บทที่ 2

ภาษาไทยกับระบบสืบค้นข้อมูลโดยใช้แฟ้มข้อมูลผกผัน

เนื่องด้วยภาษาไทยมีลักษณะการเขียนแตกต่างจากภาษาอังกฤษ โดยเฉพาะมีการเขียนคำติดกัน และเนื่องด้วยระบบสืบค้นข้อมูล ที่ใช้แฟ้มข้อมูลผกผัน ต้องนำเอาคำที่ได้จากเอกสารมาทำดัชนี การแบ่งคำในภาษาไทยมีอยู่ด้วยกัน 2 วิธี คือการใช้กฎ และการใช้พจนานุกรม การใช้กฎในการแบ่งคำนั้นสิ่งที่ได้คือพยางค์ [1], [13], [15] ถึงแม้ว่าวิธีนี้จะมีความถูกต้องสูง แต่ไม่เหมาะสำหรับการนำมาทำดัชนี ซึ่งจะต้องใช้เป็นคำ อีกวิธีหนึ่งคือการแบ่งคำโดยใช้พจนานุกรม คำที่ได้จากการใช้พจนานุกรมสามารถใช้วิธีเลือกเฉพาะคำมีลักษณะเป็นคำยาว [16] มีจำนวนคำน้อย [17] หรือใช้วิธีทางสถิติ [7] แต่การใช้พจนานุกรมในการแบ่งคำมีข้อเสียเช่นกัน เช่น พจนานุกรมนั้นไม่สามารถจัดเก็บคำได้ทั้งหมด เนื่องจากคำในภาษาไทยมีคำเกิดขึ้นใหม่ได้อยู่เสมอ

การแบ่งพยางค์โดยใช้กฎ

ในการค้นหาจุดแบ่งพยางค์นั้น อัลกอริทึมจะทำงานโดยใช้กฎจำนวนหนึ่งที่ผู้พัฒนารวบรวมขึ้นมา เรียกว่า กฎแบ่งพยางค์ ตัวอย่างเช่น กฎการแบ่งพยางค์ที่ใช้ในโปรแกรมซียูไรท์เตอร์ (ภาคผนวก ก.) ยกตัวอย่างข้อความ "นายเจมส์มาร์ตินต้องการผลิตรายการโทรทัศน์" ผลที่ได้จากการแบ่งพยางค์ คือ นาย + เจมส์ + มาร์ + ติน + ต้อง + การ + ผลิต + ราย + การ + โทร + ทัศน์ เป็นต้น

ปัญหาของอัลกอริทึมแบ่งพยางค์ด้วยกฎ คือความซับซ้อน การที่จะแบ่งพยางค์ได้ดีนั้นหมายความว่า จะต้องมีกฎแบ่งพยางค์ที่ละเอียด มีจำนวนมาก และต้องมีการจัดการคำยกเว้นซึ่งมีอยู่มากตามไปด้วย คือในตัวโปรแกรมจะเต็มไปด้วยการจัดการกับกรณีต่างๆ และเต็มไปด้วยความซับซ้อน

การแบ่งคำด้วยพจนานุกรม

การแบ่งคำด้วยพจนานุกรม

การแบ่งคำโดยวิธีนี้จะใช้พจนานุกรมซึ่งโดยปกติแล้วจะเก็บอยู่ในหน่วยความจำของเครื่อง ขนาดของพจนานุกรมจะเป็นตัวกำหนดว่าการแบ่งคำโดยวิธีนี้จะมีประสิทธิภาพมากเพียงใด ถ้าพจนานุกรมเก็บคำไว้มาก จะสามารถแบ่งคำได้มาก แต่จะใช้หน่วยความจำในการเก็บมากด้วยเช่นกัน ถ้าพจนานุกรมเก็บคำไว้น้อย จะทำให้ประหยัดหน่วยความจำได้ แต่จะสามารถแบ่งคำได้น้อย

