธีในการประกอบคำเพื่อให้ได้คำใหม่ ๆ มาใช้เพิ่มเติมในภาษาให้เพียงพอต่อความจำเป็นในการสื่อสาร ตำราไวยากรณ์ภาษาไทยต่าง ๆ ได้อธิบายชนิดของคำตามลักษณะการประกอบคำเป็นประเภทต่าง ๆ ได้แก่

1) คำมูล (Simple Word) คือ คำดั้งเดิมในภาษาไทยที่ยังมิได้ประสมกับคำอื่น จะเป็นคำที่มาจากภาษาใดก็ได้ ซึ่งอาจประกอบขึ้นจากคำพยางค์เดียวหรือหลายพยางค์ ลักษณะของคำมูลมีดังนี้

- คำมูลพยางค์เดียวเป็นคำไทยแท้ เช่น พ่อ แม่ พี่ น้อง กิน และดื่ม เป็นต้น
- คำมูลพยางค์เดียวที่เป็นคำยืมจากภาษาอื่น เช่น เดิน บวช นาค และกอล์ฟ เป็นต้น
- คำมูลหลายพยางค์ที่เป็นคำไทยแท้ เช่น กระจอก คะนอง และจำเพาะ เป็นต้น
- คำมูลหลายพยางค์ที่เป็นคำยืมภาษาอื่น เช่น สติ นาฬิกา และกาแฟ เป็นต้น

2) คำประสม (Compound Word) เกิดจากคำมูลตั้งแต่สองคำมาประสมกัน ทำให้เกิดคำใหม่ที่มีความหมายใหม่ แต่ยังคงเค้าความหมายเดิม เช่น ม้านั่ง มีดพับ ดวงตรา กันชน ห้องรับแขก ไม้จิ้มฟัน และผ้าปูโต๊ะ เป็นต้น คำประสมในภาษาไทยมีประเด็นที่น่าพิจารณาดังนี้

(ก) ลักษณะทางความหมายของคำที่นำมาประสมกันมี 2 ลักษณะ คือ

- คำประสมที่มีความหมายตรง คำประสมประเภทนี้ความหมายสำคัญอยู่ที่คำหลัก ส่วนขยายมีความสำคัญรองลงไป เช่น *ผล*ผลิต *สระ*น้ำ *แม่*น้ำ *ทาง*เดิน *เรือ*หาง *พ่อ*ตา *เมือง*นอก *พัด*ลม และ*ลาย*มือ เป็นต้น
- คำประสมที่มีความหมายโดยนัย หรือความหมายเปรียบเทียบ เช่น มือเท้า ไก่อ่อน หัวสูง เส้นสาย ตาขาว วิ่งราว หัวหมุน และหัวแข็ง เป็นต้น

(ข) วิธีการสร้างคำประสม กระทำโดยนำคำมูลที่อาจเป็นภาษาเดียวกันหรือต่างภาษากันก็ได้ ได้แก่

- นำคำที่ไม่สามารถปรากฏตามอิสระได้ เช่น ช่าง ชาว เครื่อง ความ การ และของ ที่เป็นคำหลักประสบกับคำที่สามารถปรากฏตามอิสระได้ เช่น ช่างทอง ชาวนา นักเขียน การกิน ความดี เครื่องเขียน และหมอสัตว์ เป็นต้น
- คำไทยแท้ประสมกับคำไทยแท้ เช่น พ่อตา โรงเรียน และคนงาน เป็นต้น
- คำไทยแท้ประสมคำบาลีสันสกฤต เช่น หลักฐาน และราชวัง เป็นต้น

(ค) ลักษณะคำประสม

- คำประสมที่ใช้เป็นคำนาม เช่น คนกรุง หน้ายักษ์ และน้ำพริก เป็นต้น
- คำประสมที่ใช้เป็นคำกริยา เช่น กินที่ และถือหาง เป็นต้น
- คำประสมที่ใช้เป็นคำวิเศษณ์ เช่น (เครื่อง) คิดเลข เป็นต้น

3) คำซ้ำ (Reduplication) เกิดจากการซ้ำเสียงคำมูล ซึ่งมักจะเป็นคำพยางค์เดียวมีลักษณะดังนี้ ออกเสียงเหมือนกัน เขียนเหมือนกัน และทำหน้าที่อย่างเดียวกันในประโยคโดยใช้เครื่องหมายไม้ยมก (ๆ) แทนคำที่ซ้ำกับคำแรก เช่น นั่ง ๆ นอน ๆ เบา ๆ และเร็ว ๆ เป็นต้น นอกจากนี้คำซ้ำยังเป็นการเปลี่ยนระดับเสียงเพื่อเน้นความหมายของคำให้เด่นชัดยิ่งขึ้น เช่น ค้าวขาว ดี๊ดี และแก๊แก่ เป็นต้น

4) คำซ้อน (Synonymous Compound Word) คือ คำประสมชนิดหนึ่งที่เกิดจากการนำคำที่มีความหมายหรือเสียงเหมือนกันหรือคล้ายคลึงกันมาซ้อนกันแล้วเกิดคำที่มีความหมายใหม่ โดยที่บางคำมีความหมายใกล้เคียงกับคำมูลเดิม เช่น ถนนหนทาง ครูบาอาจารย์ และบ้านเรือน เป็นต้น

5) คำสมาส เกิดจากการนำคำที่มาจากภาษาบาลี-สันสกฤต ตั้งแต่ 2 คำขึ้นไปมาประสมกันให้เกิดเป็นคำใหม่ คำสมาสคล้ายกับคำประสม ต่างกันตรงที่คำประสมนั้นคำหลักมักอยู่ข้างหน้า คำขยายจะอยู่ข้างหลัง ส่วนคำสมาสนั้นคำหลักมักจะอยู่ข้างหลัง คำขยายมักจะวางอยู่ข้างหน้า เวลาแปลความคำสมาสจะแปลจากคำหลังมาหาคำหน้า เช่น

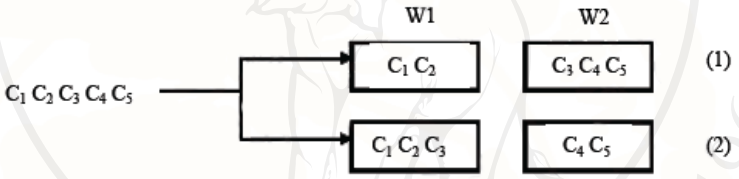
ประวัติ + ศาสตร์	เป็น	ประวัติศาสตร์
พละ + ศึกษา	เป็น	พลศึกษา
วัฒนะ + ธรรม	เป็น	วัฒนธรรม

6) คำสนธิ เกิดจากการนำคำที่มาจาภาษาบาลี-สันสกฤตมารวมกัน แต่การมารวมกันนั้น มิได้นำคำมาเรียงติดต่อกันอย่างคำสมาส แต่จะมีการเชื่อมเสียงให้กลมกลืนตามหลักภาษาบาลี-สันสกฤต เช่น

วิทยา + อาลัย	เป็น	วิทยาลัย
ธนู + อาคม	เป็น	ธันวาคม
มนัส + ภาพ	เป็น	มโนภาพ
สุข + อุทัย	เป็น	สุโขทัย

จากมโนทัศน์ของคำตามที่กล่าวมาข้างต้น ทำให้มองเห็นปัญหาที่สำคัญอย่างหนึ่งของภาษาไทยในการประมวลผล คือ

1) ความกำกวมในเรื่องขอบเขตของคำ ประโยคของคำภาษาไทยถูกจัดรูปแบบด้วยลำดับของคำที่มาก ซึ่งปราศจากการแยกแยะคำที่ชัดเจน จึงเป็นสาเหตุของความกำกวมในเรื่องขอบเขตของคำ ตามรูป 2.1



รูป 2.1 แสดงกลุ่มตัวอักษรที่เป็นไปได้สำหรับการตัดคำ

2) หน้าที่คำที่กำกวม คำในภาษาไทยมีได้มากกว่าหนึ่งหน้าที่ เป็นสาเหตุทำให้หน้าที่ตัดคำที่วิเคราะห์ได้มีขนาดใหญ่ เช่น ในประโยคหนึ่งมีทั้งหมด 12 คำ สามารถสร้างเป็นรูปแบบไวยากรณ์ได้ถึง 3,027 รูปแบบ ดังนั้นทั้งความกำกวมในเรื่องขอบเขตของคำและหน้าที่คำจะสร้างความซับซ้อนให้กับการวิเคราะห์ไวยากรณ์ภาษาไทย

