บทที่ 7

การตัดคำภาษาไทยโดยใช้คุณลักษณะ

จากบทที่แล้วได้กล่าวถึงขบวนการเรียนรู้ของเครื่องแบบต่างๆ ไปแล้ว ในบทนี้จะกล่าวถึงวิธีการที่จะนำเอาการเรียนรู้ของเครื่องเข้ามาประยุกต์ใช้ในการเลือกคุณลักษณะต่างๆ จากคลังข้อความที่สามารถจะนำมาใช้ในการแก้ไขปัญหาคำกำกวมและปัญหาชื่อเฉพาะที่ไม่ปรากฏในพจนานุกรมได้ นิยามของคุณลักษณะในที่นี้หมายถึงข้อมูลใดๆ ที่สามารถจะนำมาใช้ในการแก้ไขปัญหาการตัดคำได้

7.1 คุณลักษณะ

คุณลักษณะที่จะนำมาใช้ในการแก้ไขปัญหาการตัดคำทั้งปัญหาความกำกวมและปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้น มีอยู่ 2 ชนิดคือ คำบริบท (Context Word) และ สิ่งที่เกิดร่วมกันโดยมีลำดับ (Collocation)

7.1.1 คำบริบท (Context Word)

คำบริบทคือ คำที่อยู่รอบๆ ข้อความหรือคำที่จะนำมาพิจารณา (Target String/ Target Word) สำหรับในงานวิทยานิพนธ์นี้จะนำบริบทที่อยู่ห่างจากข้อความหรือคำที่จะนำมาพิจารณาภายใน 10 คำก่อนหน้าหรือหลังข้อความที่จะนำมาพิจารณา ส่วนข้อความที่จะนำมาพิจารณาในที่นี้จะขึ้นอยู่กับลักษณะของปัญหาที่จะนำมาประยุกต์ใช้ ซึ่งในที่นี้จะนำไปประยุกต์ใช้กับ 2 ปัญหาคือ ปัญหาความกำกวมและ ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

ในการแก้ปัญหาความกำกวม ข้อความหรือคำที่จะนำมาพิจารณาคือข้อความที่กำกวมและคำบริบทสำหรับปัญหานี้ก็คือคำรอบๆ ข้อความที่กำกวมภายใน +/- 10 คำ สำหรับปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม ข้อความหรือคำที่จะนำมาพิจารณาคือ ตัวเลือกของคำที่ไม่ปรากฏในพจนานุกรม (Unknown Word Candidate) ส่วนบริบทสำหรับปัญหานี้ก็คือ คำรอบๆ ตัวเลือกของคำที่ไม่ปรากฏในพจนานุกรมภายใน +/- 10 คำ

ตัวอย่างของคำบริบทที่นำมาใช้ในการแก้ปัญหาความกำกวม เช่นถ้ามีข้อความกำกวม "ตากลม" และมีคำบริบทอยู่ด้านขวามือของข้อความนี้เป็นคำว่า "แป้ว" ก็จะตัดสินใจให้ตัดคำเป็น "ตา กลม"

7.1.2 สิ่งที่เกิดร่วมกันโดยมีลำดับ (Collocation)

สิ่งที่เกิดร่วมกันโดยมีลำดับคือ คำหรือหน้าที่คำที่ติดกับข้อความหรือคำที่จะนำมาพิจารณา สำหรับในงานวิทยานิพนธ์นี้จะนำคำหรือหน้าที่คำก่อนหน้าหรือหลังข้อความที่พิจารณาเพียง 2 คำ ส่วนข้อความหรือคำที่นำมาพิจารณาในที่นี้ จะขึ้นอยู่กับลักษณะของปัญหา สำหรับในที่นี้ปัญหาที่จะนำมาแก้ไข คือ ปัญหาความกำกวม และ ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม โดยวิธีการเลือกข้อความหรือคำที่จะนำมาพิจารณานั้น จะมีลักษณะเหมือนกับที่พิจารณาในคำบริบทดังที่ได้กล่าวไปแล้ว

ตัวอย่างของการนำสิ่งที่เกิดร่วมกันโดยมีลำดับมาใช้ในการแก้ปัญหาความกำกวม ตัวอย่างการแก้ความกำกวมของข้อความ "มากว่า" ซึ่งสามารถตัดคำได้เป็น "มา กว่า" หรือ "มาก ว่า" เช่น

ถ้า มากว่า *ตัวเลข* *CMTR* แล้วให้ตัดเป็น "มา กว่า"