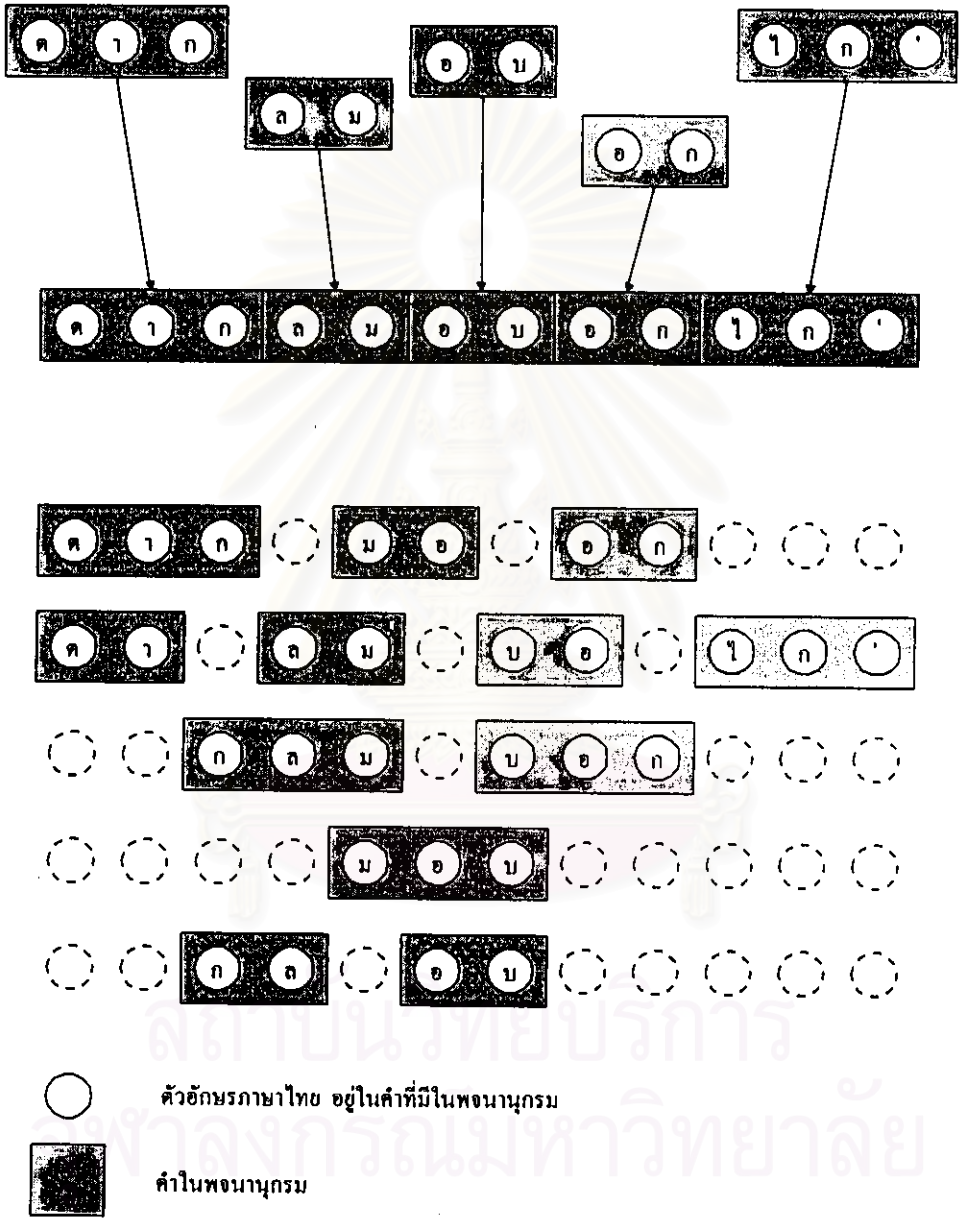
ในการแบ่งคำโดยใช้พจนานุกรมนั้น ถ้าคำเหล่านั้นไม่มีอยู่ในพจนานุกรมของโปรแกรม เช่น คำที่เป็นชื่อคน คำเฉพาะต่าง ๆ คำทับศัพท์ภาษาต่างประเทศ รวมถึงคำที่สะกดผิด จะไม่สามารถทำการแบ่งคำนั้นได้

การใช้พจนานุกรมในการดึงคำภาษาไทย

การใช้พจนานุกรมในการดึงคำภาษาไทย จะแบ่งออกเป็น 2 กรณี คือในกรณีที่ประโยคเหล่านั้นเกิดจากคำที่อยู่ในพจนานุกรมทั้งหมด กับประโยคที่เกิดจากการผสมกันระหว่างคำที่อยู่ในพจนานุกรม กับคำที่ไม่อยู่ในพจนานุกรม ในกรณีแรกจะมีอยู่ 1 รูปแบบ ส่วนกรณีที่ 2 จะมีอยู่ 5 รูปแบบ ดังต่อไปนี้