3) การปรากฏการสะกดผิด การสะกดผิดจำแนกได้ 2 ชนิด คือ การสะกดผิดที่เห็นได้ชัดเจนซึ่งสามารถตรวจสอบโดยการใช้พจนานุกรม และการสะกดผิดที่ทำให้กลายเป็นคำอื่นจะทำให้ความหมายของคำเดิมผิดไปจากเดิม ในกรณีไม่สามารถตรวจสอบโดยการใช้พจนานุกรมได้

2.1.8 การตัดคำ (Word Segmentation)

ไพศาล เจริญพรสวัสดิ์ (2541) กล่าวไว้ว่า การตัดคำ คือ การแบ่งข้อความหรือสายอักขระของตัวอักษรที่ต่อเนื่องกันออกเป็นหน่วยคำ (Morpheme) หรือลักษณะของการรู้จำ (Recognition) คำหนึ่ง ๆ เพื่อหาขอบเขตของแต่ละหน่วยคำ การตัดคำในข้อความที่ยาวต่อเนื่องกันนั้นเริ่มจะมีความหมายมากขึ้นเป็นลำดับเมื่อมีการนำเอาคอมพิวเตอร์เข้ามาช่วยในการประมวลผลข้อมูลทางภาษามากยิ่งขึ้น ความยากง่ายหรือวิธีการที่จะนำมาใช้ในการตัดคำนั้นต้องขึ้นอยู่กับลักษณะเฉพาะของภาษานั้น ๆ เป็นอย่างมาก

โดยปกติแล้ว ความมุ่งหวังที่จะให้คอมพิวเตอร์สามารถประมวลผลภาษาไทยได้อย่างมีประสิทธิภาพนั้นมีปัญหาเบื้องต้น คือ ลักษณะการเขียนของภาษาไทยจะเขียนติดต่อกันโดยไม่มีการใช้เครื่องหมายวรรคตอนใด ๆ แสดงการแบ่งคำดังเช่นภาษาอังกฤษ ยกเว้นแต่จะมีการเว้นวรรคเป็นระยะ ๆ เพื่อให้ผู้อ่านได้หยุดพักและทำความเข้าใจในความหมายเป็นตอน ๆ ไปเท่านั้น แม้ว่าการเว้นวรรคในการเขียนบทความไม่ได้มีกฎเกณฑ์ที่ชัดเจน แต่ถ้าใช้การเว้นวรรคด้วยความระมัดระวังแล้วก็จะสามารถช่วยลดความคลุมเครือของคำหรือประโยคได้ การตัดคำเพื่อให้ได้ขอบเขตของคำที่ถูกต้องเป็นอุปสรรคอย่างหนึ่งที่ต้องการการศึกษาวิจัยและพัฒนาเพื่อทำให้คอมพิวเตอร์สามารถคำนวณเพื่อแบ่งสายอักขระไทยออกเป็นคำ ๆ ซึ่งจะส่งผลให้การทำงานของคอมพิวเตอร์ในการค้นหาคำใด ๆ สามารถทำได้อย่างถูกต้องและแม่นยำ

ไม่ว่าจะด้วยวิธีการใดก็ตาม ถ้าหากสามารถรู้ขอบเขตของคำแต่ละคำได้แล้ว การจัดการกับข้อความนั้น ๆ ก็จะเป็นไปได้อย่างสะดวกและถูกต้อง ในระบบคอมพิวเตอร์จึงจำเป็นอย่างยิ่งที่จะต้องคำนวณหาขอบเขตของแต่ละคำให้ได้ เพื่อที่จะสามารถลดภาระของผู้ใช้หรือเพื่อที่จะให้กระบวนการที่อยู่ในระดับที่ลึกกว่าสามารถทำงานต่อไปได้ เช่น ฟังก์ชั่นการจัดขอบขวา (Word Wrapping) ในโปรแกรมไมโครซอฟท์ออฟฟิศเวิร์ด การตรวจคำผิด การค้นหาคำในข้อความ หรือเป็นตัวช่วยในการกำหนดคำเพื่อทำการวิเคราะห์ต่อไปในระบบของเครื่องแปลภาษา เป็นต้น

ที่ผ่านมาจนถึงปัจจุบัน งานด้านการตัดคำในภาษาไทยได้รับการพัฒนาจากหน่วยงานวิจัยต่างๆ ทั้งของภาครัฐและภาคเอกชน โดยมีวิธีพัฒนาแนวคิดวิธีการต่าง ๆ เพื่อใช้ในการตัดคำมาเป็นลำดับ แต่ละวิธีการต่าง ๆ ก็ให้ผลในด้านความถูกต้อง ความรวดเร็วของการทำงาน และปริมาณการใช้ทรัพยากรต่าง ๆ แตกต่างกัน วิธีการตัดคำภาษาไทยสามารถแบ่งได้เป็น 3 หลักการใหญ่ ๆ คือ หลักการตัดคำโดยใช้กฎ หลักการตัดคำโดยใช้พจนานุกรมและหลักการตัดคำโดยใช้คลังข้อมูล ดังนี้

(ก) หลักการตัดคำโดยใช้กฎ (Rule Based Approach) วิรัช ศรเลิศล้ำวานิช (2536) กล่าวไว้ว่า การตัดคำโดยใช้กฎเป็นความพยายามในขั้นเริ่มต้นของการพัฒนาระบบตัดคำภาษาไทย โดยที่ได้เสนอวิธีการตรวจสอบกฎเกณฑ์ทางอักขรวิธีที่กำหนดลักษณะของการประสมอักษร การเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการบ่งชี้ขอบเขตของคำ อาทิเช่น อักษรกลุ่มที่เป็นตัวการันต์ที่มีทัณฑฆาตบังคับข้างบน เช่น การเว้นวรรคจะมีความเป็นไปได้ของการสิ้นสุดของคำหรือประโยค และการขึ้นย่อหน้าจะเป็นตัวบ่งชี้การสิ้นสุดข้อความ เป็นต้น

พิสิทธิ์ พรมจันทร์ (2540) กล่าวว่า วิธีการนี้มีข้อจำกัดมาก คือ ผลของการตัดคำอาจได้เป็นกลุ่มคำที่สามารถตัดแยกย่อยออกไปได้อีก ดังนั้นความถูกต้องของคำที่ได้จึงต่ำ แต่มีข้อดี คือ ความเร็วในการทำงานสูง ใช้ทรัพยากรน้อย ซึ่งวิธีการนี้ใช้กับงานบางประเภทที่ไม่ต้องจำเป็นต้องตัดแยกคำให้ย่อยที่สุด เช่น งานจัดรูปแบบเอกสาร เป็นต้น

(ข) หลักการตัดคำโดยใช้พจนานุกรม (Dictionary Approach) การตัดคำโดยใช้พจนานุกรมเป็นแนวคิดที่ได้รับการพัฒนาในยุคต่อมา โดยเก็บคำภาษาไทยไว้ในพจนานุกรม แล้วนำข้อความที่ป้อนเข้า (Input) ไปค้นหาและเทียบสายอักขระกับคำในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวควรตัดคำในบริเวณใด และประกอบด้วยคำใดบ้าง การนำพจนานุกรมมาใช้ในการตัดคำภาษาไทยจะมีการทำงานอยู่ 2 ขั้นตอน คือ ขั้นตอนแรกจะทำการตัดคำโดยเทียบระหว่างข้อความกับคำในพจนานุกรม และขั้นตอนที่สองจะทำการเปรียบเทียบระหว่างคำที่ได้ในขั้นตอนแรกกับคำในพจนานุกรมอีกครั้งหนึ่ง เพื่อหาคำที่สามารถตัดในรูปแบบอื่นได้อีกหรือไม่ ถ้าได้จะใช้คำใหม่ที่ได้เป็นผลลัพธ์ของการตัดคำ

พิสิทธิ์ พรมจันทร์ (2540) กล่าวไว้ว่า วิธีการนี้จะสิ้นเปลืองทรัพยากรหน่วยความจำหลักค่อนข้างมาก เนื่องจากทั้งพจนานุกรมและอะเรย์ของคำที่ตัดได้จะเก็บไว้ในหน่วยความจำหลักทั้งหมด ประสิทธิภาพเชิงความถูกต้องของคำขึ้นอยู่กับปริมาณคำในพจนานุกรม ส่วนความเร็วขึ้นอยู่กับขั้นตอนวิธี (Algorithm) ที่ใช้ในการค้นหาคำจากพจนานุกรม และผลที่ได้จากการตัดคำวิธีนี้อาจมีได้มากกว่าหนึ่งทางเลือก ดังนั้นจึงต้องมีการเลือกทางเลือกของคำที่ถูกต้องต่อไปอีก