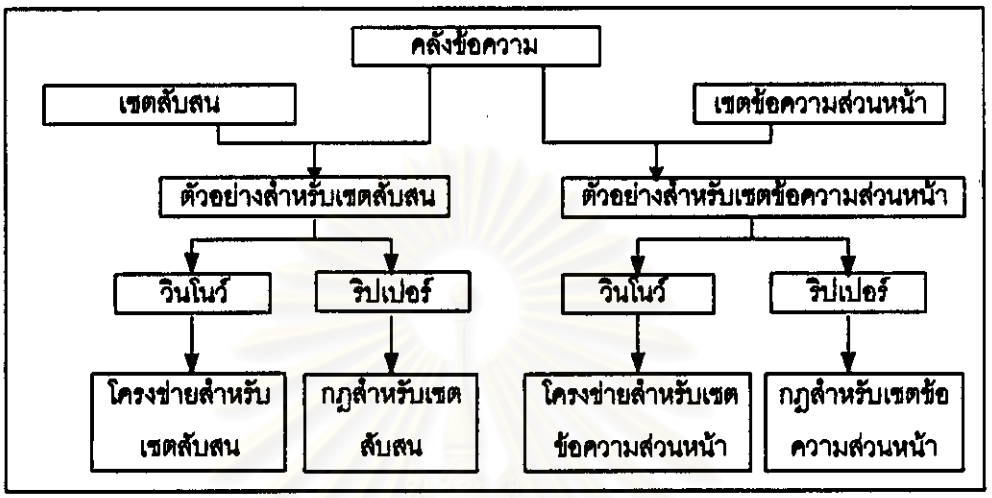
จากตัวอย่างข้างบนสิ่งที่เกิดร่วมกันโดยมีลำดับคือ ตัวเลข และ CMTR จากตัวอย่างข้างต้นหมายความว่า ถ้าพบข้อความ "มากว่า" แล้วตามด้วยตัวเลข และคำถัดไปมีหน้าที่คำเป็น CMTR แล้วให้ตัดคำเป็น "มา กว่า" โดยที่ CMTR หมายถึงหน่วยในการวัด เช่น ปี กิโลกรัม ชั่วโมง เป็นต้น

7.2 การแก้ไขปัญหาความกำกวม

เนื่องจากปัญหาความกำกวมนั้นเป็นปัญหาที่สำคัญในการตัดคำ และการแก้ปัญหาความกำกวมนั้นมีแนวทางในการแก้ปัญหาหลายๆ รูปแบบดังที่ได้กล่าวไปแล้วในบทที่ 2 สำหรับในงานวิทยานิพนธ์นี้ได้เสนอวิธีการใหม่ในการแก้ปัญหาความกำกวม โดยจะนำเอาคุณลักษณะแบบต่างๆ มาประยุกต์ใช้ โดยคุณลักษณะที่จะนำมาใช้คือคำบริบทและสิ่งที่เกิดร่วมกันโดยมีลำดับ ตามที่ได้กล่าวไปแล้วในตอนต้น

การแก้ปัญหาความกำกวมโดยใช้คุณลักษณะ จะแบ่งวิธีการแก้ไขปัญหานี้ออกเป็น 2 แบบโดยที่วิธีการทั้ง 2 แบบนี้สามารถจะนำไปใช้ในการแก้ปัญหาได้ ซึ่งในแต่ละแบบนั้นจะมีข้อดีและข้อเสียแตกต่างกัน โดยจะได้ทำการอธิบายในลำดับถัดไป สำหรับการแก้ปัญหาความกำกวมนั้นสามารถแบ่งวิธีการแก้ปัญหาได้ดังต่อไปนี้คือ เซตสับสน (Confusion Set) และเซตข้อความส่วนหน้า (Prefix Set)

สำหรับขั้นตอนการสร้างการเรียนรู้ให้กับวินโนว์และริปเปอร์ในการแก้ปัญหาความกำกวม โดยแบบที่ใช้เซตสับสน หรือเซตข้อความส่วนหน้า ดังแสดงในรูปที่ 7-1 ส่วนรายละเอียดการสร้างเซตสับสน และเซตข้อความส่วนหน้านั้นจะอธิบายในหัวข้อ 7.2.1 และ 7.2.2 ตามลำดับ



รูปที่ 7-1 ขั้นตอนการเรียนรู้คุณลักษณะเพื่อนำมาใช้ในการแก้ปัญหาความกำกวม

7.2.1. เซตสับสน (Confusion Set)

ลักษณะของปัญหาความกำกวมในการตัดคำคือ การที่สามารถตัดคำได้หลายๆ แบบ สำหรับข้อความหนึ่งๆ ซึ่งจะเป็นสาเหตุทำให้เกิดความสับสนขึ้นว่าแบบไหนจะเป็นแบบที่ถูกต้องที่สุด ดังนั้น ในวิทยานิพนธ์นี้จะแก้ปัญหาความกำกวมโดยมีการนำเซตสับสนเข้ามาประยุกต์ใช้ สำหรับเซตสับสนนั้นจะต้องสร้างขึ้นสำหรับทุกๆ ข้อความที่กำกวม ตัวอย่างเช่น ถ้ามีข้อความ "ตากลม" "มากว่า" และ "ขนมอบ" ซึ่งเป็นข้อความที่กำกวม วิธีการนี้จะต้องสร้างเซตสับสนของข้อความเหล่านี้ทั้งหมด

นิยามของเซตสับสนคือ เซตของข้อความที่กำกวม โดยสมาชิกภายในเซตนั้นประกอบไปด้วยข้อความที่ตัดคำที่เป็นไปได้ทุกๆ แบบสำหรับข้อความกำกวมนั้นๆ ตัวอย่างเช่น ข้อความ "มากว่า" เป็นข้อความกำกวม ดังนั้นวิธีการนี้จะต้องสร้างเซตสับสนสำหรับข้อความนี้ โดยที่ข้อความนี้เมื่อทำการตัดคำแล้วสามารถจะตัดได้เป็น "มาก ว่า" กับ "มา กว่า"ดังนั้นเซตสับสนของข้อความ "มากว่า" จะได้ดังแสดงตัวอย่างที่ 7-1

$$C_{\text{มากว่า}} = \{\text{มาก ว่า, มา กว่า}\} \qquad (7\text{-}1)$$

จากตัวอย่าง 7-1 $C_{มากว่า}$ คือเซตสับสนของข้อความ "มากว่า" ซึ่งสมาชิกภายในเซตนี้จะประกอบไปด้วย "มาก ว่า" และ "มา กว่า" ซึ่งจะหมายความว่าข้อความ "มากว่า" สามารถจะตัดคำได้ทั้งหมด 2 แบบคือ "มาก ว่า" หรือ "มา กว่า"