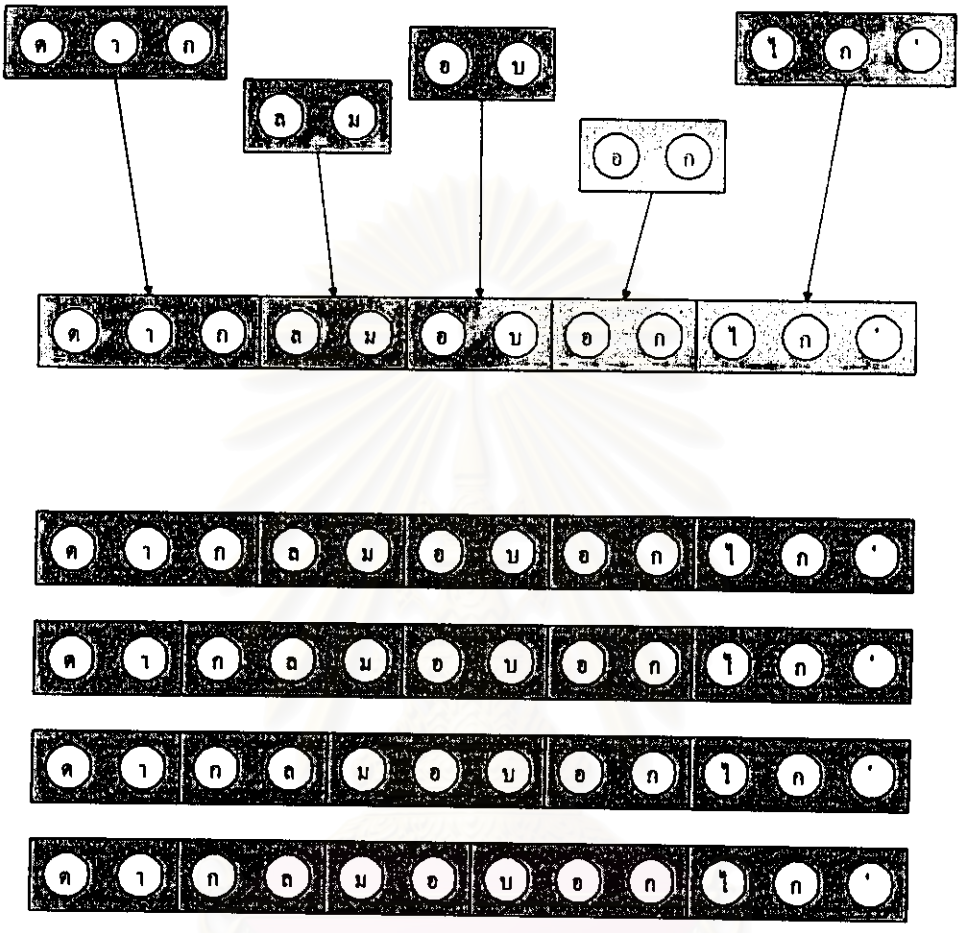
กรณีที่ 1

ประโยคที่เกิดจากการนำเอาคำที่อยู่ในพจนานุกรมทั้งหมดมารวมกัน ลักษณะกรณีที่ 1 ปรากฏดังรูปที่ 2.1



รูปที่ 2.1 ลักษณะของกรณีที่ 1 ประโยคที่เกิดจากคำที่มีในพจนานุกรมทั้งหมด

จากรูปที่ 2.1 คำต่าง ๆ ที่นำมารวมกันเป็นประโยคขึ้นมา เป็นคำที่อยู่ในพจนานุกรมทั้งหมด 5 คำ แต่เมื่อมีการดึงคำออกมา กลับได้คำที่อยู่ในพจนานุกรมออกมาจำนวน 12 คำ นำคำที่ได้มาจัดเรียงกันให้เป็นประโยคเดิม สามารถจัดได้ตามรูปที่ 2.2 ดังนี้



รูปที่ 2.2 การนำคำที่ดึงได้ มาจัดเรียงให้ได้ประโยคเดิม

รูปที่ 2.2 นำผลที่ได้จากรูปที่ 2.1 ซึ่งมีจำนวนคำทั้งหมด 12 คำมาจัดเรียงกันเพื่อให้ได้ประโยคเดิม ได้ทั้งหมด 4 รูปแบบ คำที่ได้จากทั้ง 4 รูปแบบนั้น จะต้องนำไปทำดัชนี เพื่อใช้ในการค้นหา เนื่องด้วยการดึงคำจะไม่มีการตรวจสอบไวยกรณ์ของภาษาไทย ดังนั้นจึงไม่สามารถระบุได้ว่าการแบ่งคำในลักษณะใดถูกต้อง จึงต้องเก็บคำทุกคำ

กรณีที่ 2

ประโยคที่เกิดจากการรวมกันของคำที่อยู่ในพจนานุกรม กับคำที่ไม่มีในพจนานุกรม แบ่งได้ทั้งหมด 5 แบบตามผลที่ได้จากการดึงคำ ดังนี้