วิรัช ศรเลิศล้ำวานิช (2536) กล่าวไว้ว่า การตัดคำโดยใช้พจนานุกรมนั้น เป็นไปไม่ได้ที่จะเก็บคำทุกคำในภาษาไทยลงในพจนานุกรมได้ โดยเฉพาะคำวิสามายนาม เช่น ชื่อคน ชื่อสถานที่ ตัวเลข รวมถึงคำแสลง หรือคำที่เกิดขึ้นใหม่

หลักการตัดคำโดยใช้พจนานุกรมนี้สามารถตัดคำได้ถูกต้องมากกว่าการใช้กฎ จึงได้รับความนิยมและมีผู้พัฒนาวิธีการตัดคำภาษาไทยอื่น ๆ โดยใช้พจนานุกรมช่วย เช่น วิธีการเทียบคำที่ยาวที่สุด และวิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่พบในพจนานุกรมน้อยที่สุด เป็นต้น

1) วิธีการเทียบคำที่ยาวที่สุด (Longest Matching) วิธีการเทียบคำที่ยาวที่สุดเป็นวิธีการตัดคำทางวิทยาการศึกษาสำนึก (Heuristic) วิธีหนึ่งซึ่งต้องใช้พจนานุกรมช่วยรู้จำคำภาษาไทย โดยวิธีนี้จะทำการตรวจสอบหรือสแกนข้อความที่ป้อนเข้าจากซ้ายไปขวา นำไปเทียบพจนานุกรมดูว่า สายอักขระดังกล่าวเป็นหนึ่งคำหรือไม่ หากไม่พบว่าสายอักขระดังกล่าวสามารถเทียบเป็นคำได้ในพจนานุกรม ก็จะทำการลดความยาวของสายอักขระลงทีละตัว จนกว่าสายอักขระที่ตรวจสอบจะสามารถเทียบเป็นคำในพจนานุกรมได้ก็จะทำเครื่องหมายเพื่อเป็นจุดย้อนกลับ จากนั้นจะเริ่มทำงานจากจุดย้อนกลับนั้นเพื่อตรวจสอบสายอักขระที่เหลือว่าจะสามารถตัดสายอักขระใดต่อไปให้เป็นคำได้ หากตัวเลือกในตอนแรกนี้สามารถทำให้ขั้นตอนวิธีค้นหาคำที่เหลือได้ ตัวเลือกนี้ก็จะเป็นคำแรกของข้อความได้จริง ไม่เช่นนั้นขั้นตอนวิธีก็จะกลับไปยังจุดย้อนกลับที่ทำเครื่องหมายไว้เพื่อแก้ไขคำแรกใหม่ จากนั้นจะเริ่มทำงานต่อไปโดยเริ่มจากจุดย้อนกลับ หากยังไม่สามารถเทียบสายอักขระกับคำในพจนานุกรมได้ก็จะทำการลดตัวอักษรลงทีละตัวจนกว่าจะเทียบคำในพจนานุกรมได้ และทำงานในรูปแบบนี้ไปจนจบข้อความ ตัวอย่างเช่น ถ้าป้อนข้อความ “ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคัญ” เข้าไป ข้อความนี้เมื่อไม่สามารถเทียบคำกับพจนานุกรมได้ก็จะลดเหลือ →“ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคั” →“ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำ” จนได้สายอักขระ “ความก้าวหน้า” ซึ่งสามารถเทียบคำในพจนานุกรมได้จึงตัดเป็นคำแรกของข้อความ และทำเครื่องหมายไว้เป็นจุดย้อนกลับ ผลการตัดคำทั้งหมดจะเป็นดังตาราง 2.1

ตาราง 2.1 ผลการตัดคำทั้งหมดด้วยวิธีเทียบคำที่ยาวที่สุด

ส่วนของคำที่ยาวที่สุดที่ตัดได้	ส่วนหลังที่เหลือที่จุดย้อนกลับ
ความก้าวหน้า	ทางด้านวิทยาศาสตร์มีบทบาทสำคัญ
ทาง	ด้านวิทยาศาสตร์มีบทบาทสำคัญ
ด้าน	วิทยาศาสตร์มีบทบาทสำคัญ
วิทยาศาสตร์	มีบทบาทสำคัญ
มี	บทบาทสำคัญ
บทบาท	สำคัญ
สำคัญ	-

ข้อด้อยของวิธีการนี้เกิดจากลักษณะความพยายามที่จะตัดคำให้ได้คำยาวที่สุดของขั้นตอนวิธี คือ การเลือกคำที่ยาวเกินไปตั้งแต่ต้นมีผลทำให้การตัดคำที่ตามมาผิดเพี้ยนไป ตัวอย่างเช่น ข้อความป้อนเข้า “ไปหามเหสี” (go to see the queen) จะตัดคำได้เป็น ไป(go) หาม(carry) เห (deviate) สี(color) เสมอ เนื่องจากขั้นตอนวิธีจะพบคำว่า “หาม” ก่อนคำว่า “หา” เสมอ ทำให้การตัดคำสำหรับคำต่อ ๆ ไปผิดด้วย

2) วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด หรือการตัดคำโดยเลือกแบบเหมือนมากที่สุด (Maximal Matching) วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรม (Unknown Word) น้อยที่สุดก็เป็นวิธีการตัดคำทางวิทยาการศึกษาสำนึกที่ใช้พจนานุกรมช่วยรู้จำคำภาษาไทย วิธีนี้พัฒนาโดยวิรัช ศรเลิศล้ำวาณิช (2536) เพื่อแก้ปัญหาที่ปรากฏในวิธีการเทียบคำที่ยาวที่สุด วิธีการนี้จะพัฒนาบนขั้นตอนวิธีการเทียบคำที่ยาวสุด เริ่มจากการหาทางเลือกของรูปแบบการตัดคำทั้งหมดที่เป็นไปได้ โดยทำการย้อนกลับ (Backtracking) ทีละคำหลังจากได้คำตอบจากวิธีการเทียบคำที่ยาวที่สุดแล้ว จึงเลือกทางเลือกที่มีจำนวนคำที่น้อยที่สุด การค้นหาทุกทางเลือกที่เป็นไปได้ทำให้ต้องเสียเวลาในการคำนวณมาก แต่สามารถลดเวลาลงได้โดยใช้โปรแกรมแบบพลวัติ (Dynamic Program) ตัวอย่างเช่น หากป้อนข้อความ “ไปหามเหสี” เข้าไปขั้นตอนวิธีนี้จะหาทางเลือกทั้งหมดของรูปแบบการตัดคำที่เป็นไปได้ ได้แก่

ไป (go) หาม (carry) เห (deviate) สี (color)

ไป (go) หา (see) [ม] เห (deviate) สี (color)

ไป (go) หา (see) มเหสี (queen)

โดยในขั้นแรก ขั้นตอนวิธีของการเทียบคำที่ยาวที่สุดจะได้คำตอบของการตัดคำเป็น “ไป-หาม-เห-สี” ก่อน หลังจากนั้นจึงเริ่มทำการย้อนกลับโดยเริ่มจากคำแรก พบว่า คำที่สอง “หาม” สามารถแบ่งเป็น “หา-ม” ได้ โดยเมื่อแบ่งเป็น “หา” แล้ว ทำให้สายอักขระที่เหลือ “มเหสี” สามารถตัดคำได้อีก 2 แบบ คือ “มเหสี” และ “[ม]-เห-สี” เมื่อทำการย้อนกลับกับทุกคำสิ้นสุดแล้ว ก็จะทำการคำนวณหาค่าต้นทุน (Cost) ให้กับแต่ละทางเลือกที่เป็นไปได้ โดยบังคับให้มีการเกิดคำที่ไม่มีในพจนานุกรมน้อยที่สุด แล้วจัดเรียงผลลัพธ์โดยให้ทางเลือกที่น่าจะเป็นไปได้มากที่สุด (ค่าต้นทุนต่ำสุด) มาเป็นอันดับแรก ตัวอย่างเช่น ข้อความที่ป้อนเข้าเป็น “กีฬาเป็นการออกกำลังกายอย่างหนึ่ง” จะให้ผลทางเลือกที่เป็นไปได้จัดเรียงตามค่าต้นทุน

