บทที่ 2

วรรณกรรมและทฤษฎีที่เกี่ยวข้อง

1. ความเบื้องต้น

งานวิทยานิพนธ์นี้ ผู้วิจัยได้มีการนำแนวคิดทางค้นหาใจความสำคัญ และการกำหนดน้ำหนักในรูปแบบของคำหรือวลีที่อ้างอิงกับหลักไวยากรณ์ภาษาไทย เพื่อการสกัดใจความสำคัญ (คำ,วลี) ที่สนับสนุนการค้นคืนสารสนเทศมาประยุกต์ใช้กับภาษาธรรมชาติที่เป็นภาษาไทย ไม่ว่าจะเป็นเรื่องของโครงสร้างของภาษา ความหมายของคำ การแยกคำในประโยค และการจบของประโยค โดยงานวิทยานิพนธ์นี้เป็นประโยชน์สำหรับประยุกต์ใช้กับงานต่าง ๆ ด้านการค้นคืนเอกสารที่มีลักษณะใกล้เคียงกัน

2. ทฤษฎีที่มีความเกี่ยวข้องกับงานวิจัย

งานวิจัย " การสกัดวลีสำคัญภาษาไทยด้วยโครงข่ายประสาทเทียม " ประกอบด้วยงานวิจัยที่เกี่ยวข้องดังต่อไปนี้

1. มโนทัศน์เรื่องคำและหลักไวยากรณ์ในภาษาไทย
2. ระบบการค้นคืนสารสนเทศ
3. การสกัดสารสนเทศ
4. การตัดคำไทยในระบบแปลภาษา
5. การจัดแบ่งหมวดคำในภาษาไทย

3. ทฤษฎีเกี่ยวข้องกับคำและวลีสำคัญภาษาไทย

ระบบการค้นหาหรือค้นคืนสารสนเทศภาษาไทยในปัจจุบันส่วนใหญ่จะทำการค้นคืนเอกสารที่ประกอบด้วยคำหรือวลีในเอกสารที่ตรงกับคำหรือวลีค้น ในระบบการค้นหาโดยที่บางครั้งนั้นผลการค้นหาเอกสารที่ได้จากระบบการค้นหาแบบเดิมนั้นอาจได้เอกสารที่ไม่ตรงกับความต้องการของผู้ใช้ โดยอาจมาจากการแทนเอกสารหรือการทำดรรชนีเอกสารด้วยคำหรือวลี

สำคัญที่ไม่ดีพอ จึงทำให้ได้ผลการค้นหามากเกินความจำเป็น ด้วยเหตุผลดังที่ได้กล่าวมานี้ เพื่อเพิ่มประสิทธิภาพของการค้นหาในด้านของการแทนเอกสารด้วยคำหรือวลีที่เหมาะสมนั้น ในส่วนนี้จึงเป็นการนำเสนอทฤษฎีของภาษาไทยที่เป็นการนิยามถึงไวยากรณ์ทางภาศาสตร์ และเพื่อเป็นแนวทางในการค้นหาใจความสำคัญของเอกสารภาษาไทยด้วย

3.1 วากยสัมพันธ์

ดังที่กล่าวใน (พระยาอุปกิตศิลปสาร ,2514) ตามตำราหลักภาษาไทยของพระยาอุปกิตศิลปสาร วลีและประโยคจัดอยู่ในส่วนที่เรียกว่า **วากยสัมพันธ์** ซึ่งเป็นไวยากรณ์ในส่วนที่กล่าวถึงความเกี่ยวข้องสัมพันธ์กันของคำพูดต่าง ๆ แล้วทำให้เกิดข้อความขึ้นมา ข้อความที่เกิดขึ้นนี้มีอยู่ 2 ลักษณะคือ เป็นเนื้อความที่เกิดจากกลุ่มคำที่มีความหมายยังไม่สมบูรณ์ครบถ้วนเรียกว่า วลี และเกิดจากกลุ่มคำที่มีความหมายสมบูรณ์ครบถ้วนเรียกว่า ประโยค

ประโยค ในภาษาบาลีแปลได้ความว่า ความเกี่ยวข้องของคำต่าง ๆ ในภาษาไทย เพราะข้อความที่เราใช้เขียนและใช้พูดนั้น ย่อมต้องเอาคำต่าง ๆ มาเรียงติดต่อกันไปจนได้ความอย่างหนึ่ง ๆ และคำต่าง ๆ ที่นำมาเรียงเป็นข้อความนั้น ล้วนมีความสัมพันธ์คือมีความเกี่ยวข้องกัน ความเกี่ยวข้องของคำต่าง ๆ ในตอนหนึ่ง ๆ นั้นเรียกว่า **วากยสัมพันธ์**

3.1.1 ถ้อยคำหรือข้อความ จัดไว้เป็น3 อย่าง คือ

3.1.1.1 คำ คือ ผู้พูดหรือผู้เขียนต้องการให้รู้เฉพาะเป็นคำ ๆ

3.1.1.2 วลี คือ เป็นการพูดหรือเขียนหลาย ๆ คำติดต่อกัน แต่ยังไม่ได้ความครบ เป็นการบอกความเพียงกลุ่มเดียวตอนเดียวเท่านั้น จึงนับว่าเป็นวลี คือยังไม่มีเนื้อความครบถ้วน

วลี คือ คำที่ติดต่อกันตั้งแต่ 2 คำขึ้นไป ซึ่งมีความหมายติดต่อเป็นเรื่องเดียวกัน แต่เป็นเพียงส่วนหนึ่ง ๆ ของประโยค และไม่มีเนื้อความครบถ้วนเป็นประโยค

3.1.1.3 ประโยค คือ กล่าวหรือเขียนข้อความครบถ้วน แบ่งออกเป็นสองภาคดังนี้

- ภาคประธาน หมายถึง ส่วนที่ผู้เขียนอ้างขึ้นก่อน เพื่อให้ผู้ผู้อ่านรู้ว่าอะไรเป็นข้อสำคัญของข้อความ ภาคประธานนี้โดยมากมักเป็นคำนามหรือสรรพนามเป็นส่วนใหญ่

- ภาคแสดง หมายถึง คำที่แสดงอาการของภาคประธานให้ได้ความครบว่าแสดงอาการอย่างนั้นอย่างนี้

3.2 คำ

คำ หมายถึง เสียงพูดหรือสายอักขระที่มีความหมายในภาษา โดย Leonard Bloomfield (1933:178) ให้คำจำกัดความว่า คำ หมายถึง หน่วยที่เล็กที่สุดที่สามารถปรากฏตามลำพังได้ (minimum free from) ส่วน R.H. Robins (1964 :185) ได้อธิบายว่า สิ่งที่เป็นคำจะมีความคงตัว (stability) ไม่สามารถจะแยกย่อยให้เล็กลงไปได้อีก และจะจัดลำดับส่วนที่อยู่ในคำเสียใหม่ก็ไม่ได้ ส่วนตำราไวยากรณ์ไทยทั้งหลาย พระยาอุปกิตศิลปสาร (2514: 59) กล่าวไว้ว่า คำ หมายถึง กลุ่มของหน่วยเสียงหรือกลุ่มของตัวอักษรที่มีความหมายในภาษา คำจำกัดความเหล่านี้มีความแตกต่างกันในด้านมุมมองในการตัดสินคำ

3.3 วลีไทย

นอกจากการใช้ตำแหน่งเรียงหน้าหลังเพื่อบอกชนิด หน้าที่ ฯลฯ ของคำ ภาษาไทยยังมีการนำเอาคำอื่นมาประกอบกันเพื่อบอกชนิดและหน้าที่ของคำ ดังนั้นภาษาไทยเราจึงเป็นกลุ่ม ๆ ที่เรียกว่า “ **วลี** ” โดยมาก

3.3.1 ชื่อของวลี

เมื่อภาษาไทยนั้นใช้วลีเป็นส่วนใหญ่ดังที่ได้กล่าวมานั้น เราจึงจำเป็นจะต้องรู้ประเภทวลีที่ใช้อยู่ในภาษาไทย งานวิจัยนี้จึงเลือกที่จะพิจารณากระบวนการสกัดใจความสำคัญทั้งในลักษณะของคำและวลีที่มีความสำคัญในแต่ละเอกสาร ชนิดของคำนั้นได้กล่าวมาแล้วในตอนต้น ในหัวข้อนี้ผู้วิจัยจะกล่าวถึงประเภทต่าง ๆ ของวลี ซึ่งมีด้วยกันดังนี้

3.3.1.1 นามวลี เช่น ไก้แจ้ , นกเขา , เมืองตรัง , รูปเดียว , ชาวนา , ข้าวสาลี

3.3.1.2 สรพนามวลี คือ คำที่ผู้เขียนแต่งขึ้นเพื่อแทนสรรพนามที่ใช้กัน เช่น เราทุกคน ท่านคณะกรรมการสภาประจำสถาบันราชภัฏ-กาญจนบุรี ข้าเบื้องยุคลบาท

3.3.1.3 กริยาวลี ที่เป็นภาคแสดง เพื่อแสดง มาลา กาล และวาจก ของภาคแสดง เช่น คงกิน , ต้องกิน , กำลังกิน , ให้ตี ฯลฯ

3.3.1.4 วิเศษณ์วลี วลีที่มีคำวิเศษณ์นำหน้า ซึ่งมีหน้าที่ประกอบคำอื่นอย่างเดียวกับคำวิเศษณ์ธรรมดา จึงใช้เป็นภาคขยายของประโยคได้อย่างเดียวกับคำวิเศษณ์โดยทั่วไป เช่น โง่บัดซบ , ดีเหลือเกิน , สายสามนาที ฯลฯ

3.3.1.5 บุพบทวลี ได้แก่วลีที่มีคำบุพบทอยู่ข้างหน้า เช่น ด้วย-ความรู้ , จาก-ที่อยู่ , ซึ่ง-กันและกัน , ใน-เราทั้งหลาย ฯลฯ

3.3.1.6 สันธานวลี หมายถึง คำสันธานที่คำประกอบให้ยืดยาวออกไป แต่ไม่ใช่คำประสม จะเป็นคำเชื่อมกับสันธานด้วยกันก็ได้ เชื่อมกับคำอื่นหรือวลีก็ได้เพื่อใช้ในหน้าที่สันธานเช่นเดิม

3.3.1.7 อุทานวลี หมายถึงคำอุทานที่มีคำอื่นประกอบท้ายให้เป็นวลียืดยาวออกไป หรือเอาบทวลีต่าง ๆ มาเป็นอุทมาน เช่น ลูกรักเจ้าแม่เอ๋ย ฯลฯ

3.3.2 หน้าที่ของวลี

ภาษาไทยนั้นนิยมใช้วลีแทนคำเป็นพื้น ดังนั้นคำมีหน้าที่อย่างไร วลีก็มีหน้าที่อย่างเดียวกับคำเหมือนกัน

3.4 ประโยคภาษาไทย

ภาคของประโยค แปลว่าส่วนของประโยค ซึ่งแบ่งได้ 2 ภาค คือ ภาคประธาน และภาคแสดง ดังแสดงในตารางที่ 2-1

ภาคประธาน		ภาคแสดง			
บทประธาน	บทขยายประธาน	บทกริยา	บทขยายกริยา	บทกรรม	บทขยายกรรม
สามี	ดี	ย่อมนับถือ	เป็นอย่างดี	ซึ่งภรรยา	ดี

ตารางที่ 2-1 ตัวอย่างประโยคพื้นฐาน (อุปกิตศิลปสาร ,2514)

รูปประโยคภาษาไทยนั้น บางครั้งอาจไม่ได้เรียงลำดับส่วนของประโยคดังที่ได้แสดงไว้ดังตาราง 2-1 เสมอไป ข้อสำคัญของไวยากรณ์ไทยอยู่ที่การเรียงลำดับคำพูดหรือเขียนเป็นหลักวินิจฉัยชนิดของคำ และหน้าที่ต่าง ๆ ของคำในการบอกสัมพันธ์ของประโยคด้วย โดยประโยคในภาษาไทยแบ่งได้เป็น 4 ประเภท ดังนี้

3.4.1 ชนิดของประโยค

3.4.1.1 ประโยคกรรตุ ส่วนสำคัญของประโยคที่จะยกขึ้นพูดมีอยู่ 3 ส่วน คือ กรรตุการก (ผู้ทำ) กริยา และกรรมการก (ผู้ถูกทำ)

3.4.1.2 ประโยคกรรม คือ ประโยคต้องการเน้นผู้ถูกทำเป็นสำคัญ

3.4.1.3 ประโยคกริยา คือ ในข้อความบางแห่ง ผู้เขียนต้องการจะให้กริยาเด่น

3.4.1.4 ประโยคการิต มีลักษณะคล้ายกับประโยคกรรม

3.4.2 โครงสร้างของประโยค

3.4.2.1 บทประธาน **บทประธานนับว่าเป็นส่วนสำคัญของประโยค** ซึ่งผู้เขียนต้องการที่จะให้เด่น จึงมักจะนำมากล่าวถึงก่อนที่ต้นประโยค ดังนั้น**บทประธานของประโยคจึงต้องเป็นคำนาม หรือสรรพนาม**

3.4.2.2 บทขยายประธาน คือ บทที่ใช้ประกอบประธานให้ได้ความชัดเจนยิ่งขึ้น มีได้ทั้ง คำ วลี และประโยค เช่น คำวิเศษณ์ , บทวิกัติการก , บทวลี

3.4.2.3 บทกริยา เป็นส่วนสำคัญของภาคแสดง ที่บอกอาการของบทประธานให้ได้ความครบในประโยค เช่น อกรรมกริยา , กริยาวลี

3.4.2.4 บทขยายกริยา หมายถึง บทที่ทำหน้าทีแต่งบทกริยาให้มีเนื้อความพิสดารยิ่งขึ้น มีอยู่ 2 ประเภท คือ คำวิเศษณ์ และ วลีต่าง ๆ ที่ทำหน้าที่วิเศษณ์

3.4.2.5 บทกรรม หมายความว่า บทที่ทำหน้าที่ผู้ถูกทำ โดยในประโยคนั้นจะต้องมีบทกรรมต่อท้ายจึงจะมีเนื้อความครบบริบูรณ์ บทกรรมนับว่าเป็นบทสำคัญคล้ายคลึงกับบทประธานเหมือนกัน โดยบทกรรมมีลักษณะคล้ายคลึงกับบทประธานทุกอย่าง ต่างกันที่หน้าที่ของกรรมนั้นเป็นผู้กระทำ

3.4.2.6 บทขยายกรรม ทำหน้าที่ในการขยายบทกรรม

3.5 คำหรือวลีสำคัญ (Keyword & KeyPhrase)

การกำหนดหน่วยของคำเกิดขึ้นจากการบัญญัติ หรือนิยาม หรือตั้งคำศัพท์ใหม่ การนิยามหรือบัญญัติคำศัพท์ใหม่ก็เพื่อใช้ในการสื่อสารในเรื่องเดียวกัน ดังนั้นคำศัพท์ใหม่อาจใช้เรียก รูปธรรม หรือ นามธรรม ซึ่งใช้วลีในการแทน รูปธรรม หรือ นามธรรม

สำหรับภาษาไทยนั้น ผู้เชี่ยวชาญต่าง ๆ มีความเห็นว่าจำเป็นต้องเป็นคำที่สมบูรณ์ในแยกคำและดีพอ เพราะถ้าแยกคำผิดความหมายก็จะผิดแปลกไปได้ ทำให้มีผลต่อศาสตร์ของการค้นคืนสารสนเทศอย่างมากในการที่จะหาคำหรือวลีสำคัญที่เป็นตัวแทนเอกสารที่ดีได้

การพิจารณาค่าความสำคัญหรือน้ำหนักของวลีสำคัญที่อ้างอิงกับหลักไวยากรณ์ภาษาไทยที่มีความเหมาะสม ควรจะต้องพิจารณาเลือกคำหรือวลีสำคัญ คือ

- กลุ่มที่ 1 เป็นกลุ่มคำหรือวลีที่มีนัยสำคัญที่สอดคล้องกับหลักไวยากรณ์ภาษาไทย ดังที่จะกล่าวในหัวข้อต่อ ๆ ไป

- กลุ่มที่ 2 จะเป็นคำที่มีการใช้มากเป็นพิเศษและไม่มีประโยชน์ในการสืบค้นเลย เช่น คำว่า และ ที่ ต้อง การ ซึ่ง ฯลฯ คำเหล่านี้เกิดขึ้นเพราะหลักไวยากรณ์ในภาษาไทยนั่นเอง ซึ่งจะเห็นว่าคำกลุ่มนี้ไม่มีคุณสมบัติเป็นคำสืบค้นได้เลยและสามารถตัดทิ้งออกจากดรรชนีคำค้นได้

- กลุ่มที่ 3 เป็นกลุ่มคำที่ประมาณการใช้โดยเฉลี่ยไม่มากหรือน้อยจนเกินไป กลุ่มดังกล่าวเป็นกลุ่มที่เหมาะที่จะใช้เป็นดรรชนีคำค้น

3.6 ลักษณะไวยากรณ์ภาษาศาสตร์ (คำและวลีสำคัญในภาษาไทย)

สำหรับในงานวิจัยนี้ผู้วิจัยต้องการนำเสนอแนวคิดในการหาใจความสำคัญของเอกสารโดยอ้างอิงกับหลักไวยากรณ์ภาษาศาสตร์ในรูปแบบของกฎฮิวริสติก ซึ่งมีหลักต่าง ๆ ด้วยกันดังนี้

3.6.1 หลักการใช้คำให้มีน้ำหนัก

หลักการใช้ความถี่ของคำบ่งบอกถึงความสำคัญของคำหรือวลีนั้น โดย ผะอบ โปษะกฤษณะ (2540) กล่าวว่า ดังนี้

3.6.1.1 ใช้คำซ้ำเพื่อเน้นหนัก

3.6.1.2 ไม่ย่อคำจนเกินไป

3.6.1.3 เว้นวรรคตอนให้ถูกที่

3.6.2 หลักการของประโยค

ในการใช้คำที่บ่งบอกถึงความสำคัญของคำหรือวลีในตำแหน่งของประโยค นั้น ดังที่กล่าวใน (ผะอบ โปษะกฤษณะ , 1997) มีหลักการในการดังนี้

3.6.2.1 การเข้าประโยค

คำต่าง ๆ ตามปกติ ถ้าอยู่แต่ลำพังไม่สู้มีน้ำหนัก ต่อเมื่อนำมาเรียงเข้าประโยคจึงจะเสดงหน้าที่ แสดงความหมายชัดเจนขึ้น แสดงน้ำหนักดียิ่งขึ้น การเรียงคำเป็นเรื่องสำคัญในภาษาไทย ฉะนั้น การเรียงคำเข้าประโยคจึงต้องมีความเหมาะสม