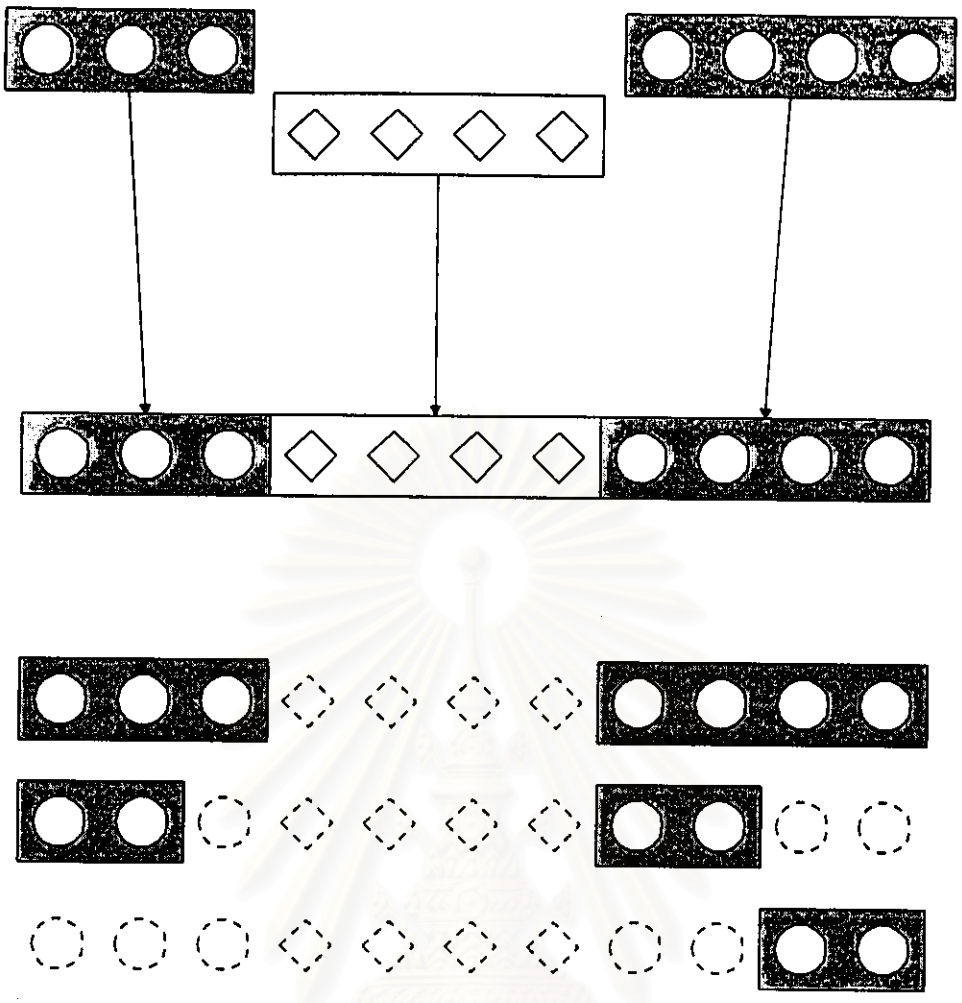
แบบที่ 1

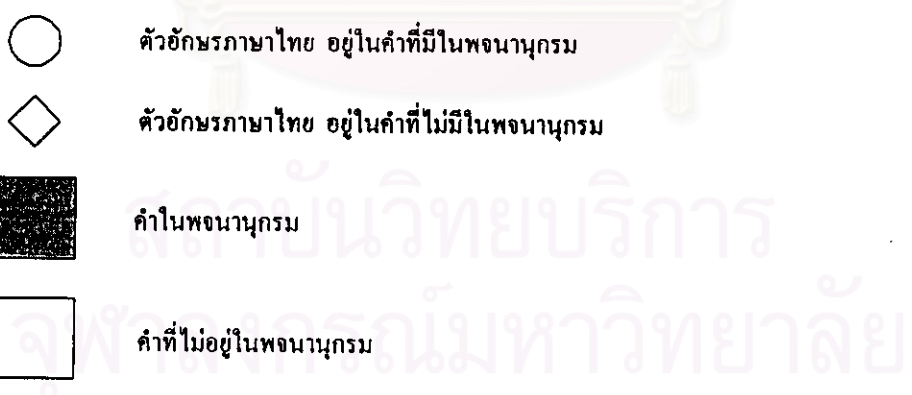
ไม่มีส่วนของคำในพจนานุกรมที่ดึงได้ ล้ำเข้าไปในส่วนของคำที่ไม่มีในพจนานุกรม ดังรูปที่ 2.3 ยกตัวอย่างเช่น "เขาได้ตำแหน่งท็อปชั้น" ซึ่งเกิดจากการรวมของคำที่มีอยู่ในพจนานุกรมคือคำว่า "เขา", "ได้", "ตำแหน่ง" และ "ชั้น" กับคำที่ไม่มีในพจนานุกรมคือคำว่า " ท็อป" สามารถดึงคำที่มีอยู่ในพจนานุกรมได้ดังนี้

เขา, เข, ขา, ได้, ได, ตำแหน่ง, ตำ, แหน, แห, หน, ชั้น

สังเกตได้ว่าไม่มีคำในพจนานุกรมคำใดที่ได้จากการดึงคำ แล้วมีส่วนหนึ่งของคำล้ำเข้าไปในคำว่า "ท็อป" ดังนั้นเมื่อนำคำที่ดึงได้มาจัดเรียงขึ้นมาใหม่ จะเกิดช่องว่างขึ้นตรงคำว่า "ท็อป" สามารถเก็บช่องว่างนั้นเพิ่มเติมเพื่อให้ประโยคสมบูรณ์

สรุปได้ว่าในประโยคที่ประกอบด้วยคำที่มีในพจนานุกรมและคำที่ไม่มีในพจนานุกรม ถ้าการดึงคำที่มีอยู่ในพจนานุกรมออกมาแล้ว คำเหล่านั้นไม่มีคำที่ล้ำเข้าไปในส่วนของคำที่ไม่มีในพจนานุกรม จะไม่เกิดปัญหา เนื่องจากจะเกิดช่องว่างของคำที่ไม่มีในพจนานุกรมนั้นขึ้น ทำให้สามารถเก็บคำนั้น ๆ เพิ่มได้อย่างถูกต้อง