เมื่อทำการสร้างเซตสับสนสำหรับข้อความกำกวมที่ปรากฏอยู่ในคลังข้อความทั้งหมดเป็นที่เรียบร้อยแล้ว ขั้นตอนต่อไปคือส่งตัวอย่างต่างๆ เข้าไปให้การเรียนรู้ของเครื่อง เพื่อที่จะให้การเรียนรู้ของเครื่องนั้นทำการเลือกคุณลักษณะต่างๆ ที่สำคัญออกมา โดยคุณลักษณะต่างๆ ที่เลือกออกมาได้นั้นสามารถนำมาใช้ในการจำแนกระหว่างสมาชิกภายในเซตนั้นๆ ได้ หรืออีกนัยหนึ่งก็คือสามารถที่จะนำคุณลักษณะต่างๆ เข้ามาใช้ในการระบุว่าข้อความที่กำกวมนี้สามารถจะตัดคำได้เป็นอย่างไร

ตัวอย่างการนำการเรียนรู้ของเครื่องเข้ามาใช้ในการแก้ไขปัญหาความกำกวม จากตัวอย่างการสร้างเซตสับสนของข้อความ "มากว่า" ซึ่งแสดงในตัวอย่างที่ 7-1 ข้อความนี้สามารถแปลงคำได้เป็น 2 แบบคือ "มาก ว่า" กับ "มา กว่า" ดังนั้นจะต้องนำประโยคที่มีคำว่า "มาก ว่า" หรือ "มา กว่า" จากคลังข้อความที่ทำการตัดคำและกำกับหน้าที่คำเรียบร้อยแล้ว มาเป็นตัวอย่างของการเรียนรู้ของเครื่องวิปเปอร์ หรือวินโนว์ โดยให้ข้อความ "มากว่า" เป็นข้อความที่พิจารณา ส่วนคำรอบๆ ข้อความนี้ให้พิจารณาเป็น คำบริบท หรือ สิ่งที่เกิดร่วมกันโดยมีลำดับ สำหรับแต่ละตัวอย่างที่ส่งเข้าไปต้องระบุด้วยว่าคุณลักษณะที่นำเข้าไปให้วิปเปอร์กับวินโนว์นั้นเป็นคุณลักษณะของ "มาก ว่า" หรือ "มา กว่า"

เมื่อวิปเปอร์และวินโนว์ได้ทำการเรียนรู้คุณลักษณะต่างๆ ที่ใช้ในการแก้ปัญหาความกำกวมในรูปแบบเซตสับสนเรียบร้อยแล้ว ดังนั้นเมื่อต้องการตัดคำสำหรับข้อความที่กำกวมที่ได้มีการสร้างเซตสับสน และทำการเรียนรู้ไปแล้ว คำรอบๆ ข้อความที่กำกวมนั้นจะถูกพิจารณาเป็นคำบริบท และสิ่งที่เกิดร่วมกันโดยมีลำดับ และเมื่อทำการส่งให้กับวินโนว์หรือวิปเปอร์ วินโนว์หรือวิปเปอร์จะพิจารณาจากคำบริบทและสิ่งที่เกิดร่วมกันโดยมีลำดับ โดยเปรียบเทียบกับโครงข่ายในกรณีที่ใช้วินโนว์ หรือกฎในกรณีที่ใช้วิปเปอร์ที่ได้มาจากการเรียนรู้ แล้วจะทำการตัดสินใจเลือกแบบการตัดทำที่ถูกต้อง

7.2.2. เซตข้อความส่วนหน้า (Prefix Set)

จากวิธีแรกเป็นการสร้างเซตสับสน ซึ่งทำการสร้างเซตของการตัดคำที่เป็นไปได้ทุกๆ แบบของข้อความที่กำกวม และเมื่อได้เซตสับสนของข้อความที่กำกวมแล้วจะทำการเรียนรู้เพื่อหาคุณลักษณะต่างๆ ที่จะสามารถนำมาจำแนกสมาชิกต่างๆ ภายในเซตได้ ซึ่งวิธีการนี้จะมีข้อจำกัดคือจะต้องมีการสร้างเซตสับสนสำหรับข้อความที่กำกวมที่เป็นไปได้ทั้งหมดก่อน ทำให้วิธีการนี้ไม่สามารถแก้ปัญหาข้อความกำกวมที่เกิดขึ้นมาใหม่ได้ แต่วิธีที่จะนำเสนอต่อไปนี้จะเป็นอีกวิธีหนึ่งที่สามารถยอมให้เกิดความกำกวมขึ้นมาใหม่ได้ ซึ่งวิธีการนี้เป็นการสร้างเซตอีกแบบหนึ่งที่จะนำมาใช้แก้ปัญหาความกำกวมเหมือนกัน โดยที่วิธีนี้จะสร้างเซตจากคำศัพท์ที่ปรากฏในพจนานุกรมทั้งหมด และเรียกเซตชนิดนี้ว่าเซตข้อความส่วนหน้า