ก. เข้าประโยคให้ได้ความชัดเจน

ข. การเข้าประโยคให้มีน้ำหนัก

ค. การเข้าประโยคต้องไม่ผิดกับหลักภาษา

3.6.2.2 การเขียนประโยคให้มีน้ำหนัก

มีความสำคัญมากในการเขียนทุกระดับ เพราะถ้าประโยคที่เขียนขาดน้ำหนักจะทำให้ใจความสำคัญขาดความหนักแน่น ผู้อ่านอาจไม่เชื่อถือหรือคล้อยตาม กล่าวได้ว่าการเขียนประโยคให้มีน้ำหนักเป็นการสร้างความประทับใจให้แก่ผู้อ่านนั่นเอง ดังนี้

3.6.2.2.1 การจัดตำแหน่งของคำในประโยคให้ตรงกับฐานน้ำหนัก

ฐานน้ำหนักในประโยค แบ่งออกเป็น 3 ตอน คือ **ตอนต้นประโยค กลางประโยค และปลายประโยค** ทั้งสามตอนนี้ปลายประโยคมีน้ำหนักมากที่สุด ต้นประโยคมีน้ำหนักรองลงมา และกลางประโยคมีน้ำหนักเบาที่สุด (เปลื้อง ณ นคร, 2507) ดังนั้นเมื่อต้องการเน้นส่วนใดให้มีน้ำหนัก ควรวางส่วนนั้นที่ปลายประโยคหรือต้นประโยค และการวางฐานน้ำหนักของประโยคดังที่กล่าวใน (อุปกิตศิลปสาร , 2514) โดยเป็นหลักการเขียนประโยคที่ตรงกับหลักจิตวิทยาประการหนึ่ง คือ “ การวางคำที่สำคัญหรือคำที่มีน้ำหนักมากที่สุดไว้ส่วนต้นและท้ายของประโยค ”

1. การใช้คำที่มีน้ำหนักเน้น
2. การเน้นโดยใช้คำกระชับหรือสรุปความ
3. การเน้นโดยการใช้คำสันธานที่เป็นคู่กัน
4. การเน้นโดยการซ้ำคำ

3.6.2.2.2 การตกแต่งประโยค

จากหลักการเขียนประโยค อุปกิตศิลปสาร (2514) มีการใช้คำสรุปความก่อนเน้นคำสำคัญในตอนท้ายประโยค คือ การใช้คำรวมข้อความต่าง ๆ ตอนต้นประโยคเข้าด้วยกัน โดยคำเหล่านั้นอาจเรียกว่า “ คำสรุปความ ” ในแต่ละประโยคเพื่อเป็นการสรุปใจความส่วนสำคัญของประโยคนั้น ๆ

3.6.3 หลักการของย่อหน้า

ในการใช้คำที่บ่งบอกถึงความสำคัญของคำหรือวลีในตำแหน่งของในย่อหน้า นั้นดังที่กล่าวใน (ผะอบ โปษะกฤษณะ , 2540) มีหลักการ ดังนี้

3.6.3.1 คำจำกัดความของย่อหน้า หรืออนุเฉท

ย่อหน้าหรืออนุเฉท ที่อธิบายโดย คณาจารย์ภาควิชาภาษาไทย คณะศิลป-ศาสตร์ มหาวิทยาลัยธรรมศาสตร์ (2546) หมายถึง ข้อความตอนหนึ่งซึ่งมีใจความสำคัญเพียงเรื่องเดียวและประโยคขยายใจความสำคัญดังกล่าวให้ได้ความชัดเจนและสมบูรณ์ อาจกล่าวได้ว่า ย่อหน้า คือความเรียงสั้น ๆ เรื่องหนึ่ง และความหมายของย่อหน้า ที่กล่าวโดย ผะอบ โปษะกฤษณะ (2540) ก็คือ ความเรียงเรื่องสั้น ๆ เรื่องหนึ่ง ซึ่งรวมประโยคหลาย ๆ ประโยคที่มีข้อความมุ่งสู่จุดเดียว การเขียนหนังสือที่จะให้ผู้อ่านติดตามได้ง่ายและให้เข้าใจได้รวดเร็ว นิยมแบ่งเป็นย่อหน้าเพื่อแสดงให้ทราบว่าหมดข้อความตอนหนึ่ง การแบ่งย่อหน้าจะช่วยให้ผู้อ่านจับความคิดได้ง่ายและยังทำให้เกิดความสวยงาม ย่อหน้าทุกย่อหน้าต้องมีความสัมพันธ์กันเป็นอย่างดี ลำดับความนึกคิดให้เป็นไปตามลำดับ

3.6.3.2 การเข้าความของย่อหน้า

เทคนิคการเข้าความย่อหน้านั้น ผะอบ โปษะกฤษณะ (2540) โดยเมื่อต้องการจะเขียนข้อความใด ๆ ก็ดี ก็ต้องเรียงประโยคให้เป็นข้อความที่ดี จะต้องให้ผู้อ่านเข้าใจแจ่มชัดและติดตามตลอดเวลา ฉะนั้นในการที่จะให้ผู้อ่านเข้าใจความคิดของผู้เขียนก็ควรแบ่งเป็นตอน ๆ เพื่อให้สามารถสรุปความคิดแต่ละตอน นั่นก็คือการแบ่งความเป็นย่อหน้า

3.6.3.3 ชนิดของย่อหน้า

ในการแบ่งชนิดของย่อหน้านั้น ปรีชา ช้างขวัญยืน (2525) ได้แบ่งไว้ดังนี้ ย่อหน้าแบ่งเป็น 4 ชนิดคือ ย่อหน้านำความคิด ย่อหน้าโยงความคิด ย่อหน้าแสดงความคิด และย่อหน้าสรุปความคิด ย่อหน้าแสดงความคิดเป็นย่อหน้าที่ใช้แสดงความคิดทั้งหมดของเรื่อง ย่อหน้าชนิดอื่นเป็นส่วนประกอบให้เรื่องสมบูรณ์ ดังนี้

3.6.3.3.1 ย่อหน้านำความคิด

ย่อหน้านำความคิดหรือที่ส่วนใหญ่เรียกว่า “ คำนำ ” หรือส่วนนำของเรื่อง เป็นย่อหน้าที่ใช้นำก่อนเข้าถึงความคิดสำคัญของเรื่อง ส่วนใหญ่มักจะระบุจุดประสงค์ของเรื่องที่เขียนหรือบอกประเด็นหลักของเรื่องทั้งหมด

3.6.3.3.2 ย่อหน้าโยงความคิด

ย่อหน้าโยงความคิดเป็นย่อหน้าที่ใช้เชื่อมโยงความคิดระหว่างย่อหน้า เพื่อให้เกิดความสัมพันธ์ต่อเนื่องกัน ส่วนใหญ่มักใช้ในกรณีที่ความคิดสำคัญหรือใจความสำคัญในย่อหน้าที่ผ่านมากับความคิดสำคัญในย่อหน้าที่กำลังจะเขียนขึ้นใหม่เป็นคนละประเด็นไม่

เกี่ยวข้องกัน ย่อหน้าโยงความคิดนี้จะเป็นตัวเชื่อมโยงความคิดสำคัญที่ไม่เกี่ยวข้องกันนั้นให้เข้ามาสัมพันธ์กันได้ ย่อหน้าชนิดนี้มักเป็นย่อหน้าสั้น ๆ

3.6.3.3.3 ย่อหน้าแสดงความคิด

หมายถึงย่อหน้าที่ต้องการสื่อถึงการแสดงความคิดเห็นของผู้เขียน

3.6.3.3.4 ย่อหน้าสรุปความคิด

ย่อหน้าสรุปความคิดหรือย่อหน้าสรุปนั่นเอง ย่อหน้าชนิดนี้ใช้เป็นย่อหน้าที่สรุปเรื่องทั้งหมด หรือสรุปความคิดสำคัญเสนอผู้อ่าน การสรุปมีกลวิธีที่จะนำมาใช้ได้หลายวิธีแล้วแต่ลักษณะของเรื่อง

3.6.3.4 ย่อหน้าที่ดี และลักษณะงานเขียนย่อหน้าที่ดี

การเขียนย่อหน้าที่ดี และจะสามารถบอกถึงใจความสำคัญของย่อหน้านั้น ๆ ได้ ดังที่กล่าวใน คณาจารย์ภาควิชาภาษาไทย คณะศิลปศาสตร์ มหาวิทยาลัยธรรมศาสตร์ (2546) และ จรูญ ตันสูงเนิน (2532) ลักษณะของการเขียนย่อหน้านั้น ควรจะเขียนดังนี้

3.6.3.4.1 มีเอกภาพ

คือ มีข้อความสำคัญเพียงประการเดียว สาระหลักเป็นจุดเดียว ส่วนจะมีขยายหรือแสดงตัวอย่างเพิ่มเติมก็ทำได้ การเขียนไม่วกวน แต่ละย่อหน้าควรรวมจุดเป็นหัวข้อหรือแนวคิดตามที่ได้ตั้งใจไว้ในการวางโครงเรื่องทำให้ผู้อ่านสามารถจับประเด็นเรื่องย่อข้อความทั้งหมดให้เหลือเพียงประโยคเดียว

3.6.3.4.2 มีความสมบูรณ์

คือ การเขียนให้ในแต่ละย่อหน้านั้น มีใจความครบถ้วนสมบูรณ์ที่ผู้อ่านสามารถเข้าได้ ว่าแต่ละย่อหน้านั้นต้องการสื่อสารเรื่องใดกับผู้อ่าน

3.6.3.4.3 มีสัมพันธภาพ

คือ มีความสัมพันธ์ต่อเนื่องกันดี เรียงลำดับประโยคให้ขยายโครงร่างรายละเอียดแสดงความคิดอย่างติดต่อกัน

ก. มีสารัตถภาพ การมีสารัตถภาพ มีหลักการดังนี้

1) การย้ำเน้นในที่ที่ควรเน้น

วิธีที่จะทำให้ย่อหน้ามีสารัตถภาพโดยการเน้นในที่ที่ควรเน้นนั้นเป็นวิธีที่ง่ายที่สุด โดย**การวางตำแหน่งประโยคใจความสำคัญในตอนต้นหรือตอนท้ายของ**

ย่อหน้า ทั้งนี้เพื่อแสดงความคิดสำคัญของผู้เขียนให้ผู้อ่านได้เห็นเด่นชัด และผู้อ่านสามารถที่จะจับความคิดสำคัญดังกล่าวได้ง่าย ตัวอย่างที่แสดงการย้ำเน้นในที่ที่ควรย้ำเน้น

2) มีการเน้นย้ำสาระที่สำคัญตามแนวคิดหลักที่สร้างสรรค์ขึ้น มีคำสำคัญและประโยคที่กระชับความหมายอย่างเด่นชัด ถ้าเป็นย่อหน้าที่มีขนาดสั้นเพียงย่อหน้าเดียวอาจสังเกตได้ว่ามักจะ**วางประโยคใจความไว้ตอนต้นย่อหน้าหรือบางทีก็เอาไว้ท้ายย่อหน้า**เพื่อสรุปความสำคัญทั้งหมดในย่อหน้านั้น

3) การมีสารัตถภาพ หมายถึง การเน้นให้เห็นความสำคัญของเรื่องอย่างชัดเจน งานเขียนที่มีสารัตถภาพ จะต้องแสดงให้ผู้อ่านเข้าใจสาระสำคัญของเรื่องอย่างชัดแจ้ง การอธิบายขยายความต้องไม่ยืดยาวเกินความจำเป็น เพราะจะทำให้ผู้อ่านสับสนหรือหลงลืมสาระสำคัญ ปรีชา ช้างขวัญยืน (2525 : 170) กล่าวว่า การย้ำความให้เห็นว่าเรื่องนั้น ๆ สำคัญกว่าเรื่องอื่น ๆ สามารถทำได้ 2 วิธี วิธีแรกคือ **การวางตำแหน่งประโยคสาระสำคัญไว้ต้นย่อหน้าหรือท้ายย่อหน้า** วิธีที่สองคือ อธิบายความคิดสำคัญให้ละเอียดมากกว่าความคิดอื่น ใช้เนื้อที่ของย่อหน้ามากกว่าความคิดอื่น

4) วิธีที่จะทำให้ย่อหน้ามีสารัตถภาพโดย**การย้ำเน้นด้วยคำวลีหรือประโยคซ้ำ ๆ กันบ่อย ๆ** ภายในย่อหน้า วิธีนี้ทำให้ผู้อ่านเกิดความสนใจคำวลีและประโยคที่พบเห็นซ้ำกันบ่อย ๆ และทำให้ผู้อ่านเข้าใจจุดมุ่งหมายหรือความคิดสำคัญที่ผู้เขียนต้องการสื่อสารถึงผู้อ่าน แต่การย้ำเน้นดังกล่าวนี้ต้องทำแต่พอดี ถ้ามีมากไปจะทำให้ผู้อ่านสะดุดความคิดดังกล่าวและเข้าใจเจตนาของผู้เขียนได้แจ่มชัดขึ้น

การเขียนย่อหน้าที่ดีนั้นตามหลักของ ผะอบ โปษะกฤษณะ (2540) กล่าวได้ว่าย่อหน้าที่ดีนั้น ควรเป็นงานเขียนที่มีลักษณะดังนี้

(1) มีข้อความเป็นเรื่อง ย่อหน้าหนึ่งจะกล่าวถึงความสำคัญแต่เพียงเรื่องเดียว ส่วนจะมีข้อความหรือแสดงตัวอย่างประกอบก็ได้

(2) ต้องมีความสัมพันธ์กันดี ประโยคต่าง ๆ แต่ละประโยคต้องเชื่อมกันให้ดี ให้ผู้อ่านจับความคิดได้ทันที ไม่สับสน การใช้คำสันธานเชื่อมประโยคใช้ให้ถูกการอ้างอิงหรือตัวอย่างวางให้ถูกที่ และต้องแน่ใจว่าตัวอย่างนั้นจะขยายข้อความให้ชัดเจน

(3) **การเน้นหนัก** จุดมุ่งหมายอันสำคัญของย่อหน้านั้นควรจะให้เด่นชัดเพื่อเน้นความสำคัญให้รู้ว่าเราพูดถึงอะไร ประโยคสำคัญเรียกว่า “ประโยคใจความ” หรือ “ประโยคกุญแจ” ผู้เขียนส่วนมากมักจะ**วางประโยคใจความไว้ข้างต้น**เพื่อให้เห็นเด่นชัดแล้วจึงเขียนขยายเรื่องนั้น แต่บางครั้งก็อาจเอาไว้**ท้ายย่อหน้าเพื่อกระชับความ**แต่บางทีเอาไว้ท้ายย่อหน้าเพื่อกระชับความ แต่บางทีไว้ทั้งข้างต้นและข้างท้าย

(4) ความแตกต่าง ประโยคที่เขียนควรจะมีความยาวสั้นต่างกัน การสร้างรูปประโยคแตกต่างกันไปประกอบจะทำให้ชวนอ่าน ถ้ามีแต่ประโยคสั้น ๆ ไปหมดหรือประโยคยาว ๆ ทั้งย่อหน้า วิธีเขียนประโยคแบบเดียวจะทำให้มองดูไม่งาม

3.6.3.5 วิธีสร้างย่อหน้า

วิธีการสร้างย่อหน้าที่ดีที่สุด คือ การกำหนดจุดมุ่งหมายหรือความคิดสำคัญของเราเป็นประโยคใจความสำคัญ แล้วหารายละเอียดต่าง ๆ มาเขียนขยายให้ได้เนื้อหาที่มีความสมบูรณ์ โดยจะต้องรู้จักใจความสำคัญไว้ในตำแหน่งต่าง ๆ โดยสำหรับ จรูญ ตันสูงเนิน (2532) ได้อธิบายไว้ว่า ประโยคใจความสำคัญ ได้แก่ ประโยคที่แสดงความคิดหลัก (Main Idea) ที่ผู้เขียนต้องการแสดงในย่อหน้านั้น หรือกล่าวอีกอย่างหนึ่งว่า ประโยคใจความสำคัญ คือ ประโยคที่สรุปสาระของย่อหน้าได้ ดังนี้

3.6.3.5.1 ตำแหน่งของประโยคใจความสำคัญ

(1) **ตอนต้นย่อหน้า**

การวางประโยคใจความสำคัญในตำแหน่งต้นย่อหน้านั้นเป็นวิธีเขียนย่อหน้าที่ง่ายสะดวกที่สุดและทำให้ผู้อ่านเข้าใจง่ายที่สุด ทั้งยังเป็นวิธีสำคัญที่ช่วยทำให้ย่อหน้ามีเอกภาพด้วย ย่อหน้าที่วางประโยคใจความสำคัญไว้ตอนต้นของย่อหน้า โดยอาจเรียกว่าย่อหน้าแบบตัว T และจรูญ ตันสูงเนิน (2532) กล่าวถึงเทคนิคการวางใจความสำคัญไว้ คือ **การวางใจความสำคัญที่ตอนต้นย่อหน้า** คือ การเขียนประโยคใจความสำคัญไว้ตอนต้นย่อหน้าเป็นวิธีที่นิยมกันมากที่สุด ทำให้การเขียนย่อหน้าง่ายที่สุด และผู้อ่านก็เข้าใจเนื้อความของย่อหน้าได้

(2) **ตอนท้ายย่อหน้า**

การวางประโยคใจความสำคัญไว้ตอนท้ายย่อหน้าเป็นการสรุปข้อความที่ได้เขียนอธิบายไว้ข้างต้น เพื่อให้เกิดความสมบูรณ์ได้เนื้อหาสาระ นับเป็นวิธีที่ย้ำเน้นความคิดสำคัญตอนท้าย เพื่อให้ผู้อ่านจับใจความสำคัญได้ง่าย การเขียนขยายประโยคใจความสำคัญวิธีนี้ผู้เขียนอาจใช้วิธีลำดับความคิดแบบแสดงผลแล้วจึงนำไปสู่เหตุ หรือยกตัวอย่างหลาย ๆ ตัวอย่างแล้วจึงจบความด้วยการสรุป หรือจะยกข้อความเปรียบเทียบให้เข้าใจแล้วจึงสรุปด้วยใจความสำคัญในตอนท้ายของย่อหน้าได้ และ จรูญ ตันสูงเนิน (2532) ยังมีการกล่าวถึงเทคนิคการวางใจความสำคัญอีกรูปแบบ คือ **การวางใจความสำคัญที่ตอนท้ายย่อหน้า** คือ ย่อหน้าที่มีความสำคัญท้ายย่อหน้า เปรียบเหมือนการอธิบายขยายความก่อนแล้วจึงให้ข้อสรุป เป็นการเน้นย้ำหรือให้ข้อสรุปในตอนท้าย

