บทที่ 5

ปัญหาความกำกวมและคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

ในบทนี้จะกล่าวถึงสาเหตุที่ทำให้การตัดคำด้วยพจนานุกรมไม่สามารถตัดคำได้ถูกต้อง โดยจะแบ่งปัญหาออกเป็นส่วนๆ เพื่อที่จะหาแนวทางในการแก้ไขปัญหาของแต่ละส่วนต่อไป สาเหตุที่ทำให้การตัดคำโดยใช้พจนานุกรมผิดพลาดมีอยู่ 2 สาเหตุคือ 1. ความกำกวม 2. คำศัพท์ที่ไม่ปรากฏในพจนานุกรม

5.1 ความกำกวม

ความกำกวมจะเกิดขึ้นเมื่อมีข้อความที่สามารถจะแบ่งได้หลายแบบ โดยทุกๆ คำที่เกิดขึ้นในแต่ละแบบจะเป็นคำศัพท์ที่พบในพจนานุกรมทั้งหมด ในกรณีนี้จะไม่พิจารณาถึงคำที่ไม่ปรากฏในพจนานุกรม จากความกำกวมที่เกิดขึ้น ในงานวิทยานิพนธ์นี้จะแบ่งประเภทของข้อความที่กำกวมออกเป็น 2 แบบตามลักษณะของข้อความที่กำกวมคือ

5.1.1 ข้อความกำกวมที่ขึ้นกับบริบท (Context Dependent Words)

คือข้อความกำกวมที่จำเป็นจะต้องพิจารณาข้อความรอบข้าง เพื่อเลือกแบบการตัดคำที่ดีที่สุด หรือกล่าวอีกนัยหนึ่งก็คือข้อความกำกวมประเภทนี้สามารถที่จะตัดคำได้หลายแบบ และแต่ละแบบก็มีความหมาย ทำให้การที่จะเลือกแบบตัดคำที่ถูกต้องนั้นจำเป็นต้องพิจารณาบริบทรอบๆ ด้วย เช่น

ตากลม สามารถตัดได้เป็น ตาก ลม หรือ ตา กลม
โคลง สามารถตัดได้เป็น โคลง หรือ โค ลง
ที่อยู่ สามารถตัดได้เป็น ที่อยู่ หรือ ที่ อยู่
มากว่า สามารถตัดได้เป็น มาก ว่า หรือ มา กว่า
สาวกลับ สามารถตัดได้เป็น สาวก ลับ หรือ สาว กลับ

5.1.2 ข้อความกำกวมที่ไม่ขึ้นกับบริบท (Context Independent Words)

คือข้อความกำกวมที่ไม่มีความจำเป็นต้องพิจารณาข้อความรอบข้าง และสามารถจะเลือกได้ทันทีว่าควรจะตัดคำแบบไหน หรือกล่าวอีกนัยหนึ่งก็คือข้อความที่สามารถตัดคำได้หลายๆ แบบแต่จะมีเพียงแบบเดียวเท่านั้นที่มีความหมาย ตัวอย่างเช่น

ขนบนอก	สามารถตัดได้เป็น	ขน บน อก หรือ ขนบ นอก
โคนกลีบดอก	สามารถตัดได้เป็น	โคน กลีบ ดอก หรือ โค นก ลีบ ดอก
นำมากลั่น	สามารถตัดได้เป็น	นำ มา กลั่น หรือ นำ มาก ลั่น
ไปหามเหลี	สามารถตัดได้เป็น	ไป หา มเหลี หรือ ไป หาม เห ลี
คอกว้าง	สามารถตัดได้เป็น	คอ กว้าง หรือ คอก ว้าง