นิยามเซตข้อความส่วนหน้า คือเซตที่ประกอบด้วยคำต่างๆ เป็นสมาชิก โดยคำที่มีจำนวนตัวอักษรน้อยกว่า จะเป็นข้อความส่วนหน้า (Prefix) ของคำศัพท์ที่มีจำนวนตัวอักษรมากกว่าที่ปรากฏภายในเซตนั้นๆ เสมอ

การสร้างเซตข้อความส่วนหน้า วิธีการนี้จะสร้างเซตข้อความส่วนหน้าของทุกๆ คำที่มีอยู่ในพจนานุกรม ยกเว้นในกรณีที่มีการสร้างเซตข้อความส่วนหน้าแล้วมีสมาชิกแค่เพียงสมาชิกเดียว ตัวอย่างการสร้างเซตข้อความส่วนหน้า สมมุติว่าคำศัพท์ที่มีอยู่ในพจนานุกรมมีดังต่อไปนี้คือ มา, มาก, มากมาย, ตา, ตาก และ ตาม ดังนั้นวิธีการนี้จะสร้างเซตข้อความส่วนหน้าของทุกๆ คำในพจนานุกรม ดังนั้นจะได้เซตข้อความส่วนหน้าของคำว่า มาก, มากมาย, ตาก และ ตาม โดยจะแสดงตามตัวอย่างที่ 7-2 , 7-3, 7-4 และ 7-5 ตามลำดับ ส่วนเซตข้อความส่วนหน้าของคำว่า มา กับ มาก นั้นไม่ต้องสร้างขึ้นมาเพราะเนื่องจากมีสมาชิกภายในเซตเพียงสมาชิกเดียว

$$P_{\text{มาก}} = \{\text{มา, มาก}\} \tag{7-2}$$

$$P_{\text{มากมาย}} = \{\text{มา, มาก, มากมาย}\} \tag{7-3}$$

$$P_{\text{ตาก}} = \{\text{ตา, ตาก}\} \tag{7-4}$$

$$P_{\text{ตาม}} = \{\text{ตา, ตาม }\} \tag{7-5}$$

จากตัวอย่างที่ 7-2, 7-3, 7-4 และ 7-5 นั้นสัญลักษณ์ P_a นั้นคือเซตข้อความส่วนหน้าของคำ a และในแต่ละตัวอย่างจะแสดงถึงสมาชิกภายในของแต่ละเซตข้อความส่วนหน้า

เมื่อมีการสร้างเซตข้อความส่วนหน้าเรียบร้อยแล้ว ขั้นตอนต่อไปคือการนำตัวอย่างไปให้วิปเปอร์กับวินโนว์เรียนรู้เพื่อที่เลือกคุณลักษณะที่สำคัญออกมา ซึ่งสามารถจะนำมาใช้ในการจำแนกระหว่างสมาชิกภายในเซตนั้น สำหรับวิธีการสร้างตัวอย่างของแต่ละเซตนั้น จะมีขั้นตอนการสร้างดังต่อไปนี้คือ 1. เลือกเซตข้อความส่วนหน้า 2. เลือกตัวอย่างประโยคจากคลังข้อความที่มีการตัดคำและกำกับหน้าที่คำเรียบร้อยแล้ว โดยให้เลือกประโยคที่มีคำที่เป็นสมาชิกภายในเซตนั้นๆ ขึ้นมา 3. ให้ส่งคำบริบท และสิ่งที่เกิดร่วมกันโดยมีลำดับของคำที่เป็นสมาชิกภายในเซตนั้นๆ แล้วต้องระบุด้วยว่าเป็นคำบริบทและสิ่งที่เกิดร่วมกันโดยมีลำดับเป็นของคำใด เมื่อวิปเปอร์หรือวินโนว์ทำการเรียนรู้ของแต่ละเซตแล้ว จะได้คุณลักษณะในรูปแบบกฎสำหรับวิปเปอร์ หรือ โครงข่ายสำหรับวินโนว์ ที่สามารถจะใช้จำแนกระหว่างสมาชิกภายในเซตข้อความส่วนหน้าได้

เมื่อมีการสร้างเซตข้อความส่วนหน้าและมีการสร้างการเรียนรู้ให้กับริปเปอร์หรือวินโนว์สำหรับแต่ละเซตข้อความส่วนหน้าแล้ว ขั้นตอนต่อไปจะแสดงการนำเซตข้อความส่วนหน้าเข้ามาใช้ในการแก้ปัญหาความกำกวม ตัวอย่างเช่นในกรณีที่พบข้อความว่า "มากมาย" ซึ่งเป็นข้อความที่กำกวม สำหรับวิธีการแก้ปัญหานี้คือการนำเซตข้อความส่วนหน้าของคำว่า "มากมาย" มาใช้ แล้วทำการส่งคำบริบท และการเกิดร่วมกันโดยมีลำดับให้กับวินโนว์หรือริปเปอร์ วินโนว์หรือริปเปอร์นั้นจะทำการเปรียบเทียบระหว่างคุณลักษณะที่ส่งเข้ามากับคุณลักษณะต่างๆ ที่ได้เคยเรียนรู้ไว้ และจะทำการตัดสินใจออกมาว่าควรจะตัดคำเป็นอย่างไร