(3) **ตอนต้นและตอนท้ายย่อหน้า**

ในการเขียนขยายประโยคใจความสำคัญให้สมบูรณ์เป็นย่อหน้านั้น ในบางครั้งเราอาจจะเริ่มต้นย่อหน้าด้วยประโยคใจความสำคัญแล้วเขียนขยายความให้ได้เนื้อหาสาระจนจบ และลงท้ายด้วยการสรุปด้วยประโยคใจความสำคัญอีกครั้งหนึ่ง วิธีนี้จะทำให้ย่อหน้ามีถ้อยคำที่ย้ำเน้นให้ความคิดสำคัญนั้นชัดเจนยิ่งขึ้น การเขียนขยายประโยคใจความสำคัญลักษณะนี้เหมาะสำหรับย่อหน้าที่มีเนื้อความมาก ย่อหน้ามักจะยาว ย่อหน้าที่เขียนขยายโดยวางประโยคใจความสำคัญไว้ทั้งตอนต้นและตอนท้ายย่อหน้าบางคนเรียกว่า ย่อหน้าแบบตัว I และ จรูญ ตันสูงเนิน (2532) ยังมีการกล่าวถึงเทคนิคการวางใจความสำคัญอีกรูปแบบ คือ **การวางใจความสำคัญที่ตอนต้นและตอนท้ายย่อหน้า** คือ ในบางกรณีผู้เขียนจะเริ่มย่อหน้าด้วยประโยคใจความสำคัญ แล้วขยายความไปจนจบย่อหน้า พร้อมกับลงท้ายสรุปใจความด้วยสำคัญอีกครั้งหนึ่ง เพื่อเป็นการเน้นย้ำให้ชัดเจนยิ่งขึ้น ประโยคใจความสำคัญที่อยู่ตอนท้ายของย่อหน้านี้นิยมเรียกว่า ประโยคสรุป

(4) **ตอนกลางย่อหน้า**

การเขียนขยายประโยคใจความสำคัญให้เป็นย่อหน้าโดยวางประโยคใจความสำคัญไว้ตอนกลางย่อหน้านั้น ทำได้โดยเริ่มเรียบเรียงข้อความที่เป็นส่วนขยายหรือรายละเอียดมาก่อนแล้วจึงตามด้วยประโยคใจความสำคัญ และต่อท้ายด้วยการขยายความเสริมเพื่อให้ย่อหน้านั้นสมบูรณ์ด้วยเนื้อหาสาระ และ จรูญ ตันสูงเนิน (2532) ยังมีการกล่าวถึงเทคนิคการวางใจความสำคัญอีกรูปแบบ คือ **การวางใจความสำคัญที่ตอนกลางย่อหน้า** คือ ย่อหน้าที่มีประโยคใจความสำคัญอยู่กลางย่อหน้า จะเริ่มต้นย่อหน้าด้วยการให้รายละเอียดกว้าง ๆ แล้วสรุปใจความสำคัญไว้กลางย่อหน้า แล้วก็ขยายความเสริมจากกลางย่อหน้าออกไปอีก

เมื่อได้รู้ว่าการเขียนย่อหน้าที่ดีทำได้ง่าย ๆ โดยกำหนดจุดมุ่งหมายเป็นประโยคใจความสำคัญ และรู้จักตำแหน่งต่าง ๆ ของประโยคใจความสำคัญในย่อหน้าแล้ว ต่อไปจะกล่าวถึงวิธีการเขียนขยายประโยคใจความสำคัญให้เป็นย่อหน้าที่ดีซึ่งมีทั้งเอกภาพ ความสมบูรณ์ สัมพันธภาพ และสารัตภถาพ

3.6.3.6 องค์ประกอบของย่อหน้า

องค์ประกอบของย่อหน้านั้น ดังที่กล่าวใน จรูญ ตันสูงเนิน (2532) ได้กล่าวไว้ว่า องค์ประกอบของย่อหน้านั้นควรประกอบไปด้วย ดังนี้

3.6.3.6.1 ประโยคใจความสำคัญ ได้แก่ ประโยคที่แสดงความคิดหลักที่ผู้เขียนต้องการแสดงในย่อหน้านั้น หรือกล่าวอีกอย่างหนึ่งว่า ประโยคใจความสำคัญ คือ ประโยคที่สรุปสาระของย่อหน้าได้ ดังนี้

(1) ตอนต้นย่อหน้า

(2) ตอนท้ายย่อหน้า

(3) ทั้งตอนต้นและตอนท้ายย่อหน้า

(4) ตอนกลางย่อหน้า

3.6.3.6.2 ประโยคขยายความ ได้แก่ ประโยคที่ประกอบขยาย หรือสนับสนุนประโยคใจความาสำคัญให้ละเอียดและเด่นชัดยิ่งขึ้น ตามปกติจำนวนประโยคขยายความมีมากกว่าหนึ่งประโยค

3.6.3.6.3 ประโยคสรุป หมายถึง ประโยคที่รวมใจความทั้งหมดของย่อหน้าไว้ ตำแหน่งของประโยคสรุป คือ ตอนท้ายหรือตอนจบของย่อหน้า

3.6.3.6.4 ประโยคส่งความหรือประโยคเชื่อมความ ได้แก่ ประโยคที่เชื่อมเนื้อหาจากย่อหน้าหนึ่งไปยังย่อหน้าต่อไป ประโยคส่งความจะปรากฏก็ต่อเมื่อมีย่อหน้าตั้งแต่สองย่อหน้าขึ้นไป การมีประโยคส่งความหรือประโยคเชื่อมความระหว่างย่อหน้า จะช่วยให้งานเขียนนั้นมีสัมพันธภาพโดยตลอด

3.6.4 หลักการเอกสารภาษาไทย

3.6.4.1 ชื่อเรื่อง เป็นการบอกถึงชื่อเรื่องของเอกสารนั้น ๆ

3.6.4.2 บทคัดย่อ เป็นการกล่าวถึงรายละเอียดของทั้งเอกสาร ในรูปแบบสรุปโดยคร่าว ๆ

3.6.4.3 คำนำ การเขียนคำนำเป็นการเตือนให้ผู้อ่านทราบว่าเขากำลังจะได้อ่านอะไรต่อไป ดังนั้นคำนำเป็นจุดสำคัญเริ่มแรกที่จะทำให้เกิดความสำเร็จ

3.6.4.4 ความเป็นมา เป็นการกล่าวถึงสาเหตุหรือเหตุผล หรือที่มาของเอกสารนั้น ๆ

3.6.4.5 เนื้อเรื่อง เนื้อเรื่องเป็นสิ่งสำคัญ ในการเขียนต้องระลึกอยู่เสมอว่าเราต้องการให้ผู้อ่านได้รับอะไร เรื่องต้องมีสาระ ผู้เขียนต้องมีความรู้พอที่จะเขียนได้ดี การเขียนจะสมความมุ่งหมายอยู่ที่การจัดเนื้อเรื่องทีได้ประมวลมาทั้งข้อเท็จจริงและเหตุผล ซึ่งเป็นหัวใจสำคัญในการเขียน

3.6.4.6 สรุป การสรุปเป็นการช่วยกระชับให้รู้แน่ชัดลงไปว่าที่เขียนมาทั้งหมดนั้นมีอะไร และอาจจะฝากความคิดนึกไว้ให้ผู้อ่านได้นำไปคิดต่อไป เป็นการขมวดปมหรือฝากความประทับใจไว้กับผู้อ่าน ว่าเรื่องที่เขียนนั้นจะมุ่งหมายไปในแบบใด

3.6.5 หลักการเว้นวรรค

ภาษาไทยใช้วิธีเว้นวรรคเพื่อให้เข้าใจว่าจบความตอนหนึ่ง ๆ การเว้นวรรคจึงเป็นสิ่งสำคัญเพราะถ้าเว้นวรรคผิดที่ก็จะทำให้เสียความ จากที่กล่าวใน กิตติพร วีรสุธีกุลและยุพิน วีรสุธีกุล (2536) โดยราชบัณฑิตยสถานได้ตั้งคณะกรรมการกำหนดหลักเกณฑ์เกี่ยวกับการใช้ภาษาไทย โดยได้พิจารณาวางหลักเกณฑ์ เรื่อง “การเว้นวรรคในการเขียนหนังสือไทย” โดยมีจุดประสงค์หลักสำคัญ 2 ประการ คือ

- เพื่อให้ถูกต้องเป็นเอกภาพสำหรับถือปฏิบัติเป็นมาตรฐานเดียวกัน
- เพื่อป้องกันมิให้ภาษาไทยต้องเสื่อมโทรม หรือพัฒนาขยายตัวออกไป อย่างไม่มีแบบแผนและหลักเกณฑ์ที่ถูกต้อง

หลักเกณฑ์การเว้นวรรคในภาษาไทย นั้น เขียนไว้ดังนี้

- วรรคเล็ก มีช่องว่างขนาด 1 ตัวอักษร กำหนดให้เป็นตัว ก
- วรรคใหญ่ มีช่องว่างขนาด 2 เท่าของตัว ก

3.6.5.1 การเขียนประโยคให้ถูกต้องในแบบแผนภาษาไทย โดยการเว้นวรรคตอน

โดยการเขียนประโยคต่าง ๆ ในหลักของภาษาไทยที่ถูกต้องนั้น ควรจะต้องมีการเว้นวรรคที่ถูกต้อง จรูญ ตันสูงเนิน (2532) ได้อธิยายไว้ คือ การเว้นวรรคตอนในประโยค หมายถึงการแยกคำหรือเนื้อหาของประโยคออกเป็นส่วนย่อย รวมทั้งการใช้เครื่องหมายวรรคตอนต่างๆ ให้ถูกต้องด้วย ทั้งนี้เพื่อให้เนื้อความดูงามตา อ่านง่ายและป้องกันมิให้ผู้อ่านสับสน

ในการเขียนประโยคนั้น ถึงแม้ผู้เขียนจะเรียงคำได้ถูกต้องตามโครงสร้างของประโยคภาษาไทยและใช้ถ้อยคำได้กะทัดรัดแล้ว แต่ถ้าหากเว้นวรรคผิดนอกจากจะทำให้ผู้อ่านเข้าใจยาก หรือบางทีก็สับสนเข้าใจผิดแล้ว ยังก่อความรำคาญใจขณะอ่าน และไม่ชวนอ่านเนื่องจากข้อความขาดความสวยงามอีกด้วย ดังนั้นการเขียนประโยคจึงควรละเอียดถี่ถ้วนในการเว้นวรรคตอน ดังนี้

3.6.5.1.1 ประโยคความเดียวสั้น ๆ ไม่มีส่วนขยาย ให้เขียนติดต่อกันโดยไม่เว้นวรรค

3.6.5.1.2 ประโยคที่มีประธาน และส่วนขยาย ก็ต้องเว้นวรค แล้วจึงถึงกริยา ถ้ามีส่วนขยายกริยาก็อนุโลมให้เว้นวรรคเช่นเดียวกัน (บุญเหลือ เทพยสุวรรณ, 2510 : 294)

3.6.5.1.3 เมื่อจบประโยคใหญ่ ควรเว้นวรรค 2 ช่วงตัวอักษร ส่วนอนุประโยคหรือวลียาว ๆ ระหว่างประโยค ควรเว้นวรรค 1 ช่วงตัวอักษร

3.6.5.1.4 ประโยคที่ต้องใช้ภาษาไทยปะปนกับภาษาต่างประเทศ จะต้องเว้นวรรคระหว่างภาษานั้น 1 ช่วงตัวอักษร

3.6.5.1.5 การแยกระหว่างสันธาน ถ้าเป็นสันธานเชื่อมคำ ห้ามแยกความออกจากกันต้องเขียนติดกันเสมอ แต่ถ้าเป็นสันธานเชื่อมประโยคให้เขียนแยกได้ โดยเขียนสันธานอยู่ติดกับประโยคหลัง

3.7 คำหยุด

คำหยุด คือ คำที่เกิดขึ้นเป็นจำนวนมากในเอกสาร เป็นคำที่เกิดขึ้นจากไวยากรณ์ภาษาไทยเพื่อให้เกิดความเชื่อมโยงและสอดคล้องกันภายในเอกสาร ดังแสดงในภาคผนวกที่ ค. (ตัวอย่างของคำหยุด)

3.8 การกำกับหน้าที่คำ ทางไวยากรณ์ภาษาศาสตร์

การกำกับหน้าที่คำ คือ การระบุหน้าที่คำของคำที่กำหนดมา ส่วนหน้าที่คำ คือ สิ่งที่ระบุว่าคำหน้าที่ทางไวยากรณ์เป็นอะไรภายในประโยคหนึ่ง ๆ โดยคำหนึ่งคำอาจจะมีหลายหน้าที่ได้ขึ้นอยู่กับตำแหน่งภายในประโยคนั้น เช่น คำว่า “ฉัน” สามารถจะมีหน้าที่ได้ 2 อย่าง คือ 1. เป็นคำกริยา 2. เป็นคำสรรพนาม ตัวอย่างเช่น “ พระฉันเพลก่อนเที่ยง” คำว่า “ฉัน” ในที่นี้ก็จะทำหน้าที่เป็นคำกริยา แต่ถ้าในประโยค “ฉันกับน้องชอบไปดูหนังด้วยกัน” คำว่า “ฉัน” ในที่นี้จะทำหน้าที่เป็นคำสรรพนาม เป็นต้น

สำหรับการวิเคราะห์ทางด้านภาษา ชุดหน้าที่คำ (POS Tag Set) ที่นำมาใช้จะมีผลต่อการวิเคราะห์เป็นอย่างมาก และในความจริงการระบุหน้าที่คำเพียงบอกว่าเป็น คำนาม คำกริยา คำสรรพนาม คำคุณศัพท์ คำวิเศษณ์ ฯลฯ นั้นไม่เพียงพอที่จะนำมาใช้ในการวิเคราะห์ทางภาษาศาสตร์ โดยลักษณะชุดหน้าที่คำของแต่ละภาษานั้นจะมีลักษณะแตกต่างกันไปตามภาษานั้น ๆ และในภาษาหนึ่ง ๆ อาจจะมีชุดหน้าที่คำได้หลายชุดโดยขึ้นอยู่กับแนวคิดของการนำชุดหน้าที่คำไปใช้ ตัวอย่างเช่นในภาษาอังกฤษได้มีการสร้างชุดหน้าที่คำออกมาหลายชุด เช่น ชุดหน้าที่คำเพนทรีแบงค์ (Penn Treebank tag set) ซึ่งในเพนทรีแบงค์นั้นได้แบ่งหมวดหมู่หน้าที่คำ

ออกเป็น 36 ชนิด (Allen, 1995) และ ชุดหน้าที่คำบราวน์ (Brown tag set) ได้แบ่งหมวดหมู่คำออกเป็น 80 ชนิด สำหรับภาษาไทยได้มีการสร้างชุดหน้าที่คำออกมาหลายชุดเช่นกัน ตัวอย่างเช่น **ชุดหน้าที่คำออร์คิด (Orchid tag set)** ซึ่งแบ่งหมวดหมู่คำเป็น 47 ชนิด (วิรัช ศรเลิศล้ำวาณิช และ Thatsanee Charoenporn and Isahara, 1997) และชุดหน้าที่คำของมหาวิทยาลัยเกษตรศาสตร์ เป็นต้น โดยงานวิจัยนี้เลือกแบ่งหมวดคำที่สอดคล้องกับชุดข้อมูลออร์คิด

4. การจัดแบ่งหมวดคำในภาษาไทย

การจัดแบ่งหมวดคำเป็นลักษณะสากล (universality) ของภาษา คำในภาษาทุกภาษาสามารถจัดประเภทเป็นหมวดหมู่ต่าง ๆ โดยคำที่อยู่ในหมวดเดียวกันก็จะคล้ายคลึงกันตามเกณฑ์ที่ใช้ในการจัดหมวดหมู่ เช่น ถ้าใช้เกณฑ์ความหมาย คำที่มีความหมายทำนองเดียวกันก็จัดอยู่ในหมวดเดียวกัน ถ้าใช้เกณฑ์ตำแหน่งในการปรากฏของคำ คำที่มักปรากฏในตำแหน่งเดียวกันก็จัดอยู่ในหมวดเดียวกัน คำ คำหนึ่งเมื่อใช้เกณฑ์หนึ่งอาจจัดอยู่ในหมวดเดียวกับอีกคำหนึ่ง แต่หากเปลี่ยนเกณฑ์ที่ใช้ในการแบ่งหมวดหมู่แล้วคำทั้งสองอาจแยกกันอยู่คนละหมวดก็ได้ แต่ละหมวดหมู่ของคำก็มีชื่อเรียกแตกต่างกัน เช่น คำนาม คำกริยา คำบุรพบท เป็นต้น ทั้งนี้เพื่อให้สะดวกในการนำคำเหล่านี้ไปใช้ในวลีและประโยคต่าง ๆ หมวดหมู่สำหรับคำต่าง ๆ เหล่านี้ในภาษาไทยเรียกว่า “หมวดคำ” หรือ “ชนิดของคำ” ซึ่งตรงกับภาษาอังกฤษว่า word class หรือ part of speech หรือ grammatical category ทุกคำในภาษาอังกฤษต้องจัดอยู่ในหมวดคำหนึ่งอย่างน้อย 1 หมวดคำเสมอ

การแบ่งหมวดคำเป็นสิ่งสำคัญสำหรับการวิเคราะห์ไวยากรณ์ภาษา เนื่องจากในระบบโครงสร้างภาษา ความสัมพันธ์ระหว่างคำในโครงสร้างไม่ใช่เป็นเพียงความสัมพันธ์ระหว่างคำดังกล่าวเท่านั้น แต่ยังแสดงให้เห็นถึงความสัมพันธ์ระหว่างหมวดคำของทั้งภาษาด้วย

4.1 วิธีการในการกำกับหมวดคำภาษาไทยที่ผ่านมา

การกำกับหมวดคำ คือ การกำหนดหมวดคำให้แก่ข้อความป้อนเข้า โดยคอมพิวเตอร์จะเป็นผู้กำกับหมวดคำ (part-of-speech tagger หรือ POS tagger) ให้กับคำแต่ละคำในข้อความที่ป้อนเข้ามา โดยหมวดคำที่ใช้กำกับอาจแตกต่างกันในแต่ละงาน ซึ่งขึ้นอยู่กับว่าผู้พัฒนาโปรแกรมกำกับหมวดคำจะกำหนดให้มีหมวดคำใดบ้างในงานของตน การกำกับหมวดคำมีประโยชน์ช่วยให้การแจงส่วนประโยค (parsing) สามารถทำได้สะดวกขึ้นและถูกต้องยิ่งขึ้น เนื่องจากว่า แต่เดิมความกำกวมจากการที่รูปคำหนึ่ง ๆ เป็นได้หลายหมวดคำทำให้การแจงส่วน