ตาราง 2.2 ผลการตัดคำด้วยวิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด

ผลจากการทำย้อนกลับ	ค่าต้นทุน
กีฬา/ เป็น/ การออกกำลังกาย/ อย่างหนึ่ง	4
กีฬา/ เป็นการ/ ออกกำลัง/ กาย/ อย่างหนึ่ง	5
กีฬา/ เป็นการ/ ออก/ กำลังกาย/ อย่างหนึ่ง	5
กีฬา/ เป็นการ/ ออกกำลัง/ กาย/ อย่าง/ หนึ่ง	6
กีฬา/ เป็นการ/ ออกกำลัง/ กา/ ยอ/ ย่าง/หนึ่ง	7
กีฬา/ เป็นการ/ ออ/ [ก]/ กำลังกาย/อย่างหนึ่ง	11
กีฬา/ เป็นการ/ ออกกำลัง/ กาย/ อย่า/ [ง]/ หนึ่ง	12
กีฬา/ เป็นการ/ ออ/ [ก]/ กำลังกาย/ อย่าง/ หนึ่ง	12
กีฬา/ เป็นการ/ ออ/ [ก]/ กำลังกาย/ อย่า/ [ง]/ หนึ่ง	18

ขั้นตอนวิธีนี้จะเลือกการตัดคำที่มีค่าต้นทุนต่ำที่สุด ขึ้นอยู่กับจำนวนคำที่ตัดออกมาได้และจำนวนคำที่ไม่มีในพจนานุกรม อย่างไรก็ตาม หากมีทางเลือกที่มีค่าต้นทุนเท่ากันและมีจำนวนคำเท่ากันมากกว่าหนึ่งทางเลือก ขั้นตอนวิธีนี้จะไม่สามารถตัดสินได้ว่าจะเลือกทางไหน ดังนั้นจึงต้องใช้วิทยาการศึกษาสำนึกรูปแบบอื่นเข้ามาช่วย โดยส่วนใหญ่มักใช้วิธีการเทียบคำที่ยาวที่สุดเข้ามาช่วย

Surapant Meknavin and Others (1997) กล่าวว่า วิธีการนี้มักจะมีประสิทธิภาพในการตัดคำสูงกว่าวิธีการเทียบคำที่ยาวที่สุด เนื่องจากวิธีการนี้ได้แก้ไขข้อจำกัดในการพยายามเลือกคำให้ยาวที่สุดตั้งแต่ต้น โดยใช้วิธีพิจารณาทางเลือกของการตัดคำที่เป็นได้ทั้งหมดก่อนที่จะเลือกการตัดคำ

(ค) หลักการตัดคำโดยใช้คลังข้อมูล (Corpus Based Approach) การตัดคำโดยใช้คลังข้อมูลเป็นการตัดคำโดยนำวิธีการทางสถิติ (Statistical Techniques) เข้ามาใช้ในการประมวลผล โดยใช้คลังข้อมูลทางภาษา (Corpus) เป็นฐานความรู้ในการตัดคำ เพื่อแก้ปัญหาของคำที่ไม่มีในพจนานุกรม เช่น ชื่อเฉพาะ คำที่ยืมมาจากภาษาต่างประเทศ เป็นต้น และความคลุมเครือในการแบ่งขอบเขตของคำได้อย่างมีประสิทธิภาพ แบ่งได้เป็น 2 แนวทาง คือ

1) การตัดคำโดยอาศัยความน่าจะเป็น (Probabilistic Word Segmentation) วิธีการนี้จะนำเอาค่าทางสถิติการเกิดของคำและลำดับของหน้าที่ของคำ (Part–of-Speech) เข้ามาช่วยในการคำนวณหาความน่าจะเป็น เพื่อที่จะใช้เลือกแบบที่มีโอกาสเกิดมากที่สุด วิธีการนี้สามารถจะตัดคำได้ดีกว่า 2 แบบแรก แต่ข้อจำกัดของวิธีการนี้ คือ จะต้องมีฐานข้อมูลที่มีการตัดคำที่ถูกต้อง และกำหนดหน้าที่ของคำให้เพื่อที่จะได้นำไปใช้ในการสร้างสถิติ

2) การตัดคำโดยอาศัยคุณลักษณะของคำ (Feature Based Word Segmentation) วิธีนี้การนี้จะพิจารณาจากบริบท (Context) และการเกิดร่วมกันของคำ หรือหน้าที่ของคำ (Collocation) เข้ามาช่วยในการตัดคำ ตัวอย่างเช่น

"ตากลม" ถ้าพบคำว่า "แป๋ว" ในบริบทก็จะสามารถตัดคำได้ว่า "ตา" "กลม"

"มากว่า" ถ้าในบริบทที่ตามมาเป็นตัวเลขก็สามารถตัดคำได้ว่า "มา" "กว่า"

วิธีการตัดคำโดยใช้คลังข้อมูลนี้จำเป็นที่จะต้องมีฐานข้อมูลเป็นจำนวนมาก และจะต้องมีการเรียนรู้การสร้างคำในบริบท หรือการเกิดร่วมกันของคำแต่ละคำ เพื่อให้มีข้อมูลที่จะนำมาใช้ในการตัดคำ

2.1.9 เอ็น-แกรม (N-Gram)

เอ็น-แกรม คือ แบบจำลองที่ใช้คำนวณค่าความน่าจะเป็นของชุดอักขระ (Character Sequence) ที่เกิดขึ้นร่วมกันเป็นคำ หรือค่าความน่าจะเป็นของคำที่เขียนเรียงกัน (Word Sequence) ที่เกิดขึ้นร่วมกันเป็นประโยค โดยค่าความน่าจะเป็นของชุดอักขระหรือคำ ประมาณได้จากคลังข้อมูลที่สร้างไว้ ซึ่งเอ็น-แกรมได้ใช้หลักการของสถิติหลาย ๆ ด้านมาประยุกต์ใช้

แกรม (Gram) คือ หน่วยที่ใช้ในการสร้างแบบจำลอง อาจจะเป็นเสียง คำ หรือ อักขระก็ได้ ซึ่งแกรมมีได้หลายขนาดแล้วแต่จะกำหนด ตั้งแต่ 1 จนถึง เอ็น (n) ในแบบจำลองเอ็น-แกรมนี้ใช้ความยาวของสายอักขระและสายคำแตกต่างกัน ได้แก่ 2-3 แกรม และ 4 แกรม เป็นต้น ถ้าจะประมาณค่าความน่าจะเป็นของสายคำหรือสายอักขระจากคลังข้อมูลโดยการใช้วิธีเอ็น-แกรมผลที่ได้มีดังนี้

การประมาณค่าด้วย 2-แกรม (Probability Bigram) คือ การประมาณค่าความน่าจะเป็นของสายอักขระ (สายคำ) ที่เกิดขึ้นร่วมกันว่ามีค่าเท่ากับผลคูณของความน่าจะเป็นที่จะพบอักขระ (คำ) ทีละ 2 ตัว (คำ) ติดกันในสายอักขระนั้น

การประมาณค่าด้วย 3-แกรม (Probability Trigram) คือ การประมาณค่าความน่าจะเป็นของสายอักขระที่เกิดขึ้นร่วมกันว่ามีค่าเท่ากับผลคูณของความน่าจะเป็นที่จะพบอักขระทีละ 3 ตัว ติดกันในสายอักขระนั้น

วิธีการใช้แบบจำลองเอ็น-แกรมนี้เป็นวิธีการทางสถิติที่นิยมใช้กันมากที่สุด เพราะเป็นวิธีที่เรียบง่าย มีประสิทธิภาพสูง และเหมาะสำหรับวิเคราะห์ภาษา สามารถใช้ระบุภาษาได้ดีกว่าวิธีอื่น ๆ

หลักการตัดคำโดยอาศัยเอ็น-แกรมเพื่อหาเศษของคำ เป็นการตัดคำอีกแนวคิดหนึ่งเพื่อแก้ไขข้อจำกัดของการหาขอบเขตของคำในภาษาไทย ซึ่งช่วยลดข้อผิดพลาดที่เกิดจากข้อความและไม่จำเป็นต้องใช้การหารากศัพท์และการตัดคำหยุด การตัดคำโดยใช้เอ็น-แกรมสามารถหาได้จากลำดับของอักขระขนาดเอ็นตัวติดกัน สร้างจากคำโดยใช้วิธีการเลื่อนไปทีละเอ็นตัวอักษร โดยไม่ต้องอาศัยหลักความรู้ทางภาษาศาสตร์ ไม่สนใจความหมายของคำ เป็นอีกวิธีการในการทำดัชนีที่