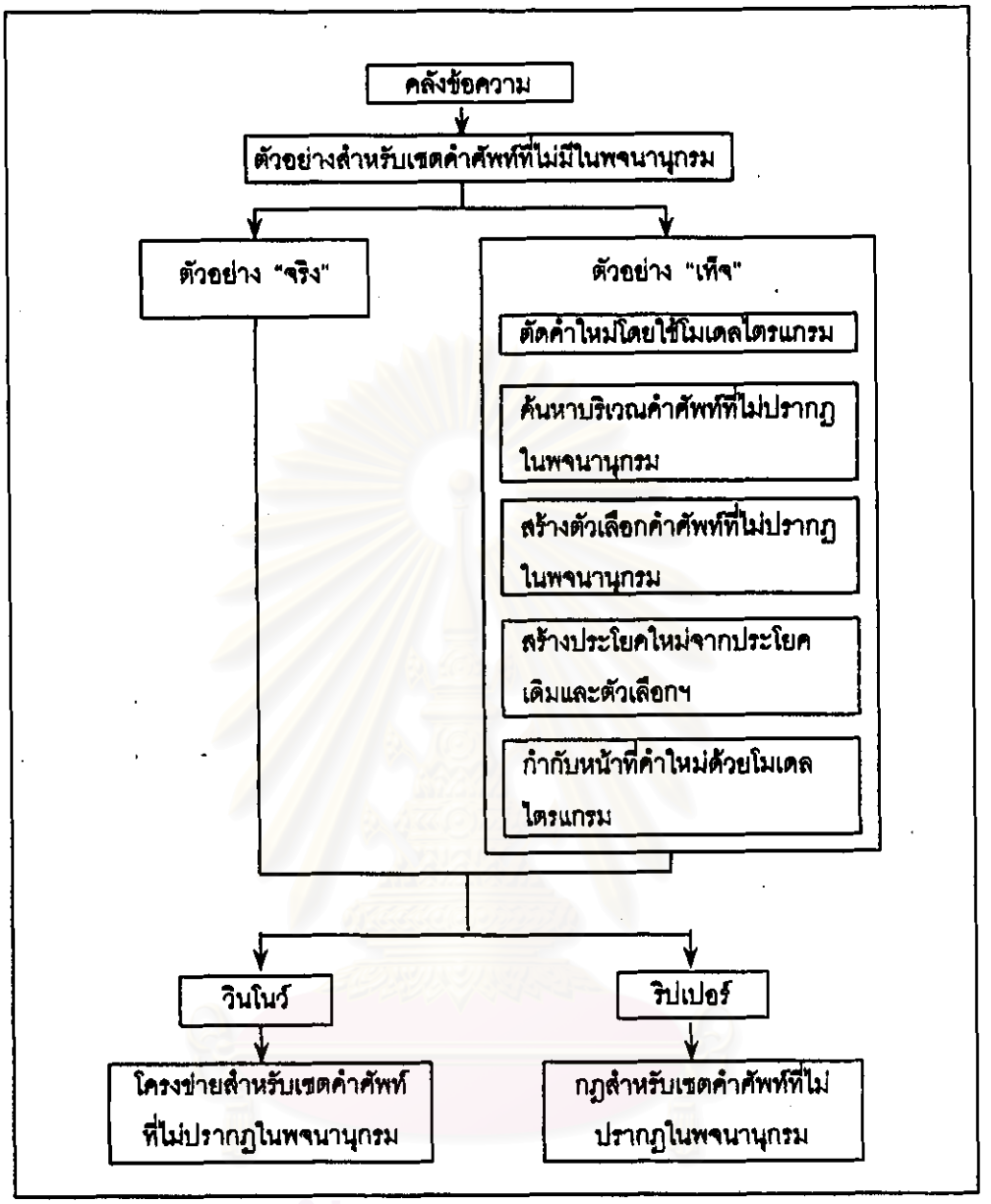
ข้อดีสำหรับวิธีการนี้คือ จะสามารถแก้ปัญหาของข้อความกำกวมที่ไม่เคยพบมาได้ เช่น สมมุติว่าพบข้อความ "มากรอง" เป็นข้อความกำกวมที่ไม่เคยพบมาก่อน ซึ่งวิธีนี้จะมีการสร้างเซตข้อความส่วนหน้าของคำว่า "มาก" (แสดงในตัวอย่างที่ 7-2) ดังนั้นวิธีการนี้จะนำเซตข้อความส่วนหน้าของคำว่า "มาก" เข้ามาใช้ในการแก้ปัญหานี้ ซึ่งเมื่อนำเซตข้อความส่วนหน้าของคำว่า "มาก" เข้ามาใช้ก็จะสามารถระบุได้ว่าข้อความนี้ควรจะตัดคำแรกให้เป็น "มา" หรือ "มาก" โดยที่ผลลัพธ์ที่ได้อาจจะเป็น "มากรอง" หรือ "มาก รอง" ซึ่งจะขึ้นอยู่กับคุณลักษณะต่างๆ ของข้อความนี้ แต่ถ้าใช้เซตสับสนแก้ปัญหาความกำกวมนี้ จะต้องมีการสร้างเซตสับสนของข้อความ "มากรอง" ก่อนถึงจะแก้ปัญหานี้ได้