ประโยคประสบปัญหา วิธีการกำกับหมวดคำสามารถแบ่งได้เป็น 2 หลักการใหญ่ คือ หลักการกำกับหมวดคำโดยใช้กฎ (rule based approach) และ หลักการกำกับหมวดคำโดยใช้แบบจำลองไตรแกรม (trigram model approach)

4.2 ชุดหมวดคำในภาษาไทย

4.2.1 การจัดแบ่งหมวดคำโดยใช้ความรู้ทางภาษาของผู้จัดแบ่ง (Intuition based approach)

การจัดแบ่งหมวดคำที่อยู่ในกลุ่มนี้ ผู้วิจัยใช้วิธีพิจารณาจากความรู้ทางภาษาของตนเองว่าควรมีหมวดคำอะไรบ้าง และแต่ละคำควรเป็นหมวดคำอะไร เกณฑ์หลักที่ใช้ในการวิเคราะห์โดยส่วนใหญ่เป็นเกณฑ์ความหมายเป็นสำคัญ

4.2.2 การจัดแบ่งหมวดคำจากการวิเคราะห์คลังข้อมูลหรือประโยคทดสอบ (corpus based approach)

การจัดแบ่งหมวดคำที่อยู่ในกลุ่มนี้ ผู้วิจัยใช้วิธีวิเคราะห์ข้อมูลการใช้ภาษาจริงจากคลังข้อมูลหรือประโยคภาษาไทย แล้วจึงสรุปเป็นชุดหมวดคำออกมา และเกณฑ์หลักที่ใช้ในการวิเคราะห์โดยส่วนใหญ่เป็นเกณฑ์ทางวากยสัมพันธ์โดยพิจารณาจากตำแหน่งในการปรากฏของคำ

งานวิจัยฉบับนี้ผู้วิจัยเลือกใช้การกำกับหมวดคำหรือกำกับหน้าที่คำโดยการแบ่งหมวดคำที่วิเคราะห์แบ่งโดยนักภาษาศาสตร์

5. คลังข้อมูลภาษา

การจัดทำคลังข้อมูลภาษาเพื่อใช้เป็นฐานความรู้สำหรับโปรแกรมในการตัดคำและกำกับหมวดคำภาษาไทย เนื่องจากโปรแกรมจำเป็นต้องเรียนรู้ค่าสถิติที่จะนำไปใช้ทั้งสำหรับการตัดคำและกำกับหมวดคำ คลังข้อมูลที่ใช้ในวิทยานิพนธ์นี้จึงมีลักษณะเป็นคลังข้อมูลภาษาไทยที่มีการตัดคำและกำกับหมวดคำ โดยเริ่มจากการกล่าวถึงบทบาทของคลังข้อมูลภาษาในงานด้านประมวลผลภาษาธรรมชาติ และการนำไปใช้ในงานวิจัยนี้

คลังข้อมูลภาษาถือเป็นทรัพยากรที่สำคัญสำหรับการวิเคราะห์ภาษาและการศึกษาทางภาษาศาสตร์ และในปัจจุบันได้ถูกนำมาใช้อย่างแพร่หลายในงานด้านประมวลผลภาษาธรรมชาติ เพื่อเป็นฐานความรู้ด้านต่าง ๆ ให้กับระบบ โดยเฉพาะอย่างยิ่งเมื่อยุคหลัง ๆ การประมวลผลภาษาธรรมชาติมักประยุกต์วิธีการทางสถิติเข้ามาใช้ คลังข้อมูลก็จะเป็นแหล่งข้อมูล

ทางภาษาขนาดใหญ่สำหรับเก็บข้อมูลทางสถิติเข้ามาใช้ เพื่อช่วยในการประมวลผลให้มีความถูกต้องแม่นยำและมีประสิทธิภาพสูงขึ้น หรือแม้แต่งานที่อาศัยกฎในปัจจุบันซึ่งได้รับการพัฒนาแก้ไขขึ้นมาใหม่ เช่น งาน Brill (1993) ก็ยังให้ระบบเรียนรู้และสรุปกฎจากคลังข้อมูลเช่นกัน

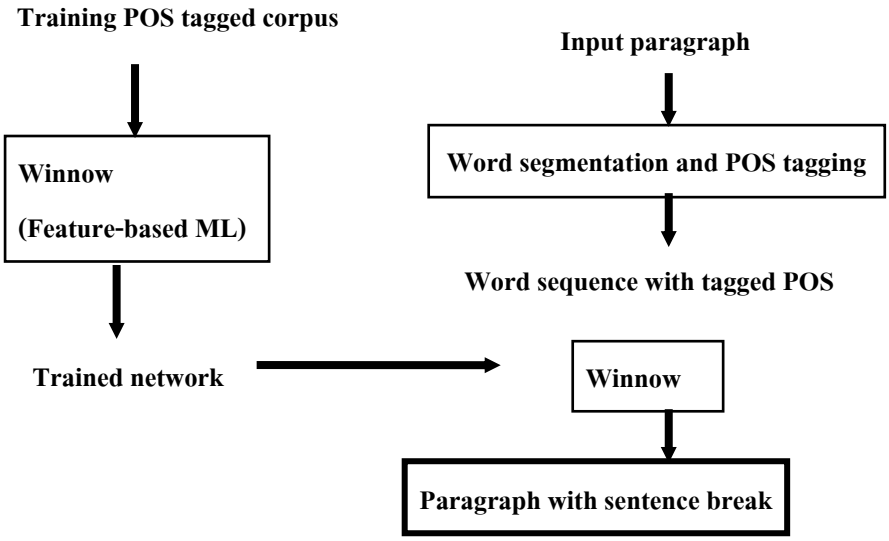
คลังข้อมูลที่นำมาใช้ในการประมวลผลภาษาธรรมชาติได้รับการพัฒนารูปแบบวิธีการเรื่อยมา ทำให้มีลักษณะแตกต่างกันไปทั้งในด้านรูปแบบและเนื้อหา มีทั้งที่เป็นคลังข้อมูลล้วน (plain corpus) ซึ่งประกอบด้วยข้อความอย่างเดียว และคลังข้อมูลที่กำกับความหมายหรือบางงานต้องการข้อมูลที่เป็นตัวแทนของทั้งภาษา ในขณะที่บางงานต้องการข้อมูลภาษาย่อย (sub-language) เฉพาะเรื่องเท่านั้น ทั้งนี้ขึ้นอยู่กับจุดประสงค์ วิธีการ และลักษณะของงานที่จะนำคลังข้อมูลไปใช้

สำหรับคลังข้อมูลภาษาไทยนั้น เพิ่งจะได้รับการพัฒนาในช่วงไม่กี่ปีที่ผ่านมา จึงทำให้ในปัจจุบันมีคลังข้อมูลภาษาไทยอยู่จำนวนน้อย คลังข้อมูลภาษาไทยที่ได้รับความนิยมนำไปใช้ในงานวิจัยต่าง ๆ ได้แก่ คลังข้อมูลออร์คิด ซึ่งสร้างและพัฒนาโดยกลุ่มวิจัยภาษาและวิทยาการความรู้ (LINKS) แห่งศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ โดยมุ่งหวังที่จะให้บริการเป็นแหล่งข้อมูลภาษาไทยสำหรับการวิจัยทางภาษาและทางการประมวลผลภาษาธรรมชาติ คลังข้อมูลออร์คิดมีลักษณะเป็นคลังข้อความภาษาไทยที่มีการตัดแบ่งประโยคและมีการกำกับหมวดคำด้วยควบคู่ด้วย

5.1 วิธีการสร้างฐานข้อมูลขนาดใหญ่ (Corpus)

การสร้างฐานข้อมูลนั้น เป็นการใช้เทคโนโลยีทางคอมพิวเตอร์รวบรวมและคัดเลือกคำ ประโยค หรือข้อความ ที่มีใช้จริงและมีอัตราการปรากฏสูงในบริบทต่างๆ ของการใช้ภาษา จากแหล่งข้อมูลและข่าวสารที่เผยแพร่ทางอินเทอร์เน็ตและแหล่งข้อมูลอื่นๆ ที่เชื่อถือได้ เช่น วรรณกรรม บทความ เอกสารทางวิชาการ ข้อมูลข่าวสารจากหนังสือพิมพ์ เป็นต้น

ลักษณะเด่นของข้อมูล คือ แสดงความหมายและประเภทของคำพร้อมทั้งประโยคตัวอย่างที่มีใช้จริงและมีอัตราการปรากฏสูงในบริบทต่างๆ ของการใช้ภาษา



ภาพประกอบที่ 2-1 ขั้นตอนกระบวนการสร้างคลังข้อมูล (ไพศาล เจริญพรสวัสดิ์, 2542)

5.2 คลังและชุดหมวดคำออร์คิด

5.2.1 คลังข้อมูลภาษา

คือ ข้อมูลภาษาเขียนหรือภาษาพูดที่เป็นภาษาที่ใช้จริงซึ่งถูกรวบรวมขึ้นมาในปริมาณที่มากเพียงพอตามข้อกำหนดหรือเงื่อนไขที่กำหนดขึ้น เพื่อนำคลังข้อมูลนั้นมาใช้ประโยชน์ในการศึกษาเรื่องราวต่าง ๆ ที่เกี่ยวข้องกับภาษา

5.2.2 คลังข้อมูลออร์คิด

ORCHID (The Open linguistic Resources CHanelled toward InterDisciplinary research) เป็นแผนงานเพื่อสนับสนุนการร่วมสร้าง การร่วมใช้ การร่วมกันพัฒนาทรัพยากรทางภาษาของภาษาไทย บนเครือข่าย World Wide Web (อรุณี โอฬารานนท์, 2546)

ORCHID POS Tagged Corpus เป็นข้อมูลพื้นฐานทางภาษาที่มีการกำกับหน่วยที่เป็นคำ หร้อมกับหน้าที่ของคำ ที่สร้างขึ้นจากเอกสารรายงานการประชุมต่าง ๆ ของเนคเทคเอง เป็นการพัฒนาคลังข้อมูลสำหรับนำไปใช้ในการศึกษาและให้ข้อมูลเกี่ยวกับการใช้คำในภาษาปริมาณ และความถูกต้อง ออร์คิดเป็นคลังข้อความภาษาไทยขนาดใหญ่ที่มีการกำกับ **หน้าที่คำ** ประกอบด้วยจำนวนข้อมูล 2 เมกะไบต์ (จำนวนคำราวๆ 400,000 คำ) พัฒนาโดย ฝ่ายวิจัยและ

พัฒนาสาขาสารสนเทศ และเป็นการนำเทคโนโลยีฐานข้อมูลขนาดใหญ่เข้ามาช่วยในการวิจัยและพัฒนาในสาขาการประมวลผลภาษาธรรมชาติ เรียกว่า การสร้างคลังข้อมูลจากฐานข้อมูลขนาดใหญ่

5.2.2.1 โครงสร้างของคลังข้อมูลออร์คิด

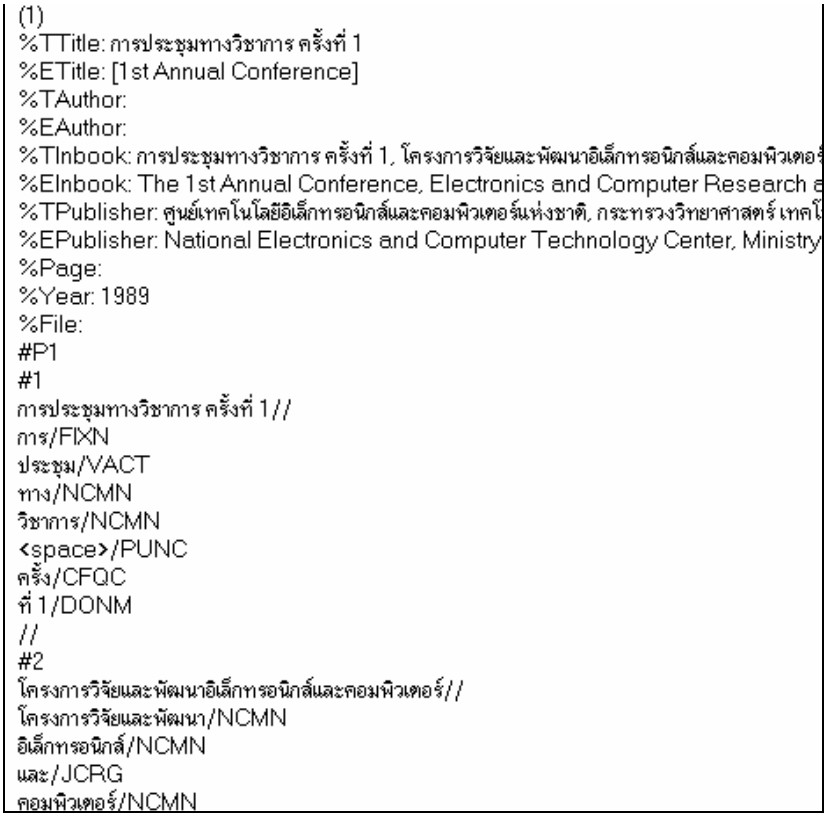
Mark-up	Description
#P[number]	Paragraph number of a text. The number in the bracket is shown in a sequence within a text.
#[number]	Sentence number of a paragraph. The number in the bracket is shown n a sequence within a paragraph.
\\	Line break symbol.
//	Sentence break symbol.
/[POS]	Tag marker for the appropriate POS of a word.
%TTitle:	Title of the document written in Thai.
%ETitle:	Title of the document written in English.
%TAuthor:	Author's name written in Thai.
%EAuthor:	Author's name written in English.
%TInbook:	Title of the book where the document exists, written in Thai.
%EInbook:	Title of the book where the document exists, written in English.

ตารางที่ 2-2 รายละเอียดโครงสร้างฐานข้อมูลออร์คิด (Virach Sorlertlamvanich , et al , 1997)

5.2.3 การนำคลังข้อความออร์คิดมาใช้สำหรับในการทดสอบ

คลังข้อความที่นำมาใช้ในการเรียนรู้และทดสอบของโครงข่ายประสาทเทียมในการสกัดคำและวลีสำคัญนั้น ได้นำมาจาก " คลังข้อความออร์คิด " (Virach Sorlertlamvanich , et al , 1997) โดยได้รับความอนุเคราะห์จาก ห้องปฏิบัติการวิจัยและพัฒนาวิศวกรรมภาษาและซอฟต์แวร์ (Software and Language Engineering Laboratory : SLL) โดยลักษณะของบทความที่นำมาใช้สร้างคลังข้อความนั้นมีจำนวนคลังข้อความนั้นมีอยู่ประมาณ 25,000 ประโยค โดยที่

ภายในคลังข้อความนี้ได้ทำการแบ่งเป็นประโยค ส่วนภายในประโยคจะแบ่งเป็นคำต่าง ๆ และยังได้มีการกำหนดหน้าที่ของคำด้วย ซึ่งทั้งหมดทำโดยนักภาษาศาสตร์ ตัวอย่างประโยคภายในคลังข้อความออร์คิด แสดงในรูปที่ 2-2


(1)
%TTitle: การประชุมทางวิชาการ ครั้งที่ 1
%ETitle: [1st Annual Conference]
%TAuthor:
%EAuthor:
%TInbook: การประชุมทางวิชาการ ครั้งที่ 1, โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์
%EInbook: The 1st Annual Conference, Electronics and Computer Research a
%TPublisher: ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, กระทรวงวิทยาศาสตร์ เทคโ
%EPublisher: National Electronics and Computer Technology Center, Ministry
%Page:
%Year: 1989
%File:
#P1
#1
การประชุมทางวิชาการ ครั้งที่ 1//
การ/FIXN
ประชุม/VACT
ทาง/NCMN
วิชาการ/NCMN
<space>/PUNC
ครั้ง/CFQC
ที่ 1/DONM
//
#2
โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์//
โครงการวิจัยและพัฒนา/NCMN
อิเล็กทรอนิกส์/NCMN
และ/JCRG
คอมพิวเตอร์/NCMN

ภาพประกอบที่ 2-2 ตัวอย่างประโยคภายในคลังข้อมูลออร์คิด
(Virach Sorlertlamvanich , et al , 1997)

หน้าที่คำที่นำมาใช้คลังข้อความนี้ของงานวิจัยนี้ ได้ถูกแบ่งออกเป็น 47 หมวด (Virach Sorlertlamvanich , et al , 1997) ซึ่งทำการแบ่งโดยนักภาษาศาสตร์ การที่ต้องแบ่งหน้าที่คำให้ละเอียดลงไปนั้น เนื่องจากถ้าแบ่งหน้าที่คำโดยแค่แบ่งเป็น คำนาม คำกริยา คุณศัพท์ ฯลฯ เท่านั้นจะไม่เพียงพอต่อการนำมาใช้ในการวิเคราะห์ทางภาษา จึงทำให้นักภาษาศาสตร์ได้มีการวิเคราะห์และแบ่งหน้าที่ของคำออกมาเป็นหมวดหมู่ต่าง ๆ

6. การตัดคำไทยในระบบแปลภาษา (Word Segmentation)

การตัดคำ เป็นการแบ่งข้อความที่ต่อเนื่องออกเป็นหน่วยคำ ๆ (morpheme) เนื่องจากภาษาไทยมีการเขียนในลักษณะที่ติดต่อกันโดยไม่มีเครื่องหมายวรรคตอนใด ๆ

6.1 วิธีในการตัดคำภาษาไทยที่ผ่านมา

ที่ผ่านมาจนถึงปัจจุบัน งานด้านการตัดคำภาษาไทยได้รับการพัฒนาจากหน่วยงานวิจัยต่าง ๆ ทั้งของภาครัฐและภาคเอกชน โดยมีการพัฒนาแนวคิดและวิธีการต่าง ๆ เพื่อใช้ในการตัดคำมาเป็นลำดับ แต่ละวิธีการต่างก็ให้ผลในด้านความถูกต้อง ความรวดเร็วของการทำงาน และปริมาณการใช้ทรัพยากรต่าง ๆ แตกต่างกัน วิธีการตัดคำภาษาไทยสามารถแบ่งได้เป็น 3 หลักการใหญ่ ๆ คือ หลักการตัดคำโดยใช้กฎ หลักการตัดคำโดยใช้พจนานุกรม (dictionary approach) และหลักการตัดคำโดยใช้คลังข้อมูล