เป็นที่ยอมรับ สำหรับวิธีการเอ็น-แกรมนั้นตัวอักษรตัวใหญ่และตัวเล็กมองเหมือนกันหมด แต่เนื่องจากภาษาไทยประกอบไปด้วยสระและวรรณยุกต์ จึงไม่สามารถกำหนดได้ว่า 1 ไบต์ (Byte) หรือ 1 ตัวอักษรนั้น คือ 1 แกรม ดังนั้น สำหรับการตัดคำภาษาไทยจึงต้องทำการตรวจสอบเบื้องต้นก่อนว่าพยัญชนะนั้นมีสระหรือวรรณยุกต์หรือไม่และทำการตัดช่องว่างทิ้ง ปัญหาในการพัฒนาขั้นตอนวิธีการตัดคำของคำที่ไม่สามารถระบุความหมายในคลังเอกสารภาษาไทย คือ ภาษาไทยไม่มีการใช้อักขระหรือเครื่องหมายพิเศษที่บอกความแตกต่างระหว่างคำเฉพาะและคำทั่วไป

2.1.10 ลำดับของหน้าที่ของคำ (Part of Speech)

ธีรศักดิ์ สังข์ศรี (2551) กล่าวไว้ว่า ลำดับของหน้าที่ของคำ หมายถึง คำศัพท์หนึ่งคำในเอกสารอาจเป็นได้ทั้งคำนามหรือคำกริยา ซึ่งขึ้นอยู่กับหน้าที่ของคำนั้น ๆ ในประโยค ตัวอย่างเช่น คำนาม คำกริยา และคำบุพบท เป็นต้น โดยในขั้นตอนนี้จะทำการแยกแยะหน้าที่ของคำในประโยคเพื่อระบุให้เกิดความชัดเจนในหน้าที่ของคำนั้น ๆ ซึ่งคำ ๆ หนึ่งทำหน้าที่ได้หลายอย่างและมีความหมายไม่เหมือนเดิม การวิเคราะห์หน้าที่ของคำจึงถือเป็นการลดความกำกวมของคำในระดับหนึ่ง

กำชัย ทองหล่อ (2525) กล่าวไว้ว่า ชนิดของคำ หรือลำดับของหน้าที่ของคำ คือ การนำคำ หรือวลีที่ผ่านการตัดคำที่ได้ออกมาเป็นหน่วยคำต่าง ๆ ซึ่งจะมีการแสดงคำออกเป็นคำตามกลุ่มคำต่าง ๆ แล้วยังได้มีการบอกถึงคุณลักษณะ หรือหน้าที่ของคำ โดยจำแนกกลุ่มคำออกเป็น 7 กลุ่มคำใหญ่ ๆ ดังนี้ กลุ่มคำนาม (Nouns) กลุ่มคำสรรพนาม (Pronouns) กลุ่มคำกริยา (Verbs) กลุ่มคำวิเศษณ์ (Modifiers) กลุ่มคำบุพบท (Prepositions) กลุ่มคำสันธาน (Conjunctions) และกลุ่มคำอุทาน (Exclamations and Interjections)

2.2 แบบจำลองเวกเตอร์สเปซ (Vector Space Model)

แบบจำลองเวกเตอร์สเปซพัฒนาโดยแซลตั่น (Salton 1989) โดยเอกสารแต่ละฉบับเปรียบเสมือนกับเวกเตอร์ของคำ โดยที่ขนาดของเวกเตอร์ขึ้นกับจำนวนของคำที่ปรากฏอยู่ในเอกสารฉบับนั้น

แบบจำลองเวกเตอร์สเปซ หรือแบบจำลองเวกเตอร์ของคำ เป็นแบบจำลองทางพีชคณิตในการนำเสนอเอกสารข้อความที่ใช้เวกเตอร์เป็นตัวระบุ ตัวอย่างเช่น การใช้ดัชนีคำศัพท์ โดยส่วนมากจะใช้แบบจำลองเวกเตอร์สเปซในงานการคัดกรองข้อมูล ระบบงานค้นคืนสารสนเทศ การทำเหมืองข้อมูล การทำดัชนีและการจัดอันดับ เป็นต้น แรกเริ่มนั้นแบบจำลองเวกเตอร์สเปซถูกนำมาใช้งานในระบบงานค้นคืนสารสนเทศสมาร์ต (SMART Information Retrieval System) ซึ่งเป็นระบบค้นคืนสารสนเทศที่พัฒนาโดยมหาวิทยาลัยคอร์แนล (Cornell University) ในปี ค.ศ. 1960 ระบบงานค้นคืนสารสนเทศสมาร์ตนี้ แนวคิดสำคัญมากมายในงานค้นคืนสารสนเทศนั้นถูกพัฒนามาจากส่วน

หนึ่งของงานวิจัยระบบงานสมาร์ตรวมถึง แบบจำลองเวกเตอร์สเปซ การคืนค่าผลลัพธ์ (Relevance Feedback) และ การแบ่งแยกออกเป็นประเภท ๆ ของรอคชิโอ (Rocchio Classification)

2.2.1 นิยามของแบบจำลองเวกเตอร์สเปซ

เอกสารและข้อสอบถาม (Query) จะถูกนำเสนอในรูปแบบของเวกเตอร์ โดยที่ในแต่ละมิติที่เหมือนกันจะแบ่งคำออกจากกัน ถ้าคำที่มีอยู่ในเอกสาร และค่าของเวกเตอร์ไม่เท่ากับศูนย์ หลายทางที่แตกต่างในการคำนวณค่าซึ่งจะรู้จักในชื่อของค่าน้ำหนักคำ (Term Weights) ได้ถูกพัฒนาขึ้น แนวทางหนึ่งที่เป็นที่รู้จักดีคือการหาค่าน้ำหนัก (TF-IDF Weighting)

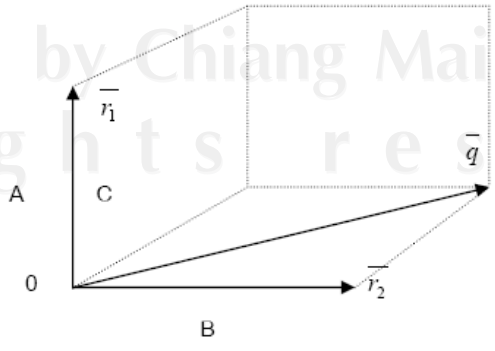
คำนิยามของคำจะขึ้นอยู่กับการประยุกต์ใช้ ซึ่งโดยปกติแล้วถ้าพูดถึงคำจะหมายถึงคำเดียว คำสำคัญ หรืออาจเป็นวลียาว ๆ ก็ได้ ถ้าคำเหล่านั้นถูกเลือกให้เป็นขอบเขต มิติของเวกเตอร์จะเป็นจำนวนของคำที่อยู่ในคำศัพท์ การดำเนินงานของเวกเตอร์สามารถนำไปใช้ในการเปรียบเทียบระหว่างเอกสารกับข้อสอบถามได้

ฆนาศัย กรึงไกร และ ชุลีรัตน์ จรัสกุลชัย (2550) ได้นำหลักการของแบบจำลองเวกเตอร์สเปซมาใช้ในการอธิบายถึงวิธีการแทนเอกสารที่ไม่มีโครงร่าง (Unstructured Text Document) โดยแทนเอกสารแต่ละฉบับซึ่งเปรียบเสมือนกับเวกเตอร์ของคำ โดยที่ขนาดของเวกเตอร์ขึ้นกับจำนวนคำที่ปรากฏอยู่ในเอกสารฉบับนั้น หลังจากนั้นจะทำการคำนวณหาความคล้ายคลึงของเอกสารคู่หนึ่ง ๆ ได้จากค่าสัมประสิทธิ์ระหว่างมุมของคู่เวกเตอร์ จากผลการทดลองพบว่าประสิทธิผลที่ได้จากการจัดกลุ่มอยู่ในเกณฑ์ดีระดับหนึ่ง

ตัวอย่าง

เอกสาร 1	ประกอบด้วย A และ C
เอกสาร 2	ประกอบด้วย B และ C
ข้อสอบถาม	ประกอบด้วย B และ C

เมื่อแทนเอกสารให้อยู่ในรูปแบบจำลองเวกเตอร์สเปซ จะได้เวกเตอร์ที่มีลักษณะดังนี้



รูป 2.2 แสดงเวกเตอร์ของเอกสาร 1 เอกสาร 2 และข้อสอบถาม

เอกสาร 1 แทนด้วยเวคเตอร์ $\overline{r_1}$

เอกสาร 2 แทนด้วยเวคเตอร์ $\overline{r_2}$

ข้อสอบถาม แทนด้วยเวคเตอร์ $\overline{q}$

เนื่องจากทั้งสองเอกสาร คือ ทั้งเอกสาร 1 และเอกสาร 2 ต่างมี C เป็นองค์ประกอบ ดังนั้นจึงให้ค่าน้ำหนักของ C เป็น 0 เมื่อนำมาเขียนอยู่ในรูปเมตริก จะได้ดังนี้

$$\overline{r_1} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \overline{r_2} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \overline{Q} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