7.3 การแก้ไขปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

จากบทที่ 5 ได้มีการอธิบายถึงลักษณะต่างๆ ของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมไปแล้ว ซึ่งจะเห็นได้ว่าคำศัพท์ประเภทนี้สามารถเกิดขึ้นได้หลายรูปแบบ และไม่มีกฎเกณฑ์ที่แน่นอน ทำให้การหาขอบเขตของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนี้เป็นไปได้ยาก ในวิทยานิพนธ์นี้จะนำเอาคุณลักษณะ เข้ามาใช้ในการหาขอบเขตของคำศัพท์ที่ไม่ปรากฏในพจนานุกรม ซึ่งคุณลักษณะต่างๆ เหล่านั้นได้มาจากการเรียนรู้ของเครื่อง ริปเปอร์ และ วินโนว์

จากวิธีการแก้ปัญหาความกำกวมนั้นจะทำการสร้างเซตสับสนหรือเซตข้อความส่วนหน้า หลังจากนั้นนำคำบริบทและสิ่งที่เกิดร่วมกันโดยมีลำดับของข้อความที่กำกวมนั้น มาส่งให้กับวินโนว์หรือริปเปอร์ทำการตรวจสอบและตัดสินว่าการตัดคำที่ถูกควรจะเป็นเช่นไร สำหรับการแก้ปัญหาเรื่องคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้นจะมีการนำตัวอย่างของคำศัพท์ที่ไม่ปรากฏในพจานุกรมที่เป็นจริง และตัวอย่างที่เป็นเท็จเข้ามาประยุกต์ใช้ในการแก้ปัญหานี้

สำหรับขั้นตอนการสร้างการเรียนรู้ให้กับวินโนว์และริปเปอร์ในการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมโดยใช้ตัวอย่างจริงกับตัวอย่างเท็จแสดงในรูปที่ 7-2 โดยที่ขั้นตอนต่างๆ จะประกอบด้วยขั้นตอนดังต่อไปนี้



รูปที่ 7-2 ขั้นตอนการเรียนรู้คุณลักษณะเพื่อนำมาใช้ในการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

- เลือกประโยคที่มีชื่อเฉพาะจากคลังข้อความมาเป็นตัวอย่าง
- สำหรับการเรียนรู้คุณลักษณะของตัวอย่างจริงนั้น ให้นำคำบริบทและสิ่งที่เกิดร่วมกันโดยมีลำดับรอบๆ ชื่อเฉพาะมาเป็นตัวอย่างให้กับการเรียนรู้เครื่องริปเปอร์และวินโนว์

- สำหรับการเรียนรู้คุณลักษณะของตัวอย่างเท็จ ขั้นแรกต้องมีการสร้างตัวเลือกของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมที่มีขอบเขตไม่ถูกต้อง แล้วจึงนำคำบริบทและสิ่งที่เกิดร่วมกันโดยมีลำดับรอบๆ ตัวเลือกที่สร้างขึ้นมาเป็นตัวอย่างให้กับการเรียนรู้ของเครื่องริปเปอร์และวินโนว์ โดยขั้นตอนการสร้างจะประกอบไปด้วยดังนี้
    - นำประโยคมาทำการตัดคำใหม่ โดยใช้โตรแกรมโมเดล ตามที่ได้กล่าวไปแล้วในหัวข้อ 2.3.2
    - ทำการค้นหาบริเวณที่น่าจะเกิดคำที่ไม่ปรากฏในพจนานุกรม
    - สร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมที่มีขอบเขตคำ ไม่ถูกต้อง
    - สร้างประโยคใหม่จากประโยคเดิมที่ได้มีการเปลี่ยนชื่อเฉพาะให้เป็นตัวเลือกอื่นๆ ที่สร้างขึ้นมา
    - กำกับหน้าที่คำโดยใช้โตรแกรมโมเดล สำหรับประโยคใหม่ที่ได้สร้างขึ้น ส่วนรายละเอียดนั้นได้กล่าวไว้ในบทที่ 3

สำหรับรายละเอียดของขั้นตอนต่างๆ ในการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้น จะกล่าวถึงในส่วนถัดไป

7.3.1 การสร้างตัวอย่างจริงกับตัวอย่างเท็จ

สำหรับการแก้ปัญหาเรื่องคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้น จะสร้างตัวอย่างจริงและตัวอย่างเท็จเพื่อนำไปให้การเรียนรู้ของเครื่องริปเปอร์หรือวินโนว์เรียนรู้คุณลักษณะต่างๆ ที่สามารถนำมาใช้ในการจำแนกระหว่างตัวอย่างจริงและตัวอย่างเท็จได้ ส่วนสาเหตุที่ต้องการให้ริปเปอร์หรือวินโนว์เรียนรู้คุณลักษณะในการจำแนกระหว่างตัวอย่างจริงกับตัวอย่างเท็จนั้น เนื่องจากคำศัพท์ที่ไม่ปรากฏในพจานานุกรมนั้นสามารถจะเกิดจากการประสมระหว่างคำได้หลายแบบ ดังนั้นในงานวิทยานิพนธ์นี้จะทำการค้นหาจุดที่น่าสงสัยที่มีโอกาสจะเกิดคำศัพท์ที่ไม่ปรากฏในพจานุกรมขึ้น แล้วจะทำการสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมขึ้นโดยการประสมระหว่างจุดที่น่าสงสัยกับคำบริเวณรอบๆ ในหลายรูปแบบ แล้วค่อยนำบริบทตัวเลือกนั้นมาป้อนให้กับริปเปอร์หรือวินโนว์ เพื่อค้นหาตัวเลือกที่มีโอกาสเป็นคำศัพท์ที่ไม่ปรากฏในพจนานุกรมมากที่สุด สำหรับการค้นหาจุดที่น่าสงสัยนั้นจะอธิบายในหัวข้อ 7.3.2