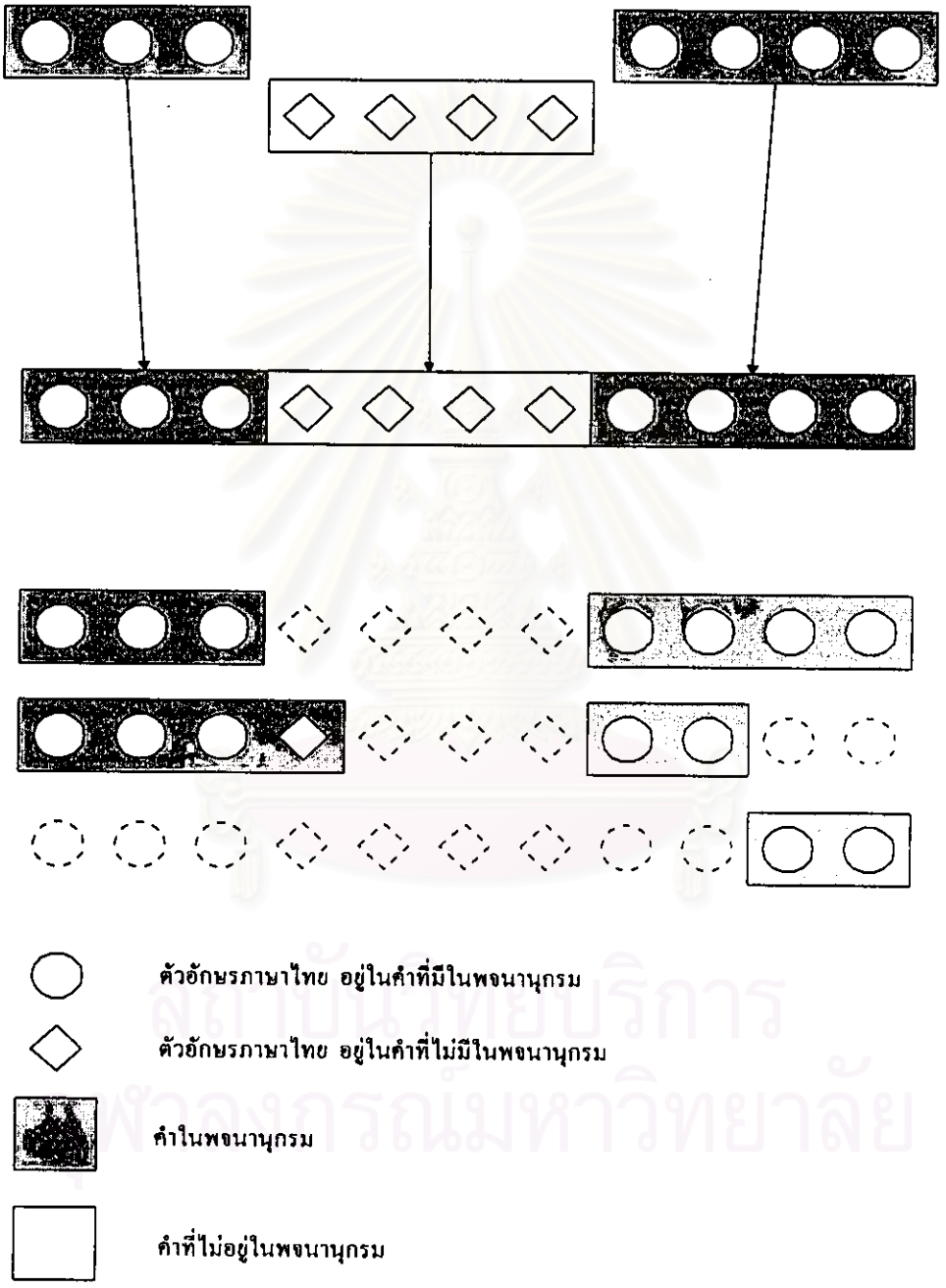




รูปที่ 2.3 ลักษณะของกรณีที่ 2 แบบที่ 1

แบบที่ 2

มีส่วนของคำในพจนานุกรมที่ดึงได้ ล้ำเข้าไปในส่วนของคำที่ไม่มีในพจนานุกรม ดังรูป ที่ 2.4



รูปที่ 2.4 ลักษณะของกรณีที่ 2 แบบที่ 2

ยกตัวอย่างประโยคที่อยู่ในแบบที่ 2 เช่น "เขาได้อันดับท็อปชั้น" ซึ่งเกิดจากการรวมกันของคำที่มีในพจนานุกรมคือคำว่า "เขา", "ได้", "อันดับ" และ "ชั้น" รวมกับคำที่ไม่มีในพจนานุกรมคือคำว่า "ท็อป" การดึงคำที่มีในพจนานุกรมจากประโยคดังกล่าว จะได้คำดังต่อไปนี้

เขา, เข, ขา, ได้, ได, อันดับ, อัน, ดับ, บท, ชั้น

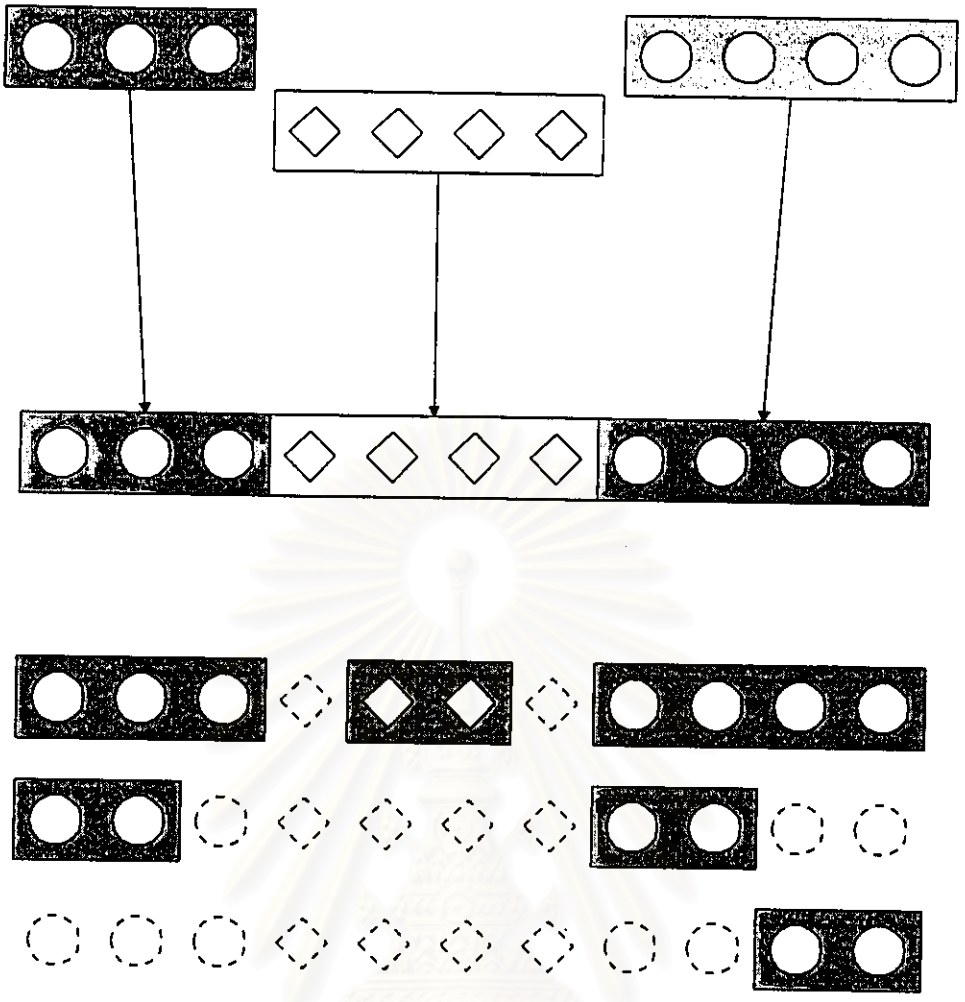
สังเกตได้ว่าคำว่า "บท" ตัวอักษร "ท" เป็นอักษรของคำว่า "ท็อป" ซึ่งเป็นคำที่ไม่มีในพจนานุกรม อีกตัวอย่างหนึ่งคำประโยคที่ว่า "ราชาเพลงป๊อประดับโลก" เกิดจากคำที่มีในพจนานุกรมคือคำว่า "ราชา", "เพลง", "ระดับ" และ "โลก" กับคำที่ไม่มีในพจนานุกรมคือคำว่า "ป๊อป" การดึงคำที่มีในพจนานุกรมจากประโยคดังกล่าว จะได้คำดังต่อไปนี้