6.1.1 หลักการตัดคำโดยใช้กฎ

การตัดคำโดยใช้กฎเป็นความพยายามในขั้นเริ่มต้นของการพัฒนาระบบตัดคำภาษาไทยโดยใช้วิธีการตรวจสอบกฎเกณฑ์ทางอักขระวิธีที่กำหนดลักษณะของการประสมอักษร การเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการบ่งชี้ขอบเขตของคำ ตัวอย่างเช่น

1. การขึ้นย่อหน้าเป็นตัวบ่งชี้ถึงการสิ้นสุดข้อความ
2. การเว้นวรรคเป็นตัวบ่งชี้ถึงความเป็นไปได้ของการสิ้นสุดคำหรือประโยค
3. กฎทางอักขระวิธีเป็นตัวบ่งชี้ถึงความเป็นไปได้ของการตัดคำในตำแหน่งนั้น ๆ โดยได้แบ่งอักขระออกเป็น 5 กลุ่ม ได้แก่

- อักขระกลุ่ม Non-spacing character คือ รูปสระ วรรณยุกต์ และเครื่องหมายที่เมื่อประสมเข้ากับพยัญชนะแล้วไม่ทำให้มีการเคลื่อนขวาของตำแหน่ง อักขระในกลุ่มนี้ไม่สามารถปรากฏเดี่ยวได้ เช่น อั , อิ , อุ , อ๊ , อ่ ,อ์
- อักขระกลุ่มที่ต้องมีพยัญชนะตามเสมอ เช่น เ , แ , โ , ไ , ใ
- อักขระกลุ่มที่ต้องมีพยัญชนะนำเสมอ เช่น อะ , อา , อำ
- อักขระกลุ่มที่เป็นตัวการันต์ที่มีทัณฑฆาตบังคับข้างบน เช่น ย์ เนื่องจากว่าตัวการันต์เป็นพยัญชนะสุดท้ายจึงไม่พิจารณาให้เป็นอักขระแรกของคำ

- อักขระที่เหลือทั้งหมด วิรัช ศรเลิศล้ำวาณิช (2536) กล่าวว่า อักขระกลุ่มที่ 2 3 และ 4 สามารถเป็นตัวจำกัด ไม่ให้มีการตัดคำในระหว่างคำอักขระนั้น ๆ

วิรัช ศรเลิศล้ำวาณิช (2536) กล่าวว่า กฎเกณฑ์ทั้ง 3 ลักษณะนี้สามารถใช้เป็นตัวช่วยในการบ่งชี้ขอบเขตของการพิจารณาในการเปรียบเทียบกับคำในพจนานุกรม

กฎทางอักขระวิธีที่นำมาใช้พิจารณาตัดคำข้างต้นไม่สามารถบ่งบอกตำแหน่งในการตัดคำได้อย่างถูกต้อง เนื่องจากอันที่จริงกฎดังกล่าวไม่ได้บ่งชี้ขอบเขตของคำ แต่เป็นตัวบ่งชี้ขอบเขตของพยางค์ ตัวอย่างเช่น “กระโดด” เป็นคำหนึ่งคำ ดังนั้นไม่สามารถตัดคำหลัง อะ (อักขระกลุ่มที่ 2) และไม่สามารถตัดคำหน้า โอ (อักขระกลุ่มที่ 3) ได้ ส่วนตัวการันต์ (อักขระกลุ่มที่ 4) ก็ไม่ได้ปรากฏเป็นตำแหน่งจบคำเสมอไป ดังตัวอย่างเช่น กอล์ฟ เป็นต้น ดังนั้น กฎเกณฑ์ทางอักขระวิธีจึงเป็นกฎเกณฑ์ที่ใช้ในการตัดพยางค์มากกว่าที่จะเป็นกฎในการตัดคำ

6.1.2 วิธีการตัดคำโดยอาศัยหลักการตัดคำโดยใช้คลังข้อมูล (corpus based approach)

วิธีการตัดคำโดยอาศัยหลักการตัดคำโดยใช้คลังข้อมูล เป็นหลักการตัดคำโดยใช้คลังข้อมูลเป็นแนวคิดที่ได้รับการพัฒนาในยุคหลัง ๆ ซึ่งเป็นยุคที่มีผู้สนใจนำวิธีการทางสถิติมาใช้ในการประมวลผลภาษาธรรมชาติมากขึ้น โดยใช้คลังข้อมูลทางภาษาเป็นฐานความรู้สำหรับเก็บค่าความถี่ที่ใช้ในการตัดคำ

6.1.2.1 วิธีการตัดคำโดยอาศัยความน่าจะเป็น (ไพศาล เจริญพรสวัสดิ์, 2542)

วิธีการตัดคำโดยอาศัยความน่าจะเป็นได้ใช้แบบจำลองไตรแกรมของคำ (word trigram model) เพื่อหารูปแบบการตัดคำที่เป็นไปได้มากที่สุดไปพร้อม ๆ กัน วิธีการนี้ต้องใช้คลังข้อมูลที่มีการตัดคำและกำกับหมวดคำเตรียมเอาไว้แล้ว โดยปัญหาการตัดคำภาษาไทยสามารถแสดงด้วยสมการดังภาพประกอบที่ 2-3 ดังต่อไปนี้

$$\arg\max_{W_{1,n}} P(W_{1,n} \mid C_{1,n}) = \arg\max_{W_{1,n}} \frac{P(C_{1,m} \mid W_{1,n}) * P(W_{1,n})}{P(C_{1,m})} = \arg\max_{W_{1,n}} P(W_{1,n}) = \arg\max_{W_{1,n}} \sum_{T_{1,n}} P(W_{1,n}, T_{1,n})$$

ภาพประกอบที่ 2-3 แสดงสมการปัญหาการตัดคำภาษาไทย

กำหนดให้ $C_{1,n}$ หมายถึง สายอักขระที่ป้อนเข้าไปตั้งแต่ตัวที่ 1 ถึงตัวที่ m

$W_{1,n}$ หมายถึง สายคำที่สามารถตัดออกมาได้ ตั้งแต่คำแรกถึงคำที่ n

$T_{1,n}$ หมายถึง สายหมวดคำที่กำกับอยู่กับแต่ละคำ ทั้งหมด n คำ

ปัญหาของการตัดคำ คือ ต้องการหาสายคำ $W_{1,n}$ ที่ทำให้ค่าความน่าจะเป็นของ $P(W_{1,n})$ มีค่าสูงที่สุด โดยเมื่อนำหมวดคำเข้ามาช่วยคำนวณด้วยแล้ว จะสามารถหาค่า $P(W_{1,n})$ ได้ดังสมการ

สมการดังกล่าว มีความหมายว่า จากสายอักขระ C1,n ที่กำหนดให้ซึ่งเป็นข้อความที่ป้อนเข้าไป เราต้องการแบ่งสายอักขระนี้ออกเป็นคำ $W_1, W_2, \ldots\ldots, W_n$ (ซึ่งเขียนสั้น ๆ ได้ว่า $W_{1,n}$) และมีหมวดคำเป็น $T_1, T_2, \ldots.., T_n$ (หรือ $T_{1,n}$) เราต้องการตัดคำให้ได้ค่าของ $P(W_{1,n} \mid C_{1,m})$ ที่สูงที่สุด ซึ่งสามารถคำนวณได้จาก $P(W_{1,n}, C_{1,m}) / P(C_{1,m})$ และเราสามารถแปลงการหาค่าความน่าจะเป็นแบบมีเงื่อนไข (condition probability) ตรงนี้ได้โดยนำกฎของเบยส์ (Bayes' rule) มาใช้ ดังนั้นจึงสามารถแปลงสมการได้เป็น $P(C_{1,m} \mid W_{1,n}) * P(W_{1,n}) / P(C_{1,m})$ (ดังแสดงในสมการ) ซึ่งสามารถลดเหลือเพียงการหาค่า $P(W_{1,n})$ (ดังแสดงในสมการ) เนื่องจาก $P(C_{1,m} \mid W_{1,n})$ มีค่าเท่ากับ 1 และ $P(C_{1,m})$ สามารถละไปได้ เนื่องจากเป็นตัวหารสำหรับทุกทางเลือกของการตัดคำที่จะนำมาเปรียบเทียบกัน จากนั้น จึงได้นำหมวดคำมาช่วยพิจารณาด้วย จึงได้สมการเป็นการหาค่า $ET_{1,n}$ $P(W_{1,n}, T_{1,n})$ (ดังแสดงในสมการ) ซึ่งสามารถแปลงสมการได้เป็น $P(W_{1,n} \mid T_{1,n}) * P(T_{1,n})$ แล้วจึงประยุกต์ใช้แนวคิดของแบบจำลองไตรแกรม (trigram) ที่สมมติฐานว่า :

(1) คำหนึ่ง ๆ สามารถปรากฏ ณ ตำแหน่งใด ๆ ในประโยคได้ โดยไม่ขึ้นกับคำหรือหมวดคำที่อยู่ก่อนหน้าหรือตามหลัง กล่าวคือ $P(W_{1,n} \mid T_{1,n})$ มีค่าประมาณเท่ากับ ผลคูณรวมของ $P(W_i \mid T_i)$ ของทุกคำ

(2) ความน่าจะเป็นที่หมวดคำหนึ่งปรากฏ ณ ตำแหน่งใด ๆ ในประโยคจะขึ้นอยู่กับหมวดคำที่ปรากฏก่อนหน้า 2 หมวดคำเท่านั้น (trigram model) กล่าวคือ $P(T_{1,n})$ คำนวณแบบประมาณค่าได้จาก $P(T_i \mid T_{i-1}, T_{i-2})$

ดังนั้น สมการในภาพประกอบที่ 2-3 จะสามารถคำนวณโดยการประมาณค่าตามแบบจำลองไตรแกรมดังสมการในภาพประกอบที่ 2-4

$$\arg\max_{W_{1,n}} \sum_{T1,n} \prod_{l=1,n} P(W_i \mid T_i) * P(T_i \mid T_{i-1}, T_{i-2})$$

ภาพประกอบที่ 2-4 แสดงสมการการประมาณค่าตามแบบจำลองไตรแกรม

จากคลังข้อมูลภาษาไทยที่มีอยู่ เราสามารถหาค่าของ $P(W_i \mid T_i)$ ได้โดยนับจำของคำ W_i ที่มีหมวดคำเป็น T_i หารด้วยจำนวนของ T_i ที่เป็นหมวดคำของคำใดๆ (จำนวนของ T_i ที่ปรากฏทั้งหมดในคลังข้อมูล) ส่วน $P(T_i \mid T_{i-1}, T_{i-2})$ เราสามารถได้โดยนับจำนวนหมวดคำ T_i ที่มีหมวดคำ T_{i-2} และ T_{i-1} นำหน้า หารด้วยจำนวนสายหมวดคำ T_{i-2} และ T_{i-1} ที่ปรากฏติดกันทั้งหมดในคลังข้อมูลขั้นตอนการคำนวณค่าความน่าจะเป็นที่ใช้ในการตัดคำวิธีนี้ ดังภาพประกอบที่ 2-5 ดังนี้

1) คำนวณ Condition prob $\arg\max_{W_{1,n}} P(W_{1,n} \mid C_{1,n}) = \arg\max_{W_{1,n}} P(C_{1,n}, W_{1,m}) * P(C_{1,m})$

2) แปลงสมการโดยใช้ Bay's rule $= \arg\max_{W_{1,n}} P(C_{1,m} \mid W_{1,n}) * P(W_{1,n}) / P(C_{1,m})$

3) $P(C_{1,m} \mid W_{1,n})$ มีค่าเท่ากับ 1 และ $P(C_{1,m})$ เป็นตัวหารที่เท่ากันทุกทาง เลือกสามารถตัดทิ้งได้ $= \arg\max_{W_{1,n}} P(W_{1,n})$

4) นำ T มาช่วย $= \arg\max_{W_{1,n}} \sum P(W_{1,n}, T_{1,n})$

5) แปลง joint prob เป็น conditional prob $= \arg\max_{W_{1,n}} \sum P(W_{1,n}, T_{1,n}) * P(T_{1,n})$

6) นำแนวคิดไตรแกรม

(1) 1exical generation prob : คำหนึ่งปรากฏ ณ ตำแหน่งใด ๆ ในประโยคได้โดยไม่ขึ้นกับสิ่งอื่น $P(W_{1,n}, T_{1,n})$

(2) tag sequence prob ความน่าจะเป็นที่หมวดคำหนึ่ง ๆ จะปรากฏ ณ ตำแหน่งใด ในประโยคจะขึ้นอยู่กับหมวดคำที่ปรากฏก่อนหน้า 2 หมวดคำเท่านั้น (trigram model)

$$P(T_{1,n}) \cong \prod_{i=1...n} P(T_i \mid T_{i=1}, T_{i=2})$$

7) $= \arg\max_{W_{1,n}} \sum T_{1,n} P(W_i \mid T_i) * P(T_i \mid T_{i-1}, T_{i-2})$

ภาพประกอบที่ 2-5 ขั้นตอนการตัดคำโดยอาศัยความน่าจะปันตามแบบจำลองไตรแกรม

อย่างไรก็ตาม หากพิจารณาตามสมการที่ใช้ดังกล่าว จะพบว่าการนำหมวดคำมาคำนวณ ไม่ได้มีผลโดยตรงใด ๆ ต่อการตัดคำ เพราะเป็นการหาเฉพาะสายคำที่ให้ค่าความน่าจะเป็นสูงที่สุด และค่าความน่าจะเป็นนี้ได้มาจากการนำค่าความน่าจะเป็นทั้งหมดของสายคำหนึ่ง ๆ ซึ่งมีการกำกับสายหมวดคำแบบต่าง ๆ กันมารวมกัน ซึ่งจะเท่ากับการหาค่าความน่าจะเป็นของสายคำนั้น สามารถแยกการตัดคำทั้งสองแบบออกจากกันได้ ตัวอย่างของคุณลักษณะที่ใช้แก้ความกำกวม ของ "มากกว่า" ได้

(1) มากกว่า number (collocation ที่บ่งปริบทควรตัดเป็น มา กว่า)

(2) พูด within – 10 words (context word ที่บ่งปริบทว่าควรตัดคำเป็นมาก ว่า)

7.1.3 หลักการตัดคำโดยใช้พจนานุกรม

การตัดคำโดยใช้พจนานุกรมเป็นแนวคิดที่ได้รับการพัฒนาในยุคต่อมา โดยเก็บคำภาษาไทยไว้ในพจนานุกรม แล้วนำข้อความที่ป้อนเข้าไปค้นหาและเทียบสายอักขระกับคำในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวควรตัดคำในบริเวณใด และประกอบด้วยคำใดบ้าง (สมปราถนา รัทยานนท์ , 2525) ได้เสนอวิธีการใช้พจนานุกรมช่วยในการตัดคำภาษาไทย ใช้การค้นหาคำจากข้อความที่ป้อนเข้าไปเทียบกับคำในพจนานุกรมเพื่อนำแต่ละคำไปจัดเก็บไว้ในแถวลำดับหรืออะเรย์ (array) ชุดหนึ่ง โดยเริ่มต้นค้นหาจากต้นข้อความ นำคำแรกที่เทียบเจอในพจนานุกรมไปจัดเก็บไว้ในอะเรย์ช่องที่หนึ่ง แล้วจึงตัดดังกล่าวออกไปจากข้อความ แล้วนำข้อความที่เหลือหลังจากตัดคำออกไป มาทำการเทียบคำกับพจนานุกรมเหมือนเดิม เพื่อนำแต่ละคำที่เทียบเจอไปจัดเก็บไว้ในอะเรย์ช่องต่อ ๆ ไปจนสิ้นสุดข้อความ (สมมติให้อะเรย์ช่องสุดท้ายเป็นช่องที่ n) แล้วจึงย้อนการทำงานกลับโดยทำการเทียบคำที่อยู่ในอะเรย์แต่ละช่องกับพจนานุกรมตั้งแต่ช่องที่ n , n-1 , n-2 , … ไปจนถึงอะเรย์ช่องที่ 1 เพื่อหาว่าในแต่ละช่องอะเรย์สามารถตัดคำในรูปแบบอื่นที่ต่างออกไปได้หรือไม่ หากอะเรย์ช่องใด (สมมติให้เป็นช่องที่ i : i<= n) สามารถตัดคำในรูปแบบอื่นได้ก็จะตัดคำในรูปแบบใหม่นั้น แล้วจึงทำการเทียบคำต่อไปในช่องอะเรย์ถัดไป (ช่องที่ i+1) จนจบข้อความ การทำงานดังกล่าวจะทำย้อนกลับไปจนถึงอะเรย์ช่องที่ 1 แล้วจึงจบการทำงาน ดังนั้นผลที่ได้จะเป็นรูปแบบความเป็นไปได้ของการตัดคำทั้งหมดของข้อความที่ป้อนเข้า

หลักการตัดคำโดยใช้พจนานนุกรมนี้สามารถตัดคำได้ถูกต้องมากกว่าการใช้กฎ เพราะฉะนั้น จึงได้รับความนิยมและมีผู้พัฒนาวิธีการตัดคำภาษาไทยอื่น ๆ โดยใช้พจนานุกรมช่วยอีก เช่น วิธีการเทียบคำที่ยาวที่สุด (longest matching) และวิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่พบในพจนานุกรมน้อยที่สุด และวิธีตัดคำแบบ Bigram ซึ่งเป็นเทคนิคการตัดคำในงานวิจัยนี้

7.1.3.1 วิธีการเทียบคำที่ยาวที่สุด

วิธีการเทียบคำที่ยาวที่สุดเป็นวิธีการตัดคำทางวิทยาการศึกษาสำนึก วิธีหนึ่ง ซึ่งต้องใช้พจนานุกรมช่วยรู้จำคำภาษาไทย โดยวิธีนี้จะทำการตรวจสอบหรือสแกนข้อความที่ป้อนเข้าจากซ้ายไปขวา นำไปเทียบกับพจนานุกรมดูว่า สายอักขระดังกล่าวเป็นหนึ่งคำหรือไม่ หากไม่พบว่าสายอักขระดังกล่าวสามารถเทียบเป็นคำได้ในพจนานุกรม ก็จะทำการลดความยาวของสายอักขระลงทีละตัว จนกว่าสายอักขระที่ตรวจสอบจะสามารถเทียบเป็นคำในพจนานุกรมได้ ก็จะทำเครื่องหมายเพื่อเป็นจุดย้อนกลับ จากนั้นก็จะเริ่มทำงานจากจุดย้อนกลับนั้นเพื่อตรวจสอบ