ในการสร้างตัวอย่างจริงให้กับการเรียนรู้เครื่องริปเปอร์หรือวินโนว์ คือนำบริบทของชื่อเฉพาะมาเป็นตัวอย่างจริง เนื่องจากวิทยานิพนธ์นี้จะทำการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมที่เป็นชื่อเฉพาะเท่านั้น ดังนั้นจึงพิจารณาให้ชื่อเฉพาะนั้น ส่วนตัวอย่างเท็จนั้นก็จะนำมาบริบทของตัวเลือกที่ไม่ปรากฏพจนานุกรมที่สร้างขึ้นมาจากชื่อเฉพาะเหล่านั้น ส่วนในรายละเอียดของการสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้นแสดงในหัวข้อ 7.3.3

7.3.2. การค้นหาบริเวณที่น่าสงสัย

เนื่องจากการเกิดคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้น มีโอกาสจะเกิดขึ้นค่อนข้างมาก โดยเฉพาะชื่อเฉพาะต่างๆ และคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้นสามารถมีได้หลายรูปแบบดังที่ได้กล่าวไปแล้วในบทที่ 5 โดยเฉพาะคำศัพท์ที่ไม่ปรากฏในพจนานุกรมแบบซ่อนเร้นทุกส่วนนั้นค่อนข้างจะตรวจสอบยากว่าบริเวณนั้นเป็นคำศัพท์ที่ไม่ปรากฏในพจนานุกรม ดังนั้นในวิทยานิพธ์นี้จะเสนอวิธีการตรวจสอบหาบริเวณที่มีโอกาสเกิดคำที่ไม่ปรากฏในพจนานุกรม ตามลักษณะประเภทของคำที่ไม่ปรากฏในพจนานุกรมดังต่อไปนี้

➢ บริเวณที่น่าสงสัยสำหรับคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างชัดเจนและอย่างซ่อนเร้นบางส่วน เนื่องจากลักษณะของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมประเภทนี้จะมีข้อความที่ไม่มีในพจนานุกรมเป็นส่วนประกอบ ดังนั้นการค้นหาบริเวณที่น่าจะเกิดคำศัพท์ที่ไม่ปรากฏในพจนานุกรมประเภทดังกล่าวทำได้โดยการค้นหาบริเวณที่ทำการตัดคำแล้วเกิดข้อความที่ไม่มีในพจนานุกรม

➢ บริเวณที่น่าสงสัยสำหรับคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นทุกส่วน เนื่องจากลักษณะของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมประเภทนี้ เกิดจากการประสมระหว่างคำศัพท์ที่ปรากฏในพจนานุกรม ทำให้การหาบริเวณที่น่าสงสัยสำหรับคำศัพท์ประเภทนี้ทำได้ยากกว่าแบบแรก ดังนั้นในวิทยานิพนธ์นี้จึงเสนอวิธีการค้นหาบริเวณที่น่าสงสัยสำหรับคำศัพท์ประเภทนี้ โดยมีวิธีการดังต่อไปนี้

1. หาค่าความน่าจะเป็นของ $P(w_i|t_i)$ ที่มีค่าน้อยกว่าค่าขีดเริ่มเปลี่ยน (Threshold)
2. หาค่าความน่าจะเป็นของ $P(t_i|t_{i-1},t_{i-2})$ ที่มีค่าน้อยกว่าค่าขีดเริ่มเปลี่ยน

เมื่อพบบริเวณที่น่าสงสัยว่าจะเกิดคำศัพท์ที่ไม่ปรากฏในพจนานุกรมแล้ว ขั้นตอนต่อไปก็คือการสร้างตัวเลือกของคำศัพท์ที่ไม่ปรากฏในพจนานุกรมซึ่งจะอธิบายในส่วนถัดไป

7.3.3 การสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรม (Generating Unknown Word Candidate)

เมื่อพบจุดที่น่าส่งสัยในการเกิดคำที่ไม่ปรากฏในพจนานุกรมตามวิธีการที่ได้กล่าวไปแล้วนั้น ขั้นตอนต่อไปคือ การสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรม โดยสาเหตุที่ต้องมีการสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรม เนื่องจากงานวิทยานิพนธ์นี้จะต้องหาขอบเขตของคำที่ไม่ปรากฏในพจนานุกรม โดยที่คำศัพท์ประเภทนี้สามารถจะเกิดขึ้นได้หลายรูปแบบ และไม่มีกฎเกณฑ์แน่นอนในการหาขอบเขตของคำศัพท์ที่ไม่ปรากฏในพจนานุกรม ดังนั้นในขั้นตอนนี้จะทำการสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมที่เป็นไปได้ทั้งหมดก่อน หลังจากนั้นจะนำคุณลักษณะที่ได้จากการเรียนรู้ของริปเปอร์หรือวินโนว์ มาประยุกต์ใช้ในการเลือกตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมที่ดีที่สุด

สำหรับวิธีการสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้น สามารถแบ่งออกได้เป็น 2 วิธี ตามประเภทของคำศัพท์ที่ไม่ปรากฏในพจนานุกรม แสดงดังต่อไปนี้