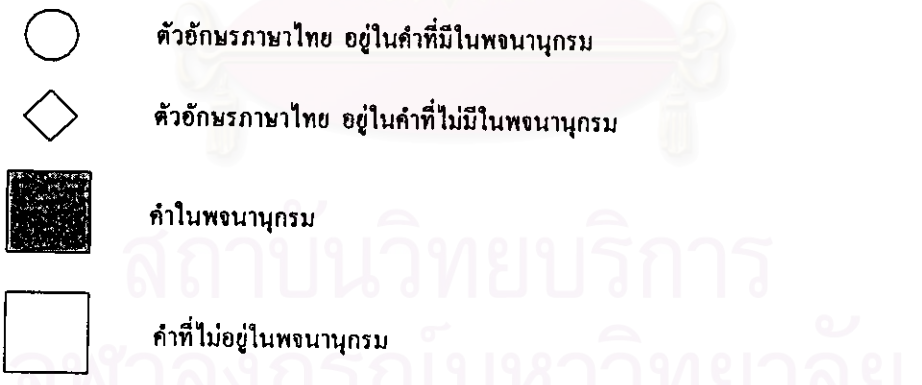
ราชา, ราช, รา, ชา, เพลง, เพล, เพ, พล, ลง,
ประดับ, ประ, ระดับ, ระ, ดับ, โลก, โล, ลก

คำว่า "ประดับ" และคำว่า "ประ" ตัวอักษร "ป" เป็นตัวอักษรของคำว่า "ป๊อป" ซึ่งเป็นคำที่ไม่มีในพจนานุกรม

แบบที่ 3

มีคำในพจนานุกรมที่ดึงได้อยู่ในคำที่ไม่มีในพจนานุกรม ดังรูปที่ 2.5 ลักษณะเช่นนี้จะทำให้เกิดช่องว่างขึ้นทั้งสองข้างของคำที่อยู่ในคำที่ไม่มีในพจนานุกรม