รูป 2.3 แสดงเมตริกของเอกสาร 1 เอกสาร 2 และข้อสอบถาม

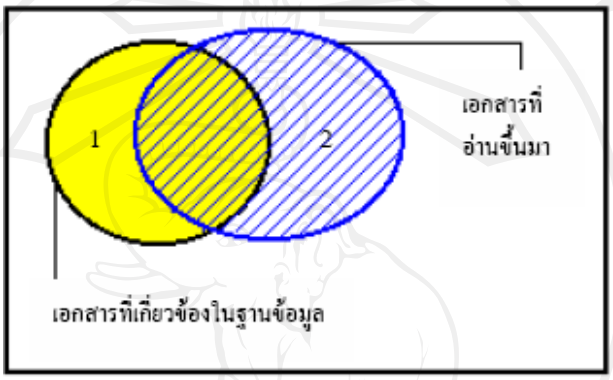
ชุลีรัตน์ จรัสกุลชัย และคณะ (2544) ได้กล่าวถึงหลักการของแบบจำลองเวกเตอร์สเปซไว้ว่า แบบจำลองเวกเตอร์สเปซจะเป็นการแทนเอกสาร และข้อสอบถาม ข้อมูลสืบค้นหรือข้อมูลนำเข้าด้วยเวกเตอร์ที่เรียกว่า เวกเตอร์ของคำ (Vector of Terms) โดยแกนแต่ละแกนของเวกเตอร์ คือ คำศัพท์ (Terms) ที่ปรากฏอยู่ในเอกสาร คำเหล่านี้ได้จากการคัดเลือกคำทั้งหมดที่ปรากฏอยู่ในทุกเอกสาร อาจบันทึกเก็บไว้ในลักษณะพจนานุกรมเพื่อใช้ในการอ้างอิงในการสร้างเวกเตอร์ของคำที่ตรงกันทั้งฐานข้อมูล และความยาวของเวกเตอร์ คือค่าความถี่ของคำที่ถูกเลือก ซึ่งปรากฏอยู่ในเอกสารนั้นประกอบ

ณัฐพร หอมเมือง (2550) ได้นำหลักการของแบบจำลองเวกเตอร์สเปซผสมผสานกับหลักการให้ค่าน้ำหนักคำมาใช้ในการพัฒนาระบบถามตอบ แล้วทำการวัดความคล้ายคลึงด้วยการหาค่าระยะทางแบบยูคลิเดียน (Euclidian Distance) ซึ่งการวัดระยะทางแบบยูคลิเดียนนั้นจะเป็นการวัดระยะห่างระหว่างข้อมูลที่ต้องการเปรียบเทียบความคล้ายคลึง ข้อจำกัดของการวัดวิธีนี้คือ ค่าระยะห่างจะมีความไว (Sensitive) กับขนาดของวัตถุที่แตกต่างกัน และค่าความห่างที่คำนวณได้ไม่ได้สนใจในความเกี่ยวข้องสัมพันธ์กัน (Correlation) และผลจากการทดลองที่ได้พบว่าในส่วนการค้นหาข้อมูลยังไม่ค่อยมีประสิทธิภาพเนื่องจากกระบวนการเปรียบเทียบกับดัชนียังอยู่ในรูปแบบที่ไม่เหมาะสม

2.3 การวัดประสิทธิภาพของระบบ (Retrieval Effectiveness)

Frankes and Baeza Yates (1992) กล่าวไว้ว่า การวัดประสิทธิภาพของระบบมีอยู่หลายวิธี แต่สองวิธีที่นิยมตามมาตรฐานของระบบค้นคืนเอกสารคือ การใช้การวัดค่าความแม่นยำของข้อมูลและค่าความถูกต้องของข้อมูล

Richard (2000) กล่าวไว้ว่า การวัดค่าความแม่นยำของข้อมูลและการวัดค่าความถูกต้องของข้อมูลนั้นเป็นวิธีการหนึ่งสำหรับการประเมินประสิทธิภาพการค้นคืนเอกสาร



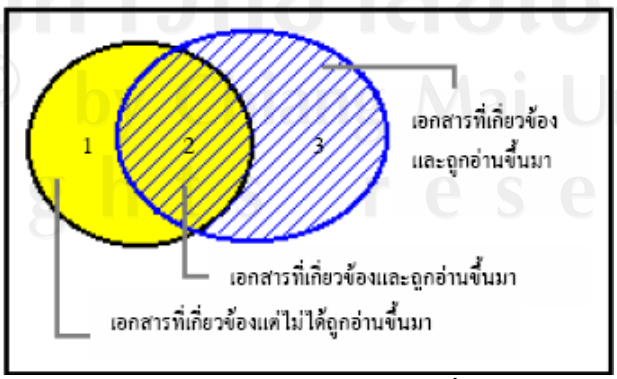
รูป 2.4 แสดงความสัมพันธ์ระหว่างกลุ่มข้อมูลในฐานข้อมูลที่เกี่ยวข้องกับข้อสอบถาม

จากรูปจะแสดงความสัมพันธ์ระหว่างกลุ่มข้อมูลในฐานข้อมูลที่เกี่ยวข้องกับข้อสอบถามโดยที่

(ก) กลุ่มของเอกสารที่เก็บไว้ในฐานข้อมูลซึ่งมีความสัมพันธ์กันกับข้อมูลของข้อสอบถามจะเป็นส่วนวงกลมหมายเลข (1)

(ข) กลุ่มของเอกสารที่ถูกเลือกขึ้นมาเป็นผลลัพธ์จะเป็นส่วนวงกลมหมายเลข (2)

ในการสืบค้นบางครั้งนั้นอาจสืบค้นเอกสารที่ไม่ได้เกี่ยวข้องแต่มีการค้นคืนมาให้ดังจะแสดงตามรูป 2.5



รูป 2.5 แสดงความสัมพันธ์ระหว่างกลุ่มของข้อมูลที่ใช้สืบค้นกับข้อสอบถาม

จากรูปจะแสดงความสัมพันธ์ระหว่างกลุ่มของข้อมูลที่ใช้สืบค้นกับข้อสอบถาม โดยที่

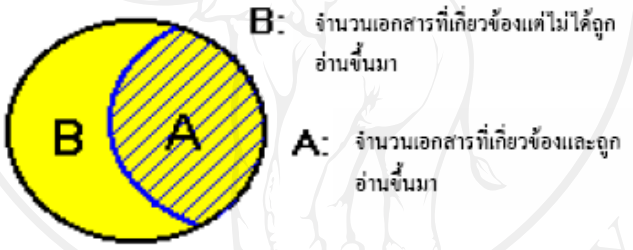
(ก) กลุ่มของเอกสารที่เกี่ยวข้องกับข้อสอบถามแต่ไม่ได้เลือกขึ้นมาคือส่วนหมายเลข (1)

(ข) กลุ่มของเอกสารที่เกี่ยวข้องกับข้อสอบถามและถูกเลือกขึ้นมาคือส่วนหมายเลข (2)

(ค) กลุ่มของเอกสารที่ไม่เกี่ยวข้องกับข้อสอบถามแต่ถูกเลือกขึ้นมาคือส่วนหมายเลข (3)

การหาค่าความถูกต้องของข้อมูล หรือค่าความระลึก (Recall) เมื่อ B คือ จำนวนของเอกสารที่เกี่ยวข้องหรือสัมพันธ์กับข้อสอบถามแต่ไม่ได้ถูกเลือกมาเป็นผลลัพธ์ และ A คือ จำนวนของเอกสารที่เกี่ยวข้องหรือสัมพันธ์กับข้อสอบถามและถูกเลือกมาเป็นผลลัพธ์ ซึ่งสามารถคำนวณได้ตามสมการ (2.1)

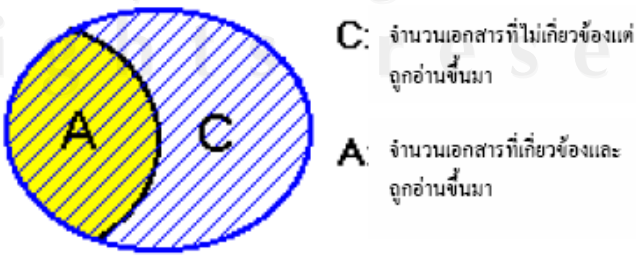
$$\text{ค่าความถูกต้อง} = \frac{A}{A+B} \times 100\% \quad \text{..... (2.1)}$$