➢ การสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมสำหรับคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างชัดเจนและอย่างซ่อนเร้นบางส่วน แสดงดังรูปที่ 7-3 (ค่า K ที่ใช้ในการทดลองเท่ากับ 4)

กำหนดให้ประโยคคือ $w_1w_2...w_aUw_b...w_n$ โดยที่ $w_i \in$พจนานุกรม $U \notin$ พจนานุกรม และ n คือจำนวนคำในประโยค

$UNK = \{\alpha U\beta \mid \alpha \in A, \beta \in B\}$ โดยที่ UNK คือเซตของตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

$A = \{w_{a-i,a-1}, i \in [0,K]\} \cup \{\varepsilon\}$

$B = \{w_{b+1,b+i}, i \in [0,K]\} \cup \{\varepsilon\}$

$w_{i,j} = w_i..w_j : i<j$

ε คือข้อความที่ว่าง (Null string) และ K คือค่าคงที่

รูปที่ 7-3 สมการการสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างชัดเจนและอย่างซ่อนเร้นบางส่วน

➢ การสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมสำหรับคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นทุกส่วน ดังแสดงในรูปที่ 7-4 (ค่า K ที่ใช้ในการทดลองเท่ากับ 4)

7.3.4 การสร้างประโยคใหม่

หลังจากการค้นหาบริเวณที่น่าสงสัยว่าจะเกิดคำศัพท์ที่ไม่ปรากฏในพจนานุกรม และได้ทำการสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมแล้ว ขั้นตอนต่อไปก็คือการสร้างประโยคใหม่จากประโยคเดิมโดยให้นำตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมที่สร้างขึ้นมา มาทำการแทนที่บริเวณที่น่าจะเกิดคำศัพท์ที่ไม่ปรากฏในพจนานุกรม แล้วค่อยนำไปกำกับหน้าที่คำโดยใช้โมเดลไตรแกรม สุดท้ายทำการส่งไปให้การเรียนรู้ของเครื่องวินโนว์หรือริปเปอร์

ตัวอย่างเช่นประโยค นางเจนนีไปเดินเล่น เมื่อทำการตัดคำโดยใช้โมเดลไตรแกรมจะได้เป็น "นาง/NTTL เจน/VACT นี/NPRP ไป/XVAE เดิน/VACT เล่น/VACT" โดยหน้าที่คำ (NTTL, VACT ฯลฯ)สามารถดูรายละเอียดได้ในภาคผนวก ข. สำหรับผลลัพธ์จากการตัดคำจะเห็นว่า "นี" เป็นข้อความที่ไม่

กำหนดให้ประโยคคือ $w_1 w_2 \ldots w_a \ldots w_n$ โดยที่ $w_i \in$ พจนานุกรม w_a คือคำที่มีความน่าจะเป็นต่ำกว่าค่าขีดจำกัดของค่า $P(w_i|t_i)$ หรือ $P(t_i|t_{i-1},t_{i-2})$ ตามที่ได้อธิบายใน 7.3.2.2 และ n คือจำนวนคำในประโยค

$UNK = \{ \alpha W \beta \mid \alpha \in A, \beta \in B \}$ โดยที่ UNK คือเซตของตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

$A = \{w_{a-i,a-1}, i \in [0,K]\} \cup \{\varepsilon\}$

$B = \{w_{a+1,a+i}, i \in [0,K]\} \cup \{\varepsilon\}$

$w_{i,j} = w_i..w_j : i<j$

$W = w_a$ ถ้า $P(w_i|t_i) <$ ค่าขีดจำกัด หรือ

$W \in \{w_a, w_{a-1}, w_{a-2}\}$ ถ้า $P(t_i|t_{i-1},t_{i-2}) <$ ค่าขีดจำกัด

ε คือข้อความที่ว่าง (Null string) และ K คือค่าคงที่

รูปที่ 7-4 สมการการสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นทุกส่วน

เกิดขึ้นในพจนานุกรม ดังนั้นบริเวณนี้จะต้องเกิดคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างแน่นอน ทำให้ต้องมีการสร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรม ดังนั้นเมื่อใช้สมการตามรูปที่ 7-3 โดยค่า K ที่ใช้มีค่าเท่า 2 ดังนั้นเซตตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรมจะได้ดังนี้ {นี, นีไป, นีไปเดิน, เจนนี, เจนนีไป, เจนนีไปเดิน, นางเจนนี, นางเจนนีไป}. ดังนั้นประโยคใหม่ที่ได้จากการแทนที่ตัวเลือกลงไปในประโยคเดิม ก็จะได้ดังต่อไปนี้

1. นาง เจ นี ไป เดิน เล่น
2. นาง เจ นีไป เดิน เล่น
3. นาง เจ นีไปเดิน เล่น
4. นาง เจนนี ไป เดิน เล่น
5. นาง เจนนีไป เดิน เล่น
6. นาง เจนนีไปเดิน เล่น
7. นางเจนนี ไป เดิน เล่น
8. นางเจนนีไป เดิน เล่น

เมื่อได้ประโยคใหม่ดังที่ได้แสดงไปแล้ว หลังจากนั้นให้นำแต่ละแบบของการตัดคำส่งไปให้กำกับหน้าที่คำโดยใช้โมเดลไตรแกรม แล้วจึงนำส่งไปให้ริปเปอร์หรือวินโนว์ในการเลือกประโยคที่เหมาะสมที่สุด