สายอักขระที่เหลือว่าจะสามารถตัดสายอักขระใดต่อไปให้เป็นคำได้ หากตัวเลือกในตอนแรกนี้สามารถทำให้ขั้นตอนวิธี (algorithm) ค้นหาคำที่เหลือได้ ตัวเลือกนี้ก็จะเป็นคำแรกของข้อความได้จริง ไม่เช่นนั้นขั้นตอนวิธีก็จะกลับไปยังจุดย้อนกลับที่ทำเครื่องหมายไว้เพื่อแก้ไขคำแรกใหม่ จากนั้นก็จะเริ่มทำงานต่อไปโดยเริ่มจากจุดย้อนกลับ หากยังไม่สามารถเทียบสายอักขระกับคำในพจนานุกรมได้ก็จะทำการลดตัวอักษรทีละตัวจนกว่าจะเทียบคำในพจนานุนกรมได้ และทำงานในรูปแบบนี้ต่อไปจนจบข้อความ

สมศักดิ์ จันวันและคณะ (2532 : 279-290) อธิบายวิธีที่ตรวจสอบข้อความทั้งประโยคโดยการนำข้อความนั้นไปตรวจสอบกับพจนานุกรม ถ้ามีคำศัพท์นั้นอยู่ก็ถือว่าได้การตัดคำ 1 ครั้ง แล้วนำข้อความที่เหลือไปตรวจสอบต่อ แต่ถ้าไม่พบคำศัพท์ก็ให้ตัดอักขระตัวสุดท้ายออกแล้วนำไปตรวจสอบในพจนานุกรมต่อ วิรัช ศรเลิศล้ำวาณิช (2538) ได้อธิบายว่ากระบวนการตัดคำแบบยาวที่สุด มีรูปแบบดังนี้

ก. ตรวจสอบข้อความทั้งหมดโดยถือว่าเป็นคำศัพท์เดียวกับคำในพจนานุกรม ซึ่งก็จะไม่พบคำศัพท์ดังกล่าว

ข. ตัดอักขระชุดสุดท้ายออกเหลือข้อความว่า “ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคัญ”

ค. นำข้อความที่เหลือไปตรวจสอบกับพจนานุกรมซึ่งก็ไม่พบคำศัพท์ดังกล่าว

ง. ตัดอักขระชุดสุดท้ายออก จะเหลือข้อความ “ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคัญ” ตัด “คัญ” ออกตามกฎของอักขระวิธี

จ. นำข้อความที่เหลืออยู่ไปตรวจสอบกับพจนานุกรมถ้ามีข้อความดังกล่าวแสดงว่าได้ผลของการตัดคำ1 คำ โดยวิธีดังกล่าวจะได้ผลลัพธ์ดังตาราง 3.1

ลำดับที่	ส่วนของคำที่ยาวที่สุด	ส่วนของข้อความที่ต้องตรวจสอบต่อ
1	ความก้าวหน้า	ทางด้านวิทยาศาสตร์มีบทบาทสำคัญ
2	ทาง	ด้านวิทยาศาสตร์มีบทบาทสำคัญ
3	ด้าน	วิทยาศาสตร์มีบทบาทสำคัญ
4	วิทยาศาสตร์	มีบทบาทสำคัญ
5	มี	บทบาทสำคัญ
6	บทบาท	สำคัญ
7	สำคัญ	

ตารางที่ 2-4 ตารางอธิบายผลลัพธ์การตัดคำแบบยาวที่สุด ในแต่ละขั้นตอนของข้อความ " ความก้าวหน้าทางวิทยาศาสตร์มีบทความสำคัญ " (วิรัช ศรเลิศล้ำวาณิช, 2538)

7.1.3.2 วิธีการตัดคำให้ได้จำนวนคำที่ไม่มีในพจนานุกรมน้อยที่สุด (Maximal Matching)

วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุดก็เป็นวิธีการตัดคำทางฮิวริสติก อีกวิธีหนึ่งที่ใช้พจนานุกรมช่วยจำรู้คำภาษาไทย วิธีการนี้พัฒนาโดย วิรัช ศรเลิศล้ำวาณิช (2538) เพื่อแก้ปัญหาที่ปรากฏในวิธีการเทียบคำที่ยาวที่สุด วิธีการนี้จะพัฒนาบนขั้นตอนวิธีการเทียบคำที่ยาวที่สุด เริ่มจากการหาทางเลือกของรูปแบบการตัดคำทั้งหมดที่เป็นไปได้เสียก่อน โดยทำการย้อนกลับทีละคำหลังจากคำตอบจากวิธีการเทียบคำที่ยาวที่สุดแล้ว แล้วจึงเลือกทางเลือกที่มีจำนวนคำน้อยที่สุด Surapant Meknavin and Boonserm Kijsirikul (2000) กล่าวว่า การค้นหาทุกทางเลือกที่เป็นไปได้นี้ทำให้ต้องเสียเวลาในการคำนวณมาก แต่ก็สามารถลดเวลาลงได้โดยใช้โปรแกรมแบบพลวัติ (dynamic programming) ตัวอย่างเช่น หากป้อนข้อความ " ไปหามเหสี " เข้าไป ขั้นตอนวิธีนี้จะหาทางเลือกทั้งหมดของรูปแบบการตัดคำที่เป็นไปได้ 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 ขั้นตอนวิธีของวิธีการเทียบคำที่ยาวที่สุดจะได้คำตอบของการตัดคำเป็น " ไป-หาม-เห-สี " ก่อน หลังจากนั้นจึงเริ่มทำงานย้อนกลับโดยเริ่มจากคำแรก

พบว่า คำที่สอง “ หาม ” สามารถแบ่งได้เป็น “ หา-ม ” ได้ โดยเมื่อแบ่งเป็น “ หา ” แล้วทำให้สายอักขระที่เหลือ “ มเหสี ” และ “ [ม]-เห-สี ” เมื่อทำการย้อนกลับกับทุกคำสิ้นสุดแล้ว ก็จะทำการคำนวณหาค่า cost ให้กับแต่ละทางเลือกที่เป็นไปได้ โดยบังคับให้มีการเกิดคำที่ไม่มีในพจนานุกรมน้อยที่สุด แล้วจัดเรียงผลลัพธ์โดยให้ทางเลือกที่น่าจะเป็นไปได้มากที่สุด (ค่า cost ต่ำที่สุด) มาเป็นอันดับแรก ตัวอย่างเช่น ข้อความที่ป้อนเข้าไปเป็น “ กีฬาเป็นการออกกำลังกายอย่างหนึ่ง ” จะได้ผลทางเลือกที่เป็นไปได้จัดเรียงตามค่า cost ดังนี้

ผลจากการทำการย้อนกลับ	ค่า cost
กีฬา/เป็น/การออกกำลังกาย/อย่างหนึ่ง	4
กีฬา/เป็นการ/ออก/กำลังกาย/อย่างหนึ่ง	5
กีฬา/เป็นการ/ออกกำลัง/กาย/อย่างหนึ่ง	5
กีฬา/เป็นการ/ออกกำลัง/กาย/อย่าง/หนึ่ง	6
กีฬา/เป็นการ/ออกกำลัง/กา/ยอ/ย่างหนึ่ง	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

ตารางที่ 2-5 ผลการตัดคำด้วยวิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด (วิรัช ศรเลิศล้ำวาณิช, 2538)

ขั้นตอนวิธีนี้จะเลือกการตัดคำที่มีค่า cost ต่ำที่สุด ซึ่งก็ขึ้นอยู่กับจำนวนคำที่ตัดออกมาได้ และจำนวนคำที่ไม่มีในพจนานุกรม อย่างไรก็ตาม หากมีทางเลือกที่มีค่า cost เท่ากัน และมีจำนวนคำเท่ากันมากกว่าหนึ่งทางเลือก ขั้นตอนวิธีจะไม่สามารถตัดสินใจได้ว่าจะเลือกทางเลือกไหน ดังนั้นจึงต้องใช้ heuristic อื่นเข้ามาช่วย โดยส่วนใหญ่มักใช้วิธีการเทียบคำที่ยาวที่สุดเข้ามาช่วย ตัวอย่าง เช่น ข้อความป้อนเข้า “ ตากลม ” จะมี 2 ทางเลือกที่มีจำนวนคำน้อยที่สุดเท่ากัน คือ

ตาก (expose) ลม (wind)

ตา (eye) กลม (round)

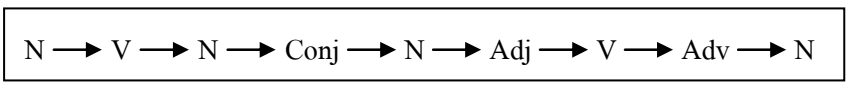
อัลกอริทึมจะเลือกทางเลือกแรกเป็นคำตอบ เนื่องจากคำแรกของทางเลือกแรก “ตาก” มีความยาวมากกว่าคำแรกของทางเลือกที่สอง “ตา”

7.1.3.3 วิธีตำคำแบบ Left Search Matching

วิธีการแยกคำประโยคภาษาไทยวิธีหนึ่ง ที่กล่าวโดย ไพฑูรย์ นุชแจ้ง และชม กิ้มปาน (2545) โดยอาศัยการตรวจสอบกับคำในพจนานุกรมกับประโยคที่มีคำที่ไม่รู้หรือคำที่ไม่มีในพจนานุกรม ซึ่งพิจารณาคำของประโยคจากทางซ้ายมือไปทางขวามือทีละหน่วยคำตามลำดับ โดยที่พิจารณาอักขระ 2 ตัวแรกของหน่วยคำก่อนแล้วจึงตรวจสอบกับคำศัพท์กับพจนานุกรม ที่อยู่ในกลุ่มอักขระ 2 ตัวแรกอันเดียวกัน ถ้าพบคำศัพท์ก็แยกออกจากประโยค Unknown ในกรณีที่สามารถแยกคำศัพท์ออกจากประโยคที่มีคำที่ไม่รู้ ได้มากกว่า 1 คำ ในอักขระ 2 ตัวแรกเดียวกันให้ทำเครื่องหมายไว้ (!*) เพื่อจะกลับมาพิจารณาคำตรงตำแหน่งนี้ใหม่ แล้วทำตรวจสอบคำต่อไปจนจบประโยค

7.1.3.4 วิธี ไบแกรม

วิธีการตัดคำที่ผู้วิจัยเลือกนำมาใช้ในงานวิจัยฉบับนี้ คือ วิธี ไบแกรม ดังที่กล่าวใน (ไพฑูรย์ นุชแจ้ง , ชม กิ้มปาน, 2545) ว่า เทคนิค ไบแกรม คือ วิธีการในการดูรูปแบบของประโยคว่าประเภทของคำต่าง ๆ มีการเรียงคำตามหลักไวยากรณ์อย่างไร โดยพิจารณาคำในประโยคทีละ 2 คำ ดังรูปที่ 2-6 โดยต้องใช้ลำดับหน่วยคำของประโยคฐานข้อมูล แล้วใช้สูตรความน่าจะเป็นในรูปคณิตศาสตร์ตามสมการที่ 2-1 เก็บค่าของลำดับหน่วยคำที่พบป้อนเข้ามาเพื่อไว้ใช้เทียบกับคำที่ไม่รู้ซึ่งได้จากการตัดคำแบบ Left Search Matching ว่าประโยคผลลัพธ์ใดถูกต้อง โดยประโยคที่ถูกต้องจะมีค่าของผลคูณของความน่าจะเป็นแบบ ไบแกรม ที่เทียบกับฐานข้อมูลแล้วทั้งประโยคมีค่าสูงสุด



ภาพประกอบที่ 2-6 การพิจารณาคำแบบ ไบแกรม (ไพฑูรย์ นุชแจ้ง และชม กิ้มปาน, 2545)

สูตรความน่าจะเป็นของ ไบแกรม

$$\text{Prtob}(X/Y) = \frac{\text{Count(Y at position i-1 and X at i)}}{\text{Count(Y at position i-1)}} \quad \text{....(2-1)}$$

ก่อนพิจารณาคำที่ไม่รู้ โดยเราจำเป็นต้องป้อนข้อมูลลงในฐานข้อมูลก่อน ด้วยโปรแกรมเรียนรู้และจดจำ ว่ารูปแบบประโยคที่ถูกต้องทางไวยากรณ์ที่พิจารณาแบบ ไบแกรม (พิจารณาทีละ 2 หน่วยคำ) มีว่าอย่างไร โดยใช้สมการที่ 2-1 ช่วยเก็บข้อมูล ซึ่งแสดงผลของการเก็บข้อมูลแบ ไบแกรม ดังตารางที่ 2-5

ตัวอย่างประโยคที่เก็บเป็นฐานข้อมูลอ้างอิง

เด็กไปโรงเรียน = /V/N

ม้าชอบกินหญ้า = N/V/V/N

เขาเป็นคนดีที่ฉันพึงปราถนา = Pron/V/N/Adj/Pron/Pron/Adj/V

Category	Count_i	Pair	Count_ii	Estimate	totalword
@	50	@Adj	24	0.48	2.46
Adj	16		2		
N	51		2		
N	51		1		
V	58		11		
V	58		10		

ตารางที่ 2-5 ฐานข้อมูล Knowledge Base ของการตัดคำแบบ ไบแกรม
(ไพฑูรย์ นุชแจ้ง และชม กิ้มปาน, 2545)

Category = ประเภทของคำโดยที่ (N = คำนาม , V = กริยา , Adj = คำวิเศษณ์ , Pron = สรรพนาม , Prep = บุพบท , Conj = สันธาน , Interj =)

Count_i = จำนวนครั้งที่พบประเภทของคำนั้น ๆ ในขณะป้อนประโยคฐานข้อมูล

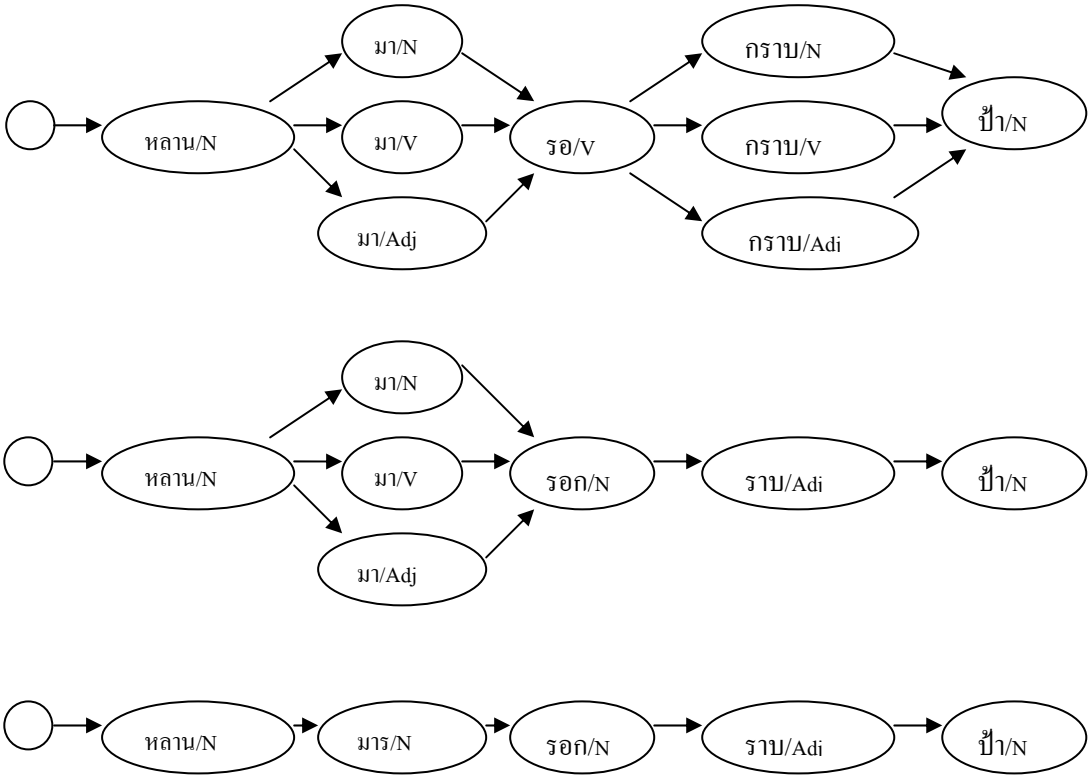
Pair = ลำดับประเภทของคำข้อมูลที่ป้อนเข้ามาเก็บแบบ 2 คำเรียงลำดับกัน

Count_ii = จำนวนครั้งที่พบ pair

$$\text{Estimate} = \frac{\text{prob}(x/y) = \text{Count\_ii}}{\text{Count\_i}} \qquad \ldots(2\text{-}2)$$

Totalword = จำนวนครั้งทั้งหมดที่ป้อนในฐานข้อมูล

Viterbi algorithm ในรูปที่ 2-7 เป็นการแสดงถึงเส้นทางที่จะหาผลลัพธ์ของความน่าจะเป็นสูงสุดของแต่ละประโยคที่ได้จากการตัดคำแบบ Left search matching โดยการพิจารณาแบบ ไบแกรม จะต้องหาเส้นทางที่ทำให้ค่าที่เป็นไปได้ที่มีค่าสูงสุดของแต่ละประโยค ซึ่งเส้นทางที่ประโยคใดมีค่าที่เป็นไปได้สูงสุดของประโยคนั้นจะเป็นประโยคที่ถูกต้อง



ภาพประกอบที่ 2-7 แผนผัง Viterbi Algorithm (ไพฑูรย์ นุชแจ้ง และชม กิ้มปาน, 2545)

ตัวอย่างการคำนวณและพิจารณาแบบ ไบแกรม จากประโยคการตัดคำแบบ left search matching โดยใช้ฐานข้อมูลจากตารางที่ 2-5 มาช่วยพิจารณา ผลลัพธ์ของประโยคที่ถูกต้องจะต้องมีค่าผลคูณของความน่าจะเป็นรวมของทั้งประโยคสูงสุด เช่น ตัวอย่าง

หลาน/มา/รอ/กราบ/ป้า = N/V/ V/V/N

$$\text{Prob(NVVVN)} = \text{Prob(N|@)} * \text{Prob(V|N)} * \text{Prob(V|V)} * \text{Prob(V|V)} * \text{Prob(N|V)}$$

$$= 0.48 * 0.39 * 0.09 * 0.09 * 0.45$$

$$= 6.82*10^{-0.4}$$

วิธีแบบ ไบแกรม จะใช้เป็นตัวตัดสินผลลัพธ์ประโยคภาษาไทยของวิธีการตัดคำแบบ Left Search Matching ที่มีผลลัพธ์ประโยคมากกว่า 1 ผลลัพธ์ได้ดี โดยยึดหลักไวยากรณ์ภาษาไทยเป็นหลักในการพิจารณาและคาดคะเน เพื่อจะให้ผลลัพธ์ประโยคภาษาไทยของการตัดคำแบบ Left Search Matching เหลือเพียงผลลัพธ์เดียวได้