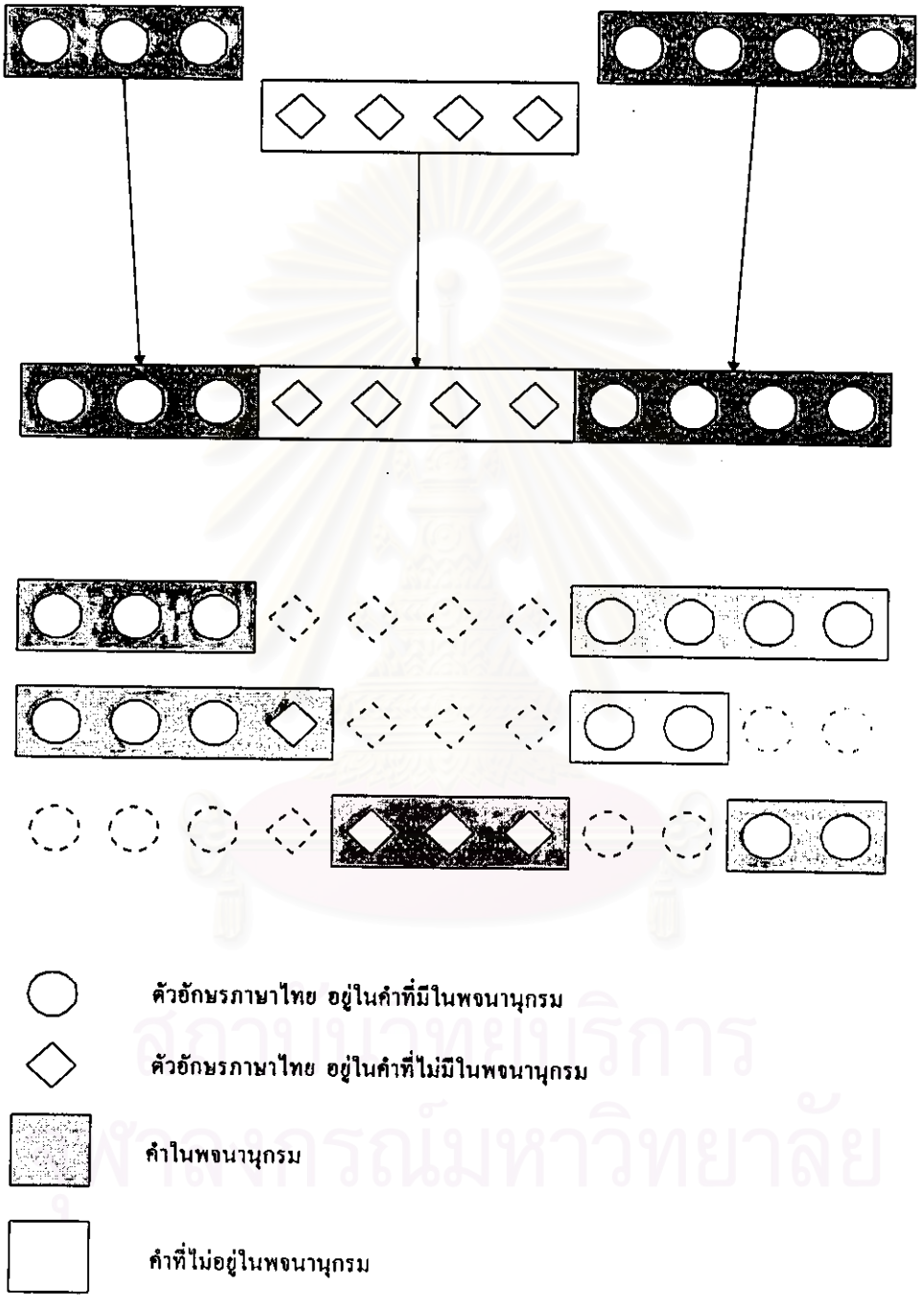




รูปที่ 2.5 ลักษณะของกรณีที่ 2 แบบที่ 3

แบบที่ 4

รูปแบบที่ 4 มีลักษณะดังรูปที่ 2.6



รูปที่ 2.6 ลักษณะของกรณีที่ 2 แบบที่ 4

จะมีคำในพจนานุกรมที่ดึงได้ล้ำเข้าไปในคำที่ไม่มีในพจนานุกรม และส่วนที่เหลือของคำที่ไม่มีในพจนานุกรมเป็นคำที่อยู่ในพจนานุกรมด้วย ยกตัวอย่างประโยคที่มีลักษณะเช่นนี้คือ " นายชาตินอยู่ไหน" เกิดจากคำที่มีในพจนานุกรมคำว่า "นาย", "อยู่" และ "ไหน" กับคำที่ไม่มีในพจนานุกรมคือคำว่า "ชาติน" ซึ่งเป็นชื่อคน การดึงคำที่มีในพจนานุกรมจากประโยคดังกล่าวจะได้ดังต่อไปนี้

นาย, นา, ชาติ, ชา, ติ, นอ, อยู่, ยู่, ไหน, ไห

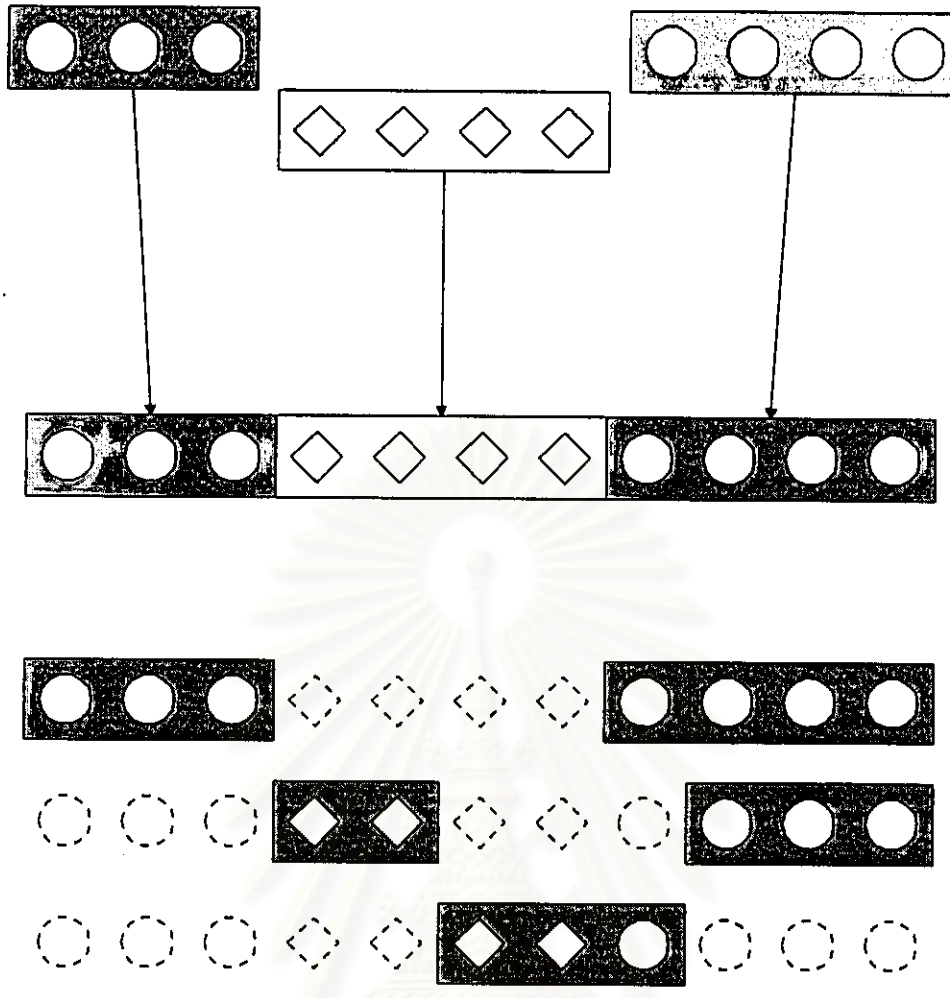
คำว่า "นอ" จะมีตัวอักษร "น" ที่เป็นของคำว่า "ชาติน" ส่วนที่เหลือคือคำว่า "ชาติ" เป็นคำในพจนานุกรม ลักษณะเช่นนำมาจัดเรียงเพื่อให้ได้ประโยคเดิมจะได้เป็นดังนี้