รูป 2.6 แสดงการคำนวณการหาค่าความถูกต้องของข้อมูล

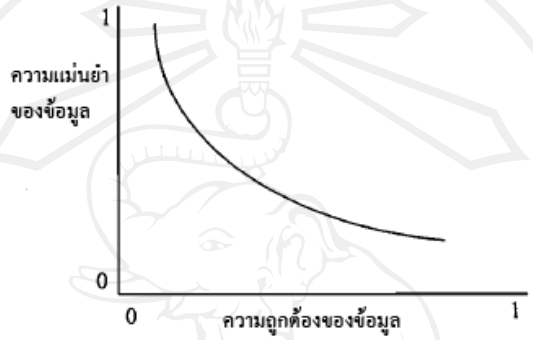
การหาค่าของความแม่นยำของข้อมูล (Precision) เมื่อ C คือ จำนวนของเอกสารที่ไม่เกี่ยวข้องหรือไม่สัมพันธ์กับข้อสอบถามแต่ถูกเลือกมาเป็นผลลัพธ์ และ A คือ จำนวนของเอกสารที่เกี่ยวข้องหรือสัมพันธ์กับข้อสอบถามและถูกเลือกมาเป็นผลลัพธ์ ซึ่งสามารถคำนวณได้ตามสมการ (2.2)

$$\text{ค่าความแม่นยำ} = \frac{A}{A+C} \times 100\% \quad \text{..... (2.2)}$$



รูป 2.7 แสดงการคำนวณหาค่าความแม่นยำของข้อมูล

ค่าความถูกต้องและค่าความแม่นยำจะมีค่าอยู่ระหว่าง 0 กับ 1 ถ้าการค้นคืนเอกสารนั้นได้เอกสารออกมาตรงกับความต้องการทั้งหมดและไม่มีเอกสารที่ไม่เกี่ยวข้องออกมาด้วยค่าความถูกต้องและค่าความแม่นยำจะมีค่าเป็น 1 โดยทั่วไปค่าความถูกต้องและค่าความแม่นยำจะมีความสัมพันธ์เป็นปฏิภาคผกผัน คือ ถ้าค่าความแม่นยำสูงขึ้น ค่าความถูกต้องก็จะมักลดลง และในทางตรงกันข้ามถ้าค่าความถูกต้องสูงขึ้น ค่าความแม่นยำก็มักจะลดลง แต่ในบางกรณีค่าความแม่นยำและค่าความถูกต้องอาจจะเพิ่มขึ้นหรือลดลงด้วยกันทั้งสองค่าก็ได้



รูป 2.8 ความสัมพันธ์ระหว่างค่าความแม่นยำและค่าความถูกต้อง

2.4 งานวิจัยที่เกี่ยวข้อง

Christian Platzer and Schahram Dustdar (2005) กล่าวไว้ว่า เว็บเซอร์วิสเริ่มที่จะมีความสำคัญมากขึ้นในด้านของคอมพิวเตอร์แบบกระจาย (Distributed Computing) แต่ความผิดพลาดบางประการและข้อจำกัดเทคโนโลยีนี้ก็เริ่มปรากฏชัดมากขึ้นเรื่อย ๆ เช่นกัน และหนึ่งข้อผิดพลาดนั้นก็ถูกค้นพบผ่านทางเมดธอด (Method) ธรรมดา ๆ โดยในงานวิจัยนี้จะกล่าวถึงในส่วนของกระบวนการใหม่ในการค้นหาและวิเคราะห์เว็บเซอร์วิส โดยใช้หลักการของระบบสืบค้นแบบเวกเตอร์สเปซ (Vector Space Search Engine) ในการทำดัชนีที่ใช้ในการอธิบายของเว็บเซอร์วิสที่มีอยู่แล้ว โดยในสารนิพนธ์ใช้รูปแบบของแบบจำลองเวกเตอร์สเปซ โดยทำการสร้างเวกเตอร์สเปซหนึ่งตัวสำหรับเอกสารแต่ละอันที่เก็บรวบรวมไว้ โดยเอกสารนั้นจะถูกนำเสนอในรูปของเทอมดัชนี (Index Term) ที่ถ่วงน้ำหนักโดยข้อสอบถามของผู้ใช้ สามารถหาความสอดคล้องโดยใช้วิธีทางเวกเตอร์ในการค้นหาและใช้วิธีเกี่ยวกับเรขาคณิตโดยการหาโคไซน์ของมุมระหว่างเวกเตอร์ของข้อสอบถามกับเอกสารที่มีอยู่ เพื่อหาความสอดคล้องกันของข้อสอบถามกับเอกสารที่มีอยู่ว่าความคล้ายคลึงกันมากน้อยเพียงใด ก่อนที่จะทำการแสดงผลแก่ผู้ใช้

Dik Lun Lee (1997) กล่าวไว้ว่า ประสิทธิภาพและการใช้งานของเทคนิคการสืบค้นนั้นมีความสำคัญในการจัดการเพิ่มจำนวนเอกสารที่มีอยู่แล้ว และแม้ว่าการสืบค้นข้อมูลจากข้อความนั้นจะเป็นงานที่ค่อนข้างยาก เพราะว่าเป็นการยากที่จะจำแนกข้อความกับภาษาธรรมชาติออกจากกัน โดยที่เทคนิคในการประมวลผลภาษาธรรมชาตินั้นจะสามารถทำได้อย่างมีประสิทธิภาพเมื่อใช้กับกลุ่มของข้อความจำนวนมาก เทคนิคที่ใช้ในการสืบค้นข้อความที่มีอยู่ส่วนใหญ่นั้นจะอาศัยการใช้ดัชนีคำในการสืบค้นเพียงอย่างเดียว และเป็นเรื่องที่ไม่ดีนักที่ว่าการใช้เพียงดัชนีคำเพียงอย่างเดียวนั้น หรือหาคำเพียงอย่างเดียวนั้นไม่สามารถพอที่จะจับเนื้อหาในเอกสารได้ ประสิทธิภาพในการค้นหานั้นต่ำ จึงได้นำหลักการของแบบจำลองเวกเตอร์สเปซมาใช้เพื่อหาจุดสมดุลที่ดีระหว่างประสิทธิภาพในการประมวลผลกับประสิทธิภาพในการค้นคืนเอกสาร โดยแสดงออกในรูปแบบของการทำระดับของเอกสารที่สัมพันธ์กัน (Relevant Document Ranking)

Raymie Stata and Others (2000) กล่าวในงานวิจัยเรื่องฐานข้อมูลเวกเตอร์ของคำ (Term Vector Database) ได้นำหลักการของแบบจำลองเวกเตอร์สเปซ โดยให้ค่าองค์ประกอบของเวกเตอร์ (Vector Element) เป็น 0 ถ้าไม่พบข้อมูล เช่น ในการค้นหาข้อมูล a b c d e และ f และเมื่อพบข้อมูลเพียง a b c และ d โดยมีค้าน้ำหนักเป็น 10 20 30 40 50 และ 60 ทำให้เวกเตอร์สเปซ เป็น (a b c d e และ f) และระยะห่างเอกสาร (Document Space) เป็น (10 20 30 40 0 และ 0) และอาจทำการคำนวณน้ำหนักโดยใช้ค่าน้ำหนักคำของ TF และ IDF โดยใช้สูตรในการคำนวณ คือ TF*IDF โดยที่ TF (Term Frequency) คือ ความถี่ที่พบข้อมูลในเอกสารยิ่งมากยิ่งดีและ IDF (Inverse Document Frequency) คือ 1/จำนวนที่พบข้อมูลทั้งหมด เช่น การพบคำว่า “algebra” มีค่ามากกว่าคำว่า “the”

Resnik P. (1999) ได้กล่าวเกี่ยวกับการให้ค่าน้ำหนักคำไว้ว่า น้ำหนักของคำแต่ละคำ (ทั้งคำที่มีในข้อสอบถามและคำในเอกสาร) จะคำนวณโดยเปรียบเทียบกับคำทั้งหมดในเอกสาร ค่าน้ำหนักคำในเอกสารจะได้รับการดำเนินการในขั้นตอนการเตรียมฐานความรู้ ส่วนค่าน้ำหนักคำในข้อสอบถามจะดำเนินการในแต่ละครั้งที่มีการสืบค้น เมื่อคำนวณค่าน้ำหนักคำของข้อสอบถามได้เรียบร้อยแล้ว ค่าน้ำหนักเหล่านั้นจะถูกนำไปแทนในสมการเพื่อคำนวณระดับความคล้ายคลึงระหว่างข้อสอบถามและเอกสาร เรียกว่า วิธีวัดระดับความเหมือนระหว่างเอกสารโดยใช้วิธีการแบบเวกเตอร์ (Vector Based Similarity Measurement)