โดยการตัดคำแบบอาศัยพจนานุกรมนี้ จะต้องมีการสืบค้นหาคำศัพท์ในพจนานุกรมเป็นจำนวนมาก ทำให้ต้องกล่าวถึงโครงสร้างของพจนานุกรมในโปรแกรมที่ใช้ในงานวิจัยนี้ โดยโปรแกรม SWATh ที่ผู้วิจัยเลือกใช้นั้น โครงสร้างของพจนานุกรมที่เก็บคำศัพท์ในการตัดคำนั้น คือ “ โครงสร้างข้อมูลแบบทรัย ” ซึ่งจากการนำโครงสร้างแบบทรัยเข้ามาใช้สามารถลดขนาดของพจนานุนกรมได้ และนอกจากนี้โครงสร้างแบบทรัยนี้ยังสามารถสืบค้นหาคำศัพท์ได้อย่างรวดเร็วและสามารถจะเพิ่มเติมคำศัพท์ได้อย่างสะดวกและรวดเร็วด้วย โดยรายละเอียดของโครงสร้างพจนานุกรมนั้น ผู้วิจัยขอกล่าวถึงในต่อไป [1]

7.2 การตัดคำภาษาไทย

ระบบการค้นคืนสารสนเทศภาษาไทยนั้น ดัง อรุณี โอฬารานนท์ (2546) กล่าวไว้ว่า จำเป็นต้องมีระบบตัดคำที่อยู่ในประโยคให้แยกจากกัน ทั้งนี้เพราะคำในภาษาไทยมีทั้งคำโดดและคำประกอบที่สร้างจากคำโดดและคำประกอบมักจะมีความหมายแตกต่างไปจากคำโดดที่นำมาประกอบหากไม่ตัดคำแล้วการค้นคืนอาจไม่ถูกต้อง ขั้นตอนนี้ว่าการตัดคำหรือการแบ่งแยกคำการตัดคำภาษาอังกฤษกระทำได้ง่าย เนื่องจากแต่ละประโยคแบ่งออกเป็นหน่วยของคำอยู่แล้ว โดยใช้ช่องว่างหรือเครื่องหมายวรรคตอนมาเป็นตัวคั่นระหว่างคำ แต่สำหรับภาษาไทยนั้นแต่ละ

---

[1] ดูโครงสร้างข้อมูลแบบทรัย ในบทที่ 2 หน้า 42-43

ประโยคจะเขียนติดกันทั้งหมด และหน่วยคำต่าง ๆ ในแต่ละประโยคก็จะเขียนติดกันทั้งหมดโดยไม่มีเครื่องหมายวรรคตอนใด ๆ มาเป็นตัวแบ่งแยกระหว่างหน่วยคำแต่ละหน่วย ดังนั้นจึงทำให้การตัดคำภาษาไทยจึงเป็นเรื่องค่อนข้างยุ่งยาก

วิธีการตัดคำภาษาไทยด้วยคอมพิวเตอร์ก็คือการหาขอบเขตของคำ แล้วเปรียบเทียบแต่ละคำกับพจนานุกรมซึ่งจะต้องมีคำศัพท์เก็บไว้เป็นจำนวนที่มากพอและต้องสามารถเพิ่มศัพท์ใหม่ลงในพจนานุกรมได้ด้วยจึงจะสามารถตอบสนองความต้องการในการใช้งานได้ เพราะเมื่อเวลาผ่านไปก็จะมีคำศัพท์ใหม่เกิดขึ้นมาตามยุคตามสมัย ปัญหาอีกประการหนึ่งของการตัดคำคือ เอกสารข้อความต่าง ๆ มีการใช้คำเฉพาะเช่นชื่อคน ชื่อสถานที่ คำผสมต่าง ๆ ทำให้ในทางปฏิบัติไม่สามารถบรรจุคำศัพท์ที่จะใช้ประกอบการตัดคำลงไปในพจนานุกรมได้ทั้งหมด ฉะนั้นจำเป็นที่จะต้องมีวิธีการอื่น ๆ มาประกอบในการตัดคำ

7.3 การเลือกประโยคที่ถูกต้องหลังการตัดคำ

วิธีการตัดคำบางวิธีจะให้ผลลัพธ์เป็นรูปแบบการตัดคำทั้งหมดที่เป็นไปได้ หรือมีทางเลือกของการตัดคำที่ได้มากกว่า 1 ทางเลือก ดังนั้นจึงจำเป็นต้องมีการเลือกรูปแบบการตัดคำที่คาดว่าถูกต้องที่สุด โดยอาจเป็นการใช้ข้อมูลความถี่ของการใช้คำภาษาไทย เข้ามาช่วยในการเลือกประโยคที่ถูกต้องหลังการตัดคำ โดยคำนวณหาค่าความน่าจะเป็นของการนำคำนั้นไปใช้ในภาษาไทย (probability of usage) ซึ่งอาศัยแหล่งข้อมูลภาษาไทยต่าง ๆ เพื่อเป็นฐานข้อมูลตัวอย่าง การคำนวณหาค่าความน่าจะเป็นของการนำคำนั้นไปใช้ในภาษาไทยสามารถหาได้จากสมการที่ 2-3

$$Pu(W_1) = f_1 / N \qquad \ldots(2\text{-}3)$$

โดยกำหนดให้

$Pu(W_1)$ หมายถึง ความน่าจะเป็นที่คำภาษาไทย W_1 จะถูกใช้ในภาษา

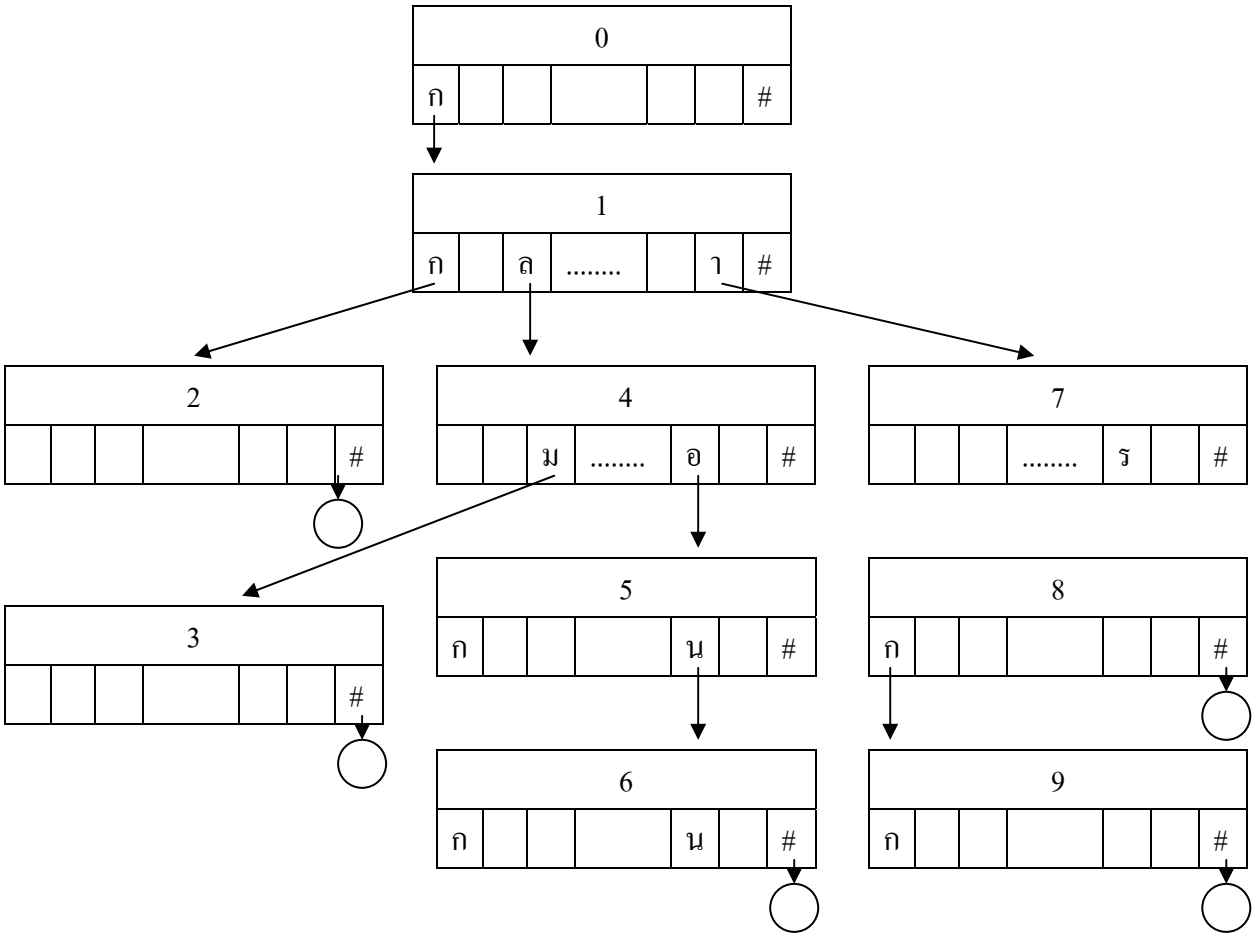
f_1 หมายถึง ความถี่หรือจำนวนครั้งของการใช้คำภาษาไทย W_1 ที่ปรากฏในคลังข้อมูล

N หมายถึง จำนวนคำทั้งหมดที่ปรากฏในคลังข้อมูล (คือ ความถี่หรือจำนวนครั้งที่ปรากฏของทุกคำในคลังข้อมูลรวมกัน)

การเลือกรูปแบบการตัดคำที่ถูกต้องสามารถทำได้โดย เปรียบเทียบค่า Pu ของคำที่อยู่ในลำดับเดียวกันในทางเลือกการตัดคำต่าง ๆ ที่ได้ แล้วเลือกทางเลือกที่มีค่าสูงกว่า หากคำในลำดับต้นมีค่าความน่าจะเป็นของการใช้เท่ากัน ก็เปรียบเทียบคำในลำดับถัดไปเรื่อย ๆ

8. โครงสร้างข้อมูลแบบทรัย

โครงสร้างข้อมูลแบบทรัย (Corman, Leiserson and Rivest, 1990: Frakes and Baeza-Yates, 1992) จะมีลักษณะคล้ายกับโครงสร้างข้อมูลแบบต้นไม้ แต่วิธีการจัดเก็บข้อมูลจะแตกต่างกัน โดยที่โครงสร้างข้อมูลแบบทรัยนี้จะจัดเก็บตัวอักษรของคำศัพท์ ซึ่งโครงสร้างข้อมูลแบบต้นไม้นั้นจะจัดเก็บข้อมูลทั้งคำ สำหรับโครงสร้างข้อมูลแบบทรัย แสดงในรูปที่ 2-7



ภาพประกอบที่ 2-8 โครงสร้างข้อมูลแบบทรัย (ไพศาล เจริญพรสวัสดิ์, 2542)

จากรูป 2-8 โครงสร้างของทรัยจะประกอบไปด้วยโหนดต่าง ๆ โดยที่ข้อมูลภายใน 1 โหนดจะประกอบไปด้วย พอยเตอร์ที่ชี้ไปยังโหนดของตัวอักษรถัดไป ซึ่งมีจำนวนพอยเตอร์เท่ากับจำนวนตัวอักขระที่จะอนุญาตให้มีได้ในพจนานุกรมบวกกับอักขระที่ใช้ระบุเป็นตัวจบคำศัพท์ (Terminator) อีก 1 ตัวอักษร ซึ่งสัญลักษณ์ที่ใช้ในนี้คือเครื่องหมาย #

สำหรับการสืบค้นในโครงสร้างข้อมูลแบบทรัยนี้จะทำโดย เริ่มต้นที่โหนด 0 ถ้าต้องการค้นหาคำศัพท์ให้นำอักษรทีละตัวจากคำศัพท์ที่ต้องการ มาดูว่าภายในโหนด 0 ถ้าต้องการค้นหาคำศัพท์ให้นำอักษรทีละตัวจากคำศัพท์ที่ต้องการ มาดูว่าภายในโหนด 0 นั้นมีพอยเตอร์ของตัวอักษรที่ต้องการชี้ไปโหนดอื่นหรือไม่ ถ้าไม่มีแสดงว่าคำนั้นไม่มีอยู่ในพจนานุกรม แต่ถ้ามีพอยเตอร์ที่ชี้ไปที่โหนดถัดไปก็ให้เดินที่โหนดที่พอยเตอร์นั้นชี้ไป แล้วนำตัวอักษรตัวถัดไปมาทำตามขั้นตอนแบบเดิมจนเหมด เมื่อนำตัวอักษรทั้งหมดจากคีย์มาเดินในทรัยแล้ว ให้เดินด้วยอักษร “ # ” แล้วดูว่าค่าพอยเตอร์มีค่าเท่ากับค่าว่าง (null) หรือไม่ ถ้าเท่าแสดงว่าไม่มีคำศัพท์นั้นใจพจนานุกรม แต่ถ้าไม่เท่ากันก็แสดงว่ามีคำศัพท์นั้นอยู่ในพจนานุกรม โดยพอยเตอร์นี้ส่วนใหญ่จะไปชี้ที่ตำแหน่งของข้อมูลของคำนั้น

ตัวอย่างการลืบค้นคำศัพท์จากโครงสร้างข้อมูลแบบทรัย จากรูปที่ 2-7 ถ้าต้องการลืบค้นคำว่า “ กลม ” มีขั้นตอนดังนี้คือ โหนด 0 จะเป็นโหนดเริ่มต้น ดังนั้นนำตัวอักษร “ ก ” เข้ามาเดินภายในทรัยก็จะไปที่โหนด 1 หลังจากนั้นก็นำตัวอักษร “ ล ” เข้ามาเดินต่อไปที่โหนด 4 แล้วก็นำตัวอักษรตัวถัดไปคือ “ ม ” เข้ามาเดินจะไปที่โหนด 3 สุดท้ายเมื่อทำการค้นหามาถึงตัวอักษรสุดท้ายของคีย์แล้ว ให้เดินด้วย “ # ” ซึ่งค่าที่ได้ไม่เท่ากับค่าว่างแสดงว่าคำว่า “ กลม ” มีอยู่ในพจนานุกรม

9. โปรแกรม SWATh

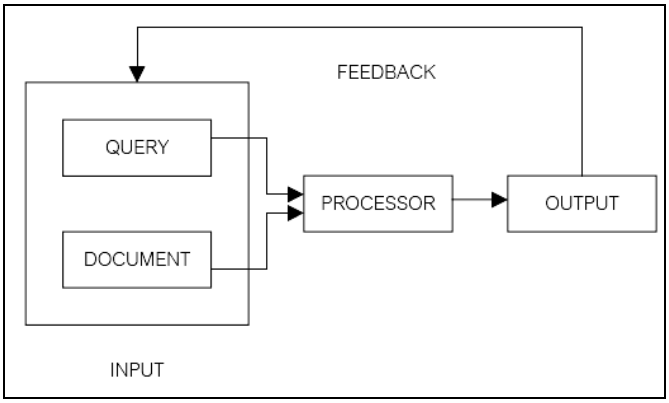
โปรแกรมแยกคำ SWATh (Smart Word Analysis for THai) ซึ่งพัฒนาโดยหน่วยปฏิบัติการวิจัยและพัฒนาวิศวกรรมภาษาและซอฟต์แวร์ ซึ่งเป็นหลักการที่อาศัยวิทยาการศึกษาสำนึกและวิธีการทางสถิติเข้ามาช่วยแก้ปัญหาความกำกวมในการตัดคำ ซึ่งวิธีการทางสถิติที่นำมาใช้คือการใช้ค่าสถิติที่เกิดจากลำดับหน้าที่คำ หรือไวยากรณ์ทางภาษา โดยมีเทคนิค คือ การตัดคำโดยใช้หน้าที่คำแบบไบแกรมโมเดล คือ การตัดคำโดยมีการนำเอาค่าสถิติ ซึ่งพิจารณาจากความต่อเนื่องของหน้าที่คำ ส่วนวิธีการเลือกแบบตัดคำที่ดีที่สุดนั้นได้โดยหาประโยคที่มีความน่าจะเป็นมากที่สุด โดยการหาความน่าจะเป็นของแต่ละประโยค

10. การค้นคืนสารสนเทศ

จุดเริ่มต้นของการพัฒนาระบบการค้นคืนสารสนเทศในภาษาต่าง ๆ คือ การคัดเลือกคำที่เหมาะสมมาเป็นดรรชนีคำค้นแบบอัตโนมัติ คำว่าอัตโนมัติ หมายความว่าให้คอมพิวเตอร์สรรหาดรรชนีคำค้นเองโดยไม่ต้องอาศัยผู้ชำนาญการ การใช้ดรรชนีคำค้นที่ไม่เหมาะสมอาจมีผลทำให้เอกสารหรือข้อมูลดังกล่าวไม่ถูกเรียกมาใช้งานเลยหรือไม่ถูกสืบค้นจากผู้ใช้ ทั้งนี้เพราะระบบค้นคืนไม่สามารถค้นหาเอกสารดังกล่าวให้กับผู้ใช้ได้ ขณะเดียวกันการสร้างดรรชนีคำค้นที่ไม่สื่อความหมาย ผู้สืบค้นเองก็อาจจะนึกไม่ถึง หรือในบางครั้งผู้สืบค้นก็จะมีความยากลำบากในการระบุคำสืบค้นเพื่อให้ตรงกับผู้กำหนดคำสืบค้น ถึงแม้จะมีผู้ชำนาญการเป็นผู้กำหนดดรรชนีคำสืบค้นก็ตาม ในบางครั้งผู้ชำนาญการก็ไม่สามารถคัดเลือกได้ หรือการคัดเลือกไม่สามารถกำหนดเป็นมาตรฐานได้ การค้นคืนเป็นเรื่องทางการหาความน่าจะเป็น มากกว่าการที่จะได้ผลที่แน่นอนตายตัว (deterministic)

10.1 โครงสร้างของระบบการค้นคืนสารสนเทศ

Rijsbergen (1979:147) ได้อธิบายว่า ระบบการค้นคืนสารสนเทศมีองค์ประกอบพื้นฐาน 3 ส่วนคือ Input, Processor และ Output ดังภาพประกอบที่ 2-9



ภาพประกอบที่ 2-9 ระบบการค้นคืนสารสนเทศ (Rijsbergen, 1979)