นาย + ชาติ + นอ + ยู่ + ไหน
นาย + ชา + ติ + นอ + ยู่ + ไหน

ไม่สามารถได้คำว่า "ชาติน" หรือ "ชา" กับ "ติน" เพื่อนำมาทำดัชนี ดังนั้นการค้นหาคำว่า "ชาติน" ในระบบสืบค้นข้อมูลก็ไม่สามารถทำได้

แบบที่ 5

มีลักษณะใกล้เคียงกับแบบที่ 4 ต่างกันตรงที่มีคำที่อยู่ในพจนานุกรมเป็นส่วนหนึ่งของคำที่ไม่มีในพจนานุกรมและมีส่วนที่ล้ำไปยังคำที่อยู่ในพจนานุกรม คำที่ดึงได้สามารถรวมเป็นประโยคที่สมบูรณ์ด้วย ดังรูปที่ 2.7



รูปที่ 2.7 ลักษณะของกรณีที่ 2 แบบที่ 5

ลักษณะอัลกอริทึมการจัดทำดัชนีที่นำมาใช้กับระบบสืบค้นข้อมูล ที่ใช้แฟ้มข้อมูลผกผัน

เนื่องด้วยอัลกอริทึมการแบ่งคำที่มีอยู่ปัจจุบัน จะนำมาใช้ในโปรแกรมประมวลผลคำเป็นหลัก โดยการหาจุดแบ่งคำในช่วงท้ายบรรทัด เพื่อทำการขึ้นบรรทัดใหม่ แต่การแบ่งคำที่จะนำมาใช้ในระบบสืบค้นข้อมูลนี้ จะต้องแบ่งออกมาทุก ๆ คำเพื่อนำมาเก็บไว้ในระบบ จึงต้องมีแนวความคิดค้นสำหรับอัลกอริทึมที่เหมาะสมขึ้นมาสำหรับระบบสืบค้นข้อมูลนี้

อัลกอริทึมในการแบ่งคำจะมีการนำพจนานุกรมมาใช้ เนื่องจากสิ่งที่เก็บอยู่ในพจนานุกรมเป็นคำหลัก ซึ่งจะเป็นสิ่งที่ใช้ในการค้นหาข้อมูล ดังนั้นการแบ่งคำที่ใช้พจนานุกรมจะแบ่งคำออกมาเป็นคำหลักที่อยู่ในพจนานุกรมนั้น แต่เนื่องจากพจนานุกรมนั้นไม่สามารถเก็บคำหลักที่มีอยู่ในปัจจุบันและในอนาคตได้ทั้งหมด จึงจำเป็นต้องมีการคิดค้นอัลกอริทึมที่ใช้ในการแบ่งคำเหล่านั้นเพิ่มเติมเข้าไป นอกจากปัญหาดังกล่าวแล้วยังมีปัญหาที่จะต้องคำนึงถึงดังต่อไปนี้

1. ในภาษาไทยคำบางคำนำมาเขียนติดกันแล้วทำให้เกิดปัญหาในการแบ่งคำ เช่นคำว่า "ตากลม" สามารถแบ่งออกได้เป็นคำว่า "ตาก" กับ "ลม" หรือคำว่า "ตา" กับ "กลม" เป็นต้น การแบ่งคำที่นำมาใช้ในระบบสืบค้นควรจะต้องแบ่งออกทุกกรณีที่เป็นไปได้ แล้วนำคำเหล่านั้นเก็บเข้าสู่ระบบสืบค้น เนื่องจากถ้าการแบ่งคำเลือกแบ่งตามรูปแบบใดรูปแบบหนึ่ง เช่นแบ่งเป็นคำว่า "ตาก" กับ "ลม" แล้วนำไปเก็บในระบบสืบค้น เมื่อผู้ใช้ค้นหาคำว่า "กลม" จะไม่สามารถหาพบ

2. เนื่องจากความพยายามดึงคำออกมาให้ได้มากที่สุด ดังนั้นจะได้คำที่ไม่ควรดึงออกมาด้วย เช่น "แปลบัญชี" คำว่า "ลบ" ซึ่งอยู่ระหว่างคำว่า "แปล" กับ "บัญชี" เป็นคำที่ไม่ควรแบ่งออกมา

3. คำยาวดีกว่าคำสั้น ตัวอย่างเช่นคำว่า "หมายเหตุ" สามารถดึงคำที่มีในพจนานุกรมได้คำว่า "หมายเหตุ", "หมาย", "หมา", "มาย", "มา", "เหตุ", "เห" และ "ตุ" คำที่ได้ทั้งหมดเก็บคำว่า "หมายเหตุ" เพียงคำเดียว ส่วนคำอื่น ๆ ไม่ควรเก็บ ทำให้เปลืองที่ และเสียเวลาในการทำดัชนี