หมายเหตุ ข้อความที่ขีดเส้นใต้คือข้อความที่ถูกต้อง

จากลักษณะของข้อความกำกวมที่กล่าวมา จะเห็นว่าความกำกวมที่เกิดขึ้นนั้นมีอยู่ทั้งหมด 2 แบบ สำหรับการเพิ่มประสิทธิภาพของการตัดคำให้ดียิ่งขึ้นนั้นมีความจำเป็นที่จะต้องหาวิธีการต่างๆ มาแก้ปัญหาความกำกวมทั้ง 2 แบบ โดยการแก้ปัญหาความกำกวมที่ขึ้นกับบริบทนั้นจะทำได้ยากกว่า และจะต้องใช้วิธีการที่ซับซ้อนกว่าการแก้ปัญหาความกำกวมแบบที่ไม่ขึ้นกับบริบท นอกเหนือจากการแก้ไขปัญหาความกำกวมแล้ว สิ่งที่จะต้องพิจารณาต่อไปคือเรื่องของคำศัพท์ที่ไม่ปรากฏในพจนานุกรม ซึ่งจะกล่าวต่อไป

5.2 คำศัพท์ที่ไม่ปรากฏในพจนานุกรม

สำหรับปัญหาเรื่องคำศัพท์ที่ไม่ปรากฏในพจนานุกรม เป็นสาเหตุสำคัญที่ทำให้การตัดคำโดยใช้พจนานุกรมไม่สามารถจะตัดคำเหล่านั้นได้ถูกต้อง เนื่องจากคำในภาษาไทยสามารถที่จะเกิดขึ้นมาได้ใหม่โดยการประสมระหว่างคำหรือพยางค์ได้ ทำให้การหาขอบเขตของคำที่ไม่ปรากฏในพจนานุกรมทำได้ยาก แต่อย่างไรก็ตามปัจจุบันได้มีผู้คิดและพัฒนาแก้ไขปัญหาดังกล่าว ซึ่งสามารถแก้ไขปัญหาเหล่านี้ได้ดีพอสมควร โดยมีการนำเรื่องสถิติ ไวยากรณ์และความหมายเข้ามาช่วยพิจารณาในการแก้ปัญหา แต่ก็ยังไม่สามารถที่จะแก้ไขปัญหาได้ทั้งหมด ทำให้ยังคงต้องมีการพัฒนาและค้นหาวิธีการที่จะแก้ไขอีกต่อไป

ประเภทต่างๆ ของคำที่ไม่ปรากฏในพจนานุกรม

คำศัพท์ที่ไม่ปรากฏอยู่ในพจนานุกรมสามารถแบ่งออกได้เป็น 6 ประเภทคือ

1. ชื่อเฉพาะ
2. คำจากภาษาต่างประเทศ
3. คำทับศัพท์
4. คำย่อ
5. คำราชาศัพท์
6. คำที่สะกดผิด

จากการรวบรวมค่าสถิติของคำศัพท์ที่ไม่มีในพจนานุกรมนั้น จะแสดงดังตารางที่ 1 โดยการรวบรวมของ (Asanee Kawtrakul et al., 1997)

ตารางที่ 5-1 ตารางค่าสถิติของคำที่ไม่มีในพจนานุกรมประเภทต่างๆ ในเอกสารต่างๆ

ชนิดของคำที่ไม่มีในพจนานุกรม	ชนิดของเอกสาร (%)		
	วิทยาศาสตร์	ข่าว	สารคดี
ชื่อเฉพาะ	3.06	73.4	51.15
คำทับศัพท์	43.6	6.88	14.39
คำย่อ	3.90	9.63	4.63
คำจากภาษาต่างประเทศ	47.49	1.15	21.34
คำราชาศัพท์	-	-	5.91

จากตารางที่ 5-1 จะเห็นว่าคำศัพท์ที่ไม่มีอยู่ในพจนานุกรมในเอกสารประเภทข่าว และสารคดี ส่วนใหญ่จะเป็นชื่อเฉพาะ ส่วนในเอกสารประเภทวิทยาศาสตร์ จะมีการใช้คำทับศัพท์และคำจากภาษาต่างประเทศเป็นจำนวนมาก เนื่องจากการแก้ปัญหาการตัดคำของคำศัพท์ที่ไม่มีอยู่ในพจนานุกรมนั้นจะขึ้นอยู่กับประเภทของคำด้วย ดังนั้นคำศัพท์ประเภทแรกที่วิทยานิพนธ์นี้จะทำการแก้ปัญหาคือคำศัพท์ประเภทชื่อเฉพาะ เพราะว่าคำประเภทนี้มีการใช้เป็นจำนวนมาก และการแก้ปัญหาจะสามารถนำไปประยุกต์ใช้ในงานด้านการสืบค้นสารสนเทศ (Information Retrieval) ได้ด้วย