การทำงานของระบบค้นคืนสารสนเทศโดยทั่วไปเริ่มต้นด้วยการจัดเก็บข้อความที่สามารถใช้เป็นตัวแทนความหมายของเอกสารทั้งฉบับ ดังนั้นในการจัดเก็บเอกสารเข้าระบบจะทำ

การคัดเลือกเฉพาะข้อความสำคัญหรือตัดข้อความที่มีประโยชน์ในการใช้งานน้อยออก เมื่อผู้ใช้ป้อนข้อมูลเข้าสู่ระบบเพื่อทำการค้นคืน ผลของการค้นคืนที่ได้จะมีส่วนช่วยในการปรับปรุงคำขอค้นคืน ซึ่งเรียกกระบวนการนี้ว่า Feedback

Processor มีหน้าที่หลักในการค้นคืนข้อมูลการจัดการในส่วนของโครงสร้างสารสนเทศให้เหมาะสมกับการใช้งาน การแจกแจงหมวดหมู่ของเอกสารการเลือกใช้กรรมวิธีในการค้นข้อมูลให้เหมาะสมกับคำค้นคืน

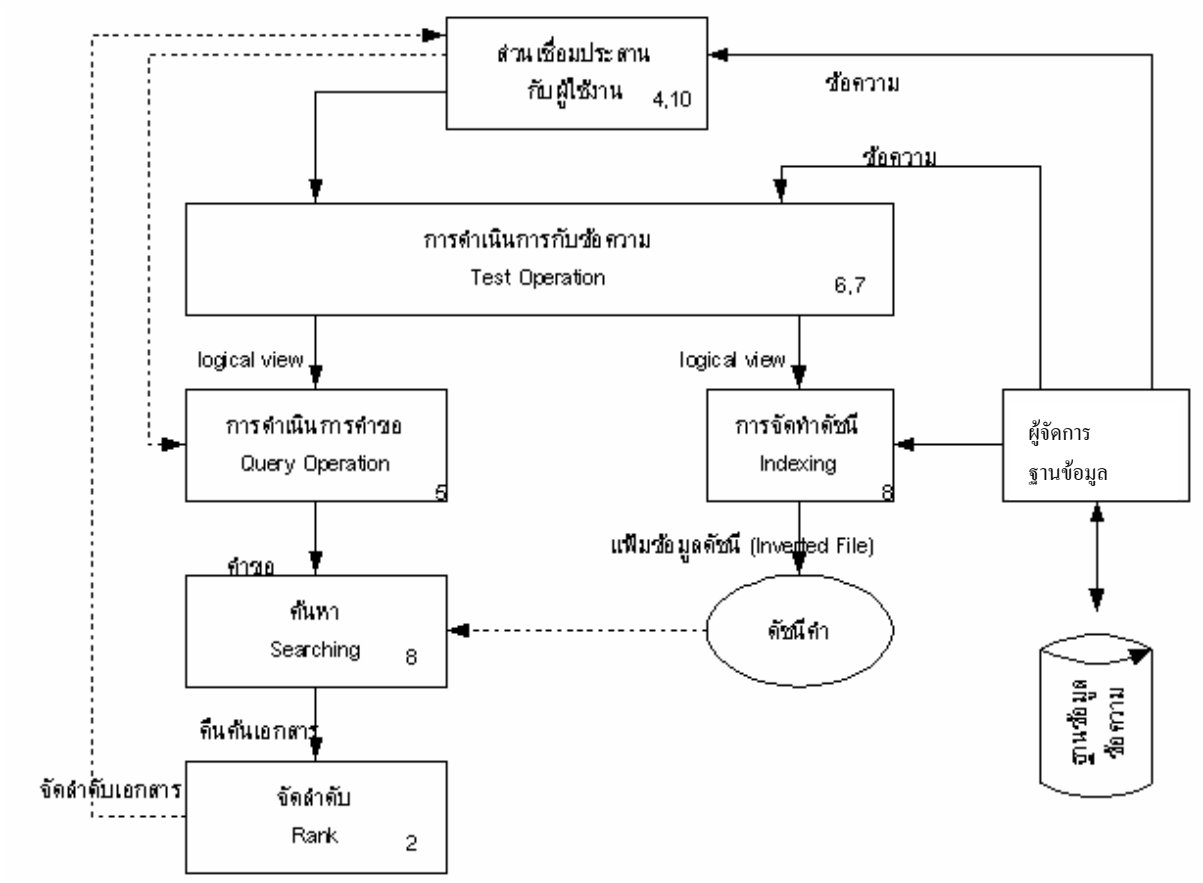
Rijsbergen (1979:147) ได้อธิบายว่าการพัฒนาระบบการค้นคืนสารสนเทศสามารถแบ่ง ออกได้เป็น 3 แนวทาง ได้แก่

1. การวิเคราะห์ข้อมูลในเอกสาร (Content Analysis) เป็นการพัฒนาในส่วนของการให้ความหมายของข้อมูลในเอกสารเพื่อนำไปสู่การประมวลผลด้วยคอมพิวเตอร์อย่างมีประสิทธิภาพ โดยในงานวิจัยนี้จะเป็นการอ้างอิงถึงการวิเคราะห์เอกสารบนเครือข่ายสารสนเทศในด้านความหมายของตำแหน่งที่อยู่ของคำหรือวลี วิเคราะห์ความหมายด้านความถี่ที่เกิดขึ้น วิเคราะห์ความหมายในด้าน “ ความสำคัญ ” ของหน้าที่ของคำตามหลักไวยากรณ์ภาษาไทยว่ามีผลต่อการระบุใจความสำคัญของเอกสารอย่างไร

2. การออกแบบโครงสร้างสารสนเทศ (Information Structures) เป็นการพัฒนาในส่วนของการสร้างความสัมพันธ์ระหว่างเอกสารเพื่อให้ได้ระบบการค้นคืนสารสนเทศที่มีประสิทธิภาพและประสิทธิผล

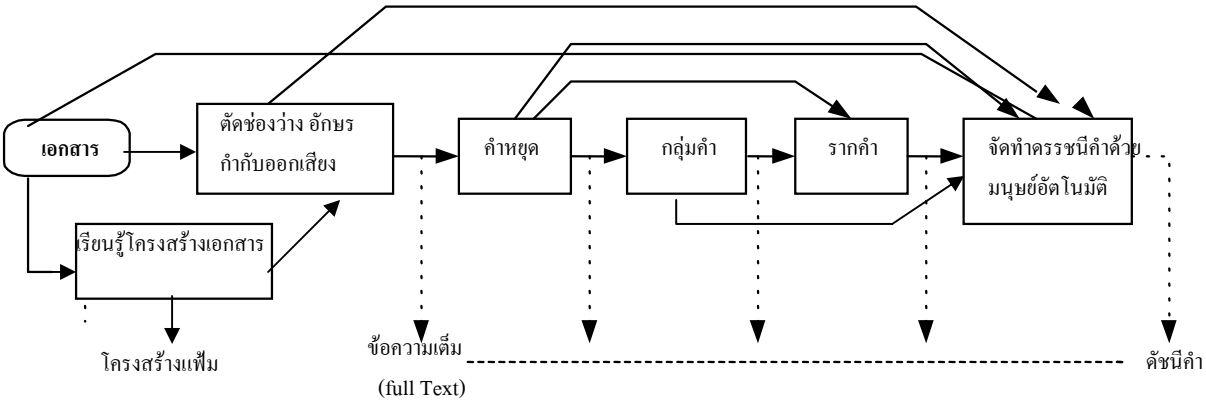
3. การประเมินประสิทธิภาพในการค้นคืน (Evaluation) เป็นการพัฒนาในส่วนของการวัดผลการทำงานของระบบการค้นคืนสารสนเทศว่าประสิทธิภาพและประสิทธิผลในการทำงานมากน้อยเพียงใด โดยในงานวิจัยนี้จะเป็นการประเมินประสิทธิภาพของการค้นคืนในด้านของค่าความแม่นยำในการสกัด และค่าความระลึกในด้านของปริมาณความของต้องของการค้นคืน

10.2 กระบวนการของการค้นคืนสารสนเทศ



ภาพประกอบที่ 2-10 กระบวนการค้นคืนสารสนเทศ (ชุลีรัตน์ จรัสกุลชัย,2549)

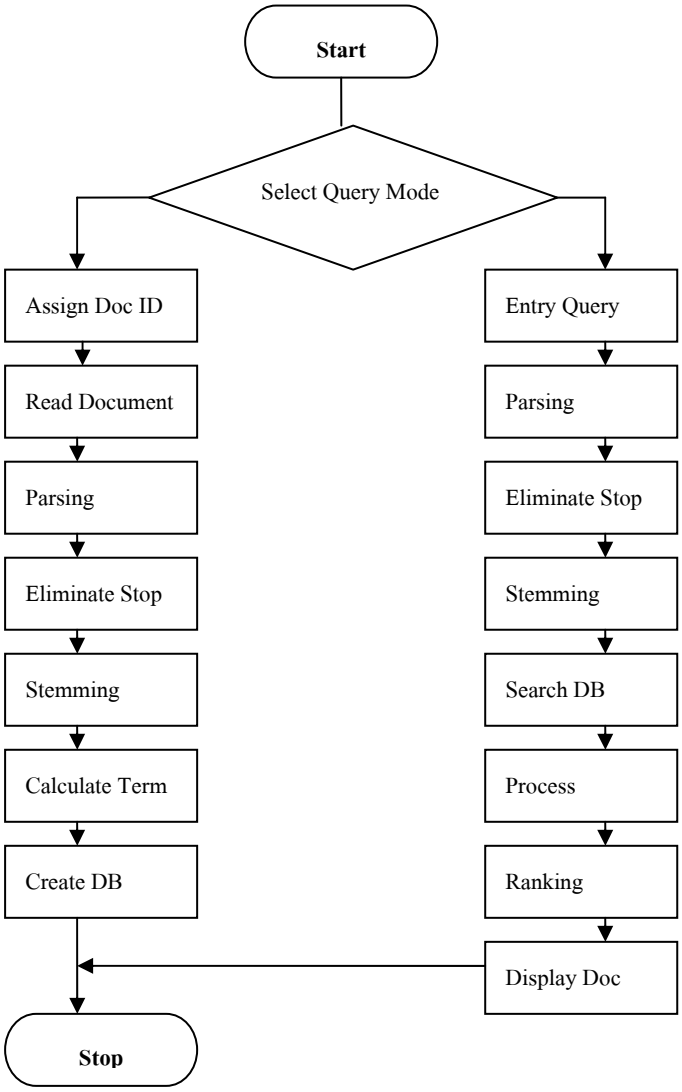
10.3 กระบวนการในการหาตัวแทนเอกสาร



ภาพประกอบที่ 2-11 กระบวนในการหาตัวแทนเอกสาร (ชุลีรัตน์ จรัสกุลชัย,2549)

10.4 ขั้นตอนการทำงานในระบบการค้นคืนสารสนเทศ

Frakes และคณะ(1992) ได้อธิบายว่าระบบการค้นคืนสารสนเทศ โดยทั่วไปประกอบด้วยขั้นตอนสำคัญ 2 ขั้นตอน คือ ขั้นตอนการจัดเก็บเอกสาร และขั้นตอนการค้นคืนเอกสาร โดยทั้ง 2 ขั้นตอนนี้จะมีกระบวนการย่อยเกิดขึ้นหลายกระบวนการ ตัวอย่างของระบบการค้นคืนสารสนเทศที่มีการค้นคืนแบบการใช้ตรรกศาสตร์ (Boolean Base Information Retrieval) สามารถแสดงเป็นผังงาน (FlowChart) ได้ดังภาพที่ 2-12



ภาพประกอบที่ 2-12 ขั้นตอนการทำงานของระบบการค้นคืนสารสนเทศ

Frakes และคณะ(1992)

ขั้นตอนการทำงานของระบบการค้นคืนเอกสารสารสนเทศตามภาพ 2-12 มีรายละเอียด ดังต่อไปนี้

10.4.1 เลือกโหมดการทำงาน

10.4.2 จัดเก็บเอกสาร ขั้นตอนนี้จะเริ่มด้วยการกำหนดหมายเลขเอกสาร (Assign Document ID) แล้วอ่าน ข้อความในเอกสาร (Read Document File) มากระจายคำ แล้วคัดเลือกคำที่ตัดได้ว่าคำใดมีประโยชน์ต่อการนำไปใช้ในการค้นคืนมาก และคำใดมีประโยชน์ต่อการค้นคืนน้อย(Eliminate Stop Word) ส่วนคำที่ไม่อยู่ในรายนามคำหยุดจะนำไปเข้ากระบวนการหาแกนคำ (Stemming) เพื่อทำการจัดรูปแบบของคำให้อยู่ในรูปแกนคำ จากนั้นจะนำคำที่ได้ไปคำนวณหาค่าความถี่ของของคำ (CalculateTerm Frequency) เพื่อหาความสำคัญของคำแต่ละคำในเอกสารและจัดสร้างเป็นฐานข้อมูล (Create Database) ซึ่งเก็บดรรชนีค่าพร้อมกับอ้างอิงไปยังเอกสารต้นฉบับที่ได้ถูกกำหนดเป็นรหัสไว้ เพื่อให้ง่ายต่อการอ้างอิงเอกสารต้นฉบับเพื่อใช้ในขั้นตอนการค้นคืน

10.4.3 ขั้นตอนการค้นคืนเอกสาร การค้นคืนเอกสารจะเริ่มต้นด้วยการรับประโยคคำขอสืบค้นคืนจากผู้ใช้ (Entry Query) ประโยคคำขอค้นคืนจากผู้ใช้จะเข้าสู่กระบวนการตัดคำ แล้วนำคำที่ได้ไปตรวจสอบกับรายนามคำหยุด คำใดที่ไม่อยู่ในรายนามคำหยุดก็จะนำไปเข้ากระบวนการหาแกนคำ แล้วนำคำที่ไปทำการค้นหาข้อมูลจากฐานข้อมูลดัชนี (Search Database) ก็จะได้เอกสารที่เกี่ยวข้องกับคำขอสืบค้นที่ผู้ใช้ต้องการ และนำเอกสารที่ได้เข้าสู่กระบวนการทางตรรกศาสตร์(Process Boolean Operation) เพื่อตรวจสอบเงื่อนไขในการเลือกเอกสารของผู้ใช้ แล้วทำการจัดลำดับความสำคัญของเอกสาร (Ranking) เพื่อนำเสนอเอกสารให้กับผู้ใช้เรียงตามลำดับตามความใกล้เคียงระหว่างคำขอสืบค้นกับเนื้อหาในเอกสาร (Display Documents)

10.5 รูปแบบของระบบการค้นคืนสนเทศ มี 3 รูปแบบ ดังนี้

10.5.1 แบบบูลีน (Boolean) ถูกแทนในรูปของ set หรือ ใช้หลักการของเซตทางคณิตศาสตร์

10.5.1.1 ฟัซซี่ (Fuzzy)

10.5.1.2 บูลีนแบบขยาย

10.5.2 แบบเวคเตอร์ (Vector) ถูกแทนด้วยเวคเตอร์หลายมิติ หรือใช้หลักของพีชคณิต (algebraic)

10.5.2.1 เวคเตอร์แบบทั่วไป (Generalized vector)

10.5.2.2 Latent semantic indexing

10.5.2.3 Neural Network

10.5.3 แบบสถิติ (Probabilistic) ถูกแทนด้วยหลักของสถิติ

10.5.3.1 เครือข่ายพยากรณ์ (Inference Network)

10.5.3.2 เครือข่ายความเชื่อ (Belief Network)

โดยงานวิจัยนี้ เป็นรูปแบบระบบการค้นคืนแบบเวกเตอร์ คือ อาศัยเทคนิคของโครงข่ายประสาทเทียม

11. การสกัดสารสนเทศ

เทคนิคการสกัดสารสนเทศ นั้น ธนดล วัฒนาสุทธิวงศ์ (2003) ได้กล่าวว่า หมายถึง การคัดเลือกข้อมูลออกมาใช้งานในส่วนที่เราต้องการ การสกัดข้อสนเทศนั้นเป็นส่วนหนึ่งของการประมวลผลภาษาธรรมชาติซึ่งมุ่งเน้นในเรื่องของการหาและกำกับความหมายของคำจากเอกสาร โดยการให้ความหมายของคำในเอกสารนี้คือการหาข้อเท็จจริงจากเอกสาร ที่สามารถมองข้อเท็จจริงเหล่านี้เป็นเอนทิตี้ ซึ่งมีความสัมพันธ์กับข้อเท็จจริงอื่น

ข้อเท็จจริงเหล่านี้จะขึ้นอยู่กับโดเมนที่มีโครงสร้างคงที่แน่นอน ซึ่งจะทำให้ข้อเท็จจริงเหล่านี้สามารถถูกเก็บอยู่ในฐานข้อมูลแบบสัมพันธ์ได้ ระบบการสกัดข้อสนเทศนั้น อาจแบ่งได้เป็น 2 แนวทางใหญ่ๆ คือ แนวทางที่ใช้การวิเคราะห์ทางภาษา และ แนวทางที่ไม่ใช้การวิเคราะห์ทางภาษา โดยที่อุปสรรค์สำคัญในการสกัดข้อสนเทศในปัจจุบันคือ ความเป็นไปได้ และประสิทธิภาพของระบบ

- ความเป็นไปได้ - ปัญหาทางด้านความสามารถของระบบสกัดข้อสนเทศที่จะถูกนำไปใช้ในโดเมนอื่น เนื่องจากระบบสกัดข้อสนเทศโดยทั่วไปสร้าง

- ประสิทธิภาพของระบบ – ข้อเท็จจริงส่วนใหญ่นั้นจะสามารถถูกสกัดได้จากรูปแบบที่ซ้ำกัน แต่ข้อเท็จจริงในส่วนที่เป็นส่วนน้อย จะมีรูปแบบที่หลากหลายและยากต่อการสกัด ซึ่งทำให้ประสิทธิภาพลดลง

ในปัจจุบันนั้น การสกัดข้อสนเทศนั้นถูกแบ่งออกเป็น 5 ระดับ คือ

1. Name Entity (NE) – การค้นหาคำนามเฉพาะในเอกสาร
2. Template Element (TE) – การเลือกและเติมเทมเพลตที่เกี่ยวข้องกับเอกสารที่ต้องการ
3. Co-Reference (CO) – การหาคำที่แตกต่างกันแต่แสดงถึงเอ็นทิตี้ชนิดเดียวกัน

4. Template Relation (TR) – การหาความสัมพันธ์ระหว่างเอ็นทิตี้
5. Scenario Template (ST) – การหาความสัมพันธ์ของเหตุการณ์