ณัฐพร หอมเมือง (2550) ได้นำหลักการให้ค่าน้ำหนักคำมาคำนวณร่วมกับการนับความถี่ของคำ โดยกล่าวถึงหลักการนี้ว่า คำที่มีความสำคัญหรือใช้เป็นตัวแทนของเอกสารที่ดีควรจะปรากฏอยู่เป็นจำนวนมากในเนื้อหาของเอกสารเฉพาะฉบับนั้น และปรากฏอยู่น้อยมากในชุดเอกสารที่เหลือทั้งหมด แต่ถ้าคำนั้นปรากฏเป็นจำนวนมากในทุก ๆ เอกสารแสดงว่าคำดังกล่าวไม่สามารถใช้เป็นตัวแทนของเอกสารใด ๆ ได้ ซึ่งเรียกคำเหล่านี้ว่าคำหยุด (Stop Word) ดังนั้น ในการให้ค่าน้ำหนัก

คำ ๆ หนึ่ง ในเอกสารฉบับหนึ่งจะพิจารณาจากความถี่ของคำ (Term Frequency) ที่ปรากฏในเอกสารนั้นและจำนวนของเอกสารทั้งหมดที่มีคำ ๆ นั้นปรากฏอยู่

บังอร กลับบ้านเกาะ (2543) ได้เสนอเกี่ยวกับวิธีการค้นคืนเอกสารสารสนเทศออนไลน์โดยประยุกต์ใช้จีเนติกอัลกอริทึมเพื่อเพิ่มประสิทธิภาพในการค้นคืนสารสนเทศ ภายใต้แบบจำลองเวกเตอร์สเปซ การค้นคืนสารสนเทศใดนั้นขึ้นอยู่กับความคล้ายคลึงระหว่างเอกสารและข้อสอบถาม เอกสารใดที่มีความคล้ายคลึงกับข้อสอบถามสูงย่อมแสดงว่าเอกสารนั้นมีความสัมพันธ์กับข้อสอบถามมากกว่า และควรจะได้รับการค้นคืนขึ้นมาก่อน ในขั้นตอนจีเนติกอัลกอริทึมนั้นข้อสอบถามจะถูกแทนด้วยโครโมโซม ซึ่งโครโมโซมเหล่านี้จะถูกนำเข้าสู่กระบวนการจีเนติกโอเปอเรเตอร์ต่าง ๆ อันได้แก่ การคัดเลือก การครอสโอเวอร์ และการมิวเตชั่น จนกระทั่งได้โครโมโซมข้อสอบถามที่เหมาะสมเพื่อนำไปค้นคืนสารสนเทศต่อไป

พิลาวัณย์ พลับรู้การ (2544) กล่าวว่าในปัจจุบันระบบที่ใช้ในการเข้าถึงเอกสารได้เพิ่มขีดความสามารถในการอำนวยความสะดวกให้แก่ผู้ใช้งาน เช่น สามารถเรียกค้นเอกสารที่คล้ายคลึงกับเอกสารที่ผู้ใช้ให้เป็นตัวอย่าง หรือการแบ่งกลุ่มเอกสารที่ได้จากระบบสืบค้น ความสามารถดังกล่าวจำเป็นต้องใช้วิธีการในการวัดความคล้ายคลึงเชิงมุม ซึ่งใช้หลักการแทนเอกสารด้วยระบบเวกเตอร์และหลักการทางสถิติในการวัด ข้อเสียของการวัดความคล้ายคลึงแบบนี้คือจะไม่มีการคำนึงถึงความหมายของคำ โดยจะถือว่าคำแต่ละคำเป็นอิสระต่อกัน ซึ่งในความเป็นจริงแล้วคำแต่ละคำอาจมีความใกล้เคียงทางความหมายต่อกัน ในงานวิจัยนี้ได้เสนอวิธีการวัดความคล้ายคลึงระหว่างเอกสารโดยใช้แนวทางด้านความหมายเข้ามาช่วย กล่าวคือจะนำโครงข่ายความสัมพันธ์ของคำมาใช้ในการคำนวณค่าความคล้ายคลึงระหว่างเอกสาร 2 เอกสาร จากผลการทดสอบกับเอกสารในโดเมนคอมพิวเตอร์พบว่าวิธีที่ใช้แนวทางด้านความหมายได้ผลที่ได้ดีกว่าการวัดความคล้ายคลึงแบบที่ใช้กันทั่วไป

มานพ จงเจริญใจ (2541) ได้กล่าวไว้ในวิทยานิพนธ์ที่เสนอการพัฒนาระบบค้นคืนสารสนเทศในรูปแบบเอกสารที่เป็นข้อความด้วยวิธีค้นคืนแบบจัดลำดับและแบบค้นคืนย้อนกลับโดยใช้แถวลำดับแพ็ตเป็นดัชนีเพื่อใช้ในการสืบค้น แถวลำดับแพ็ตเป็นโครงสร้างที่เหมาะกับข้อความภาษาไทยที่การแบ่งคำยังไม่ถูกต้องสมบูรณ์ โดยแถวลำดับแพ็ตจะจัดเก็บดัชนีในรูปของสายอักขระแบบกึ่งอนันต์ที่เรียกว่า ซีสตริง การพัฒนาโปรแกรมค้นคืนแบ่งออกเป็น 3 ส่วน คือ ส่วนของการสร้างดัชนีของค่าน้ำหนักคำ ส่วนของการจัดลำดับผลการค้นคืน และส่วนของการค้นคืนย้อนกลับ สำหรับส่วนของการสร้างดัชนีค่าน้ำหนักคำจะเก็บค่าตัวชี้ตำแหน่งซีสตริงที่ไม่ซ้ำกันและค่าความถี่ของแต่ละซีสตริงในเอกสารทั้งหมดไว้ในแถวลำดับแพ็ตเพื่อลดขั้นตอนการประมวลผลในช่วงค้นคืน การค้นคืนจะเปรียบเทียบข้อสอบถามที่ผู้ใช้ป้อนกับคำที่ได้จากซีสตริงซึ่งเป็นคำที่ถูกต้องตาม

หลักภาษาศาสตร์ สำหรับส่วนของการจัดลำดับผลการค้นคืนนั้น เมื่อได้ผลลัพธ์การค้นคืนจะนำผลลัพธ์นั้นมาคำนวณหาค่าตามสูตรคำนวณน้ำหนักคำ เพื่อให้ได้ค่าน้ำหนักคำรวมของแต่ละเอกสาร แล้วนำผลน้ำหนักคำที่ได้มาทำการจัดลำดับตามค่าน้ำหนักคำ และส่วนของการค้นคืนย้อนกลับจะนำเอกสารที่ผู้ใช้แสดงว่าเอกสารนั้นตรงตามความต้องการมาสร้างคำใหม่ เพื่อให้ผู้ใช้นำคำใหม่นี้ไปใช้ค้นคืนซ้ำอีกครั้ง เพื่อให้ผลการค้นคืนใหม่มีค่าถูกต้องสูงขึ้นกว่าเดิม ในการวิจัยนี้ได้เลือกสูตรคำนวณน้ำหนักคำมาทั้งหมด 5 สูตร พบว่ามี 2 สูตรที่ให้ผลเฉลี่ยค่าความถูกต้องสูงที่สุด คือ สูตรคำนวณค่าน้ำหนักคำที่ประกอบด้วยค่าความถี่ของคำที่ปรากฏในเอกสาร และสูตรคำนวณค่าน้ำหนักคำที่ประกอบไปด้วยค่าความถี่ของคำที่ปรากฏในเอกสารคูณกับค่าความถี่เอกสารแบบผกผัน ส่วนผลการทดลองการค้นคืนแบบค้นคืนย้อนกลับ พบว่าการเลือกใช้คำที่มีค่าความถี่คำอยู่ในช่วงขีดจำกัดที่เหมาะสม ช่วยให้ระบบเสนอข้อสอบถามใหม่ที่ช่วยให้ผลการค้นคืนมีผลเฉลี่ยค่าความถูกต้องสูงขึ้นกว่าเดิมได้