5.2.1 ลักษณะของคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

เนื่องจากคำในภาษาไทยนั้นสามารถจะเกิดขึ้นใหม่โดยอาจจะเกิดจากการประสมระหว่างคำ หรือระหว่างพยางค์เป็นต้น ดังนั้นจึงเป็นสาเหตุให้การหาขอบเขตของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้นทำได้ยาก โดยจะต้องมีการนำคำหรือข้อความรอบๆ บริเวณคำศัพท์ที่ไม่ปรากฏในพจนานุกรมมาช่วยในการหาขอบเขต ดังนั้นก่อนที่จะทำการอธิบายขั้นตอนในการหาขอบเขตของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้น ในส่วนนี้จะอธิบายถึงลักษณะของคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

ลักษณะของคำศัพท์ที่ไม่ปรากฏในพจนานุกรม จะเห็นว่าคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอาจจะประกอบไปด้วยข้อความที่มีในพจนานุกรม (Known String) กับข้อความที่ไม่ปรากฏในพจนานุกรม (Unknown String) และเมื่อพิจารณาจากลักษณะของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมสามารถที่จะแบ่งคำที่ไม่ปรากฏในพจนานุกรมได้เป็น 2 ประเภทใหญ่ๆ คือ

➢ คำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างชัดเจน (Explicit Unknown Word) คำศัพท์ประเภทนี้คือคำศัพท์ที่ไม่ปรากฏในพจนานุกรม โดยภายในคำนั้นๆ จะไม่มีข้อความส่วนใดๆ ภายในคำนั้นที่เป็นคำที่พบอยู่ในพจนานุกรม ตัวอย่างเช่นคำว่า "โลตัส", "แฮรี่", "ดุนีย์" ฯลฯ

➢ คำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้น (Hidden Unknown Word) คำศัพท์ประเภทนี้คือคำศัพท์ที่ไม่ปรากฏในพจนานุกรม โดยภายในคำนั้นๆ จะมีข้อความส่วนหนึ่งส่วนใดภายในคำนั้นที่เป็นคำที่พบอยู่ในพจนานุกรม ตัวอย่างเช่นคำว่า "สุมานี", "คธาพงษ์", " สม ชาย", "สม ศักดิ์" เป็นต้น และคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นนี้สามารถจะแบ่งเป็นประเภทย่อยๆ ได้อีก 2 ประเภทคือ

1. คำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นบางส่วน (Partially Hidden Unknown Word) คือคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นที่เกิดจากการประกอบ ระหว่างคำที่ปรากฏในพจนานุกรมกับข้อความที่ไม่ปรากฏในพจนานุกรม ตัวอย่างเช่น "สุมานี" และ "คธาพงษ์"

2. คำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นทุกส่วน (Fully Hidden Unknown Word) คือคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นที่เกิดจากการประกอบไปด้วยคำที่ปรากฏในพจนานุกรมทั้งหมด หรืออาจกล่าวได้ว่าเป็นคำที่สร้างขึ้นใหม่โดยมีการนำคำศัพท์ต่างๆ มาประกอบกัน ตัวอย่างเช่น "สมชาย" เกิดจากการประสมระหว่าง "สม" กับ "ชาย" และคำว่า "สมหญิง" เกิดจากการประสมคำระหว่าง "สม" กับ "หญิง" เป็นต้น

จากลักษณะของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมที่ได้อธิบายมาในข้างต้นแล้ว จะเห็นว่าคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้นสามารถจะเกิดได้หลายๆ รูปแบบ ดังนั้นการแก้ไขปัญหานี้จะทำได้ยากเพราะจะไม่สามารถรู้ขอบเขตที่แน่นอนของคำได้ ทำให้การแก้ปัญหาจะต้องมีการนำคำบริบทเข้ามาช่วย ซึ่งในวิทยานิพนธ์นี้จะอธิบายวิธีแก้ปัญหาเรื่องคำศัพท์ที่ไม่ปรากฏในพจนานุกรมในบทที่ 7

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย