บทที่ 8

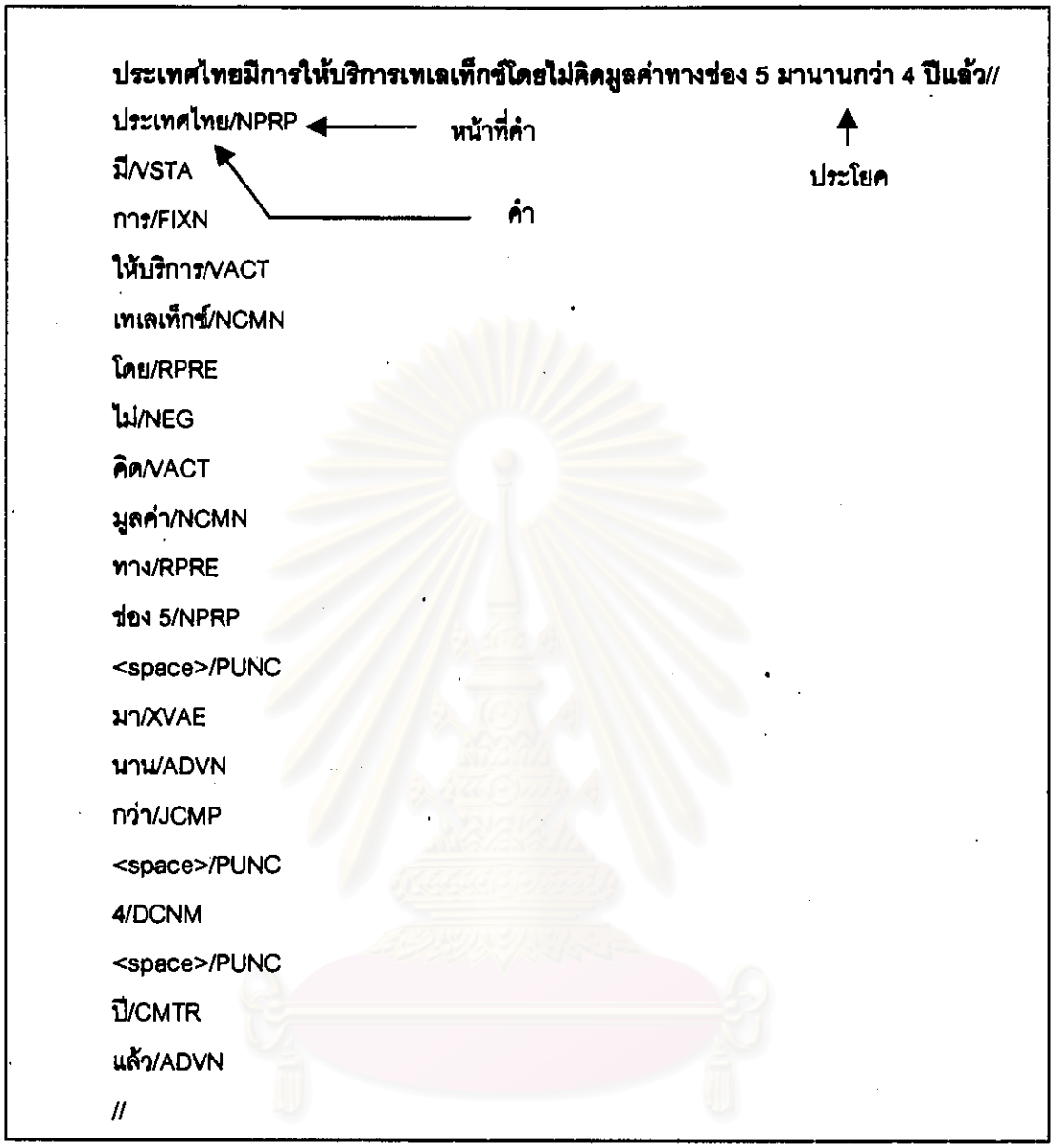
ประสิทธิภาพการตัดคำโดยใช้คุณลักษณะ

จากบทที่แล้วได้มีการอธิบายขั้นตอนการนำเอาการเรียนรู้ของเครื่องในแบบต่างๆ เข้ามาใช้ในการเรียนรู้คุณลักษณะต่างๆ ที่สามารถนำมาใช้แก้ปัญหาความกำกวม และคำศัพท์ที่ไม่ปรากฏในพจนานุกรม ในบทนี้จะแสดงถึงผลการทดลองเปรียบเทียบระหว่างการเรียนรู้ของเครื่องริปเปอร์, วินโนว์ และมีการเปรียบเทียบการแก้ปัญหาความกำกวมระหว่างการใช้คุณลักษณะกับวิธีการต่างๆ ที่ผ่านมา

คลังข้อความที่นำมาใช้ในการเรียนรู้ของเครื่องได้นำมาจาก คลังข้อความออร์คิด (Orchid Corpus) (Virach Sornlertlamvanich et al., 1997) โดยได้รับความอนุเคราะห์จาก ห้องปฏิบัติการวิจัยและพัฒนาวิศวกรรมภาษาและซอฟต์แวร์ (Software and Language Engineering Laboratory: SLL) ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (National Electronics and Computer Technology Center: NECTEC) โดยลักษณะของบทความที่นำมาใช้สร้างคลังข้อความนั้นได้นำมาจากรายงานการประชุมวิชาการของศูนย์ฯ เอง และจำนวนคลังข้อความนั้นมีอยู่ประมาณ 25,000 ประโยค โดยที่ภายในคลังข้อความนี้ได้ทำการแบ่งเป็นประโยค ส่วนภายในประโยคจะแบ่งเป็นคำต่างๆ และยังได้มีการกำหนดหน้าที่คำด้วย ซึ่งทั้งหมดทำโดยนักภาษาศาสตร์ ตัวอย่างประโยคภายในคลังข้อความออร์คิด แสดงในรูปที่ 8-1

จากรูปที่ 8-1 จะแสดงตัวอย่างประโยค "ประเทศไทยมีการให้บริการเทเลเท็กซ์โดยไม่คิดมูลค่าทางช่อง 5 มานานกว่า 4 ปีแล้ว" ซึ่งประโยคนี้ได้ทำการตัดคำและกำกับหน้าที่ของคำเรียบร้อยแล้ว โดยภายในหนึ่งบรรทัดจะประกอบไปด้วย 1 คำและจะบอกด้วยว่าคำนี้ทำหน้าที่อะไรอยู่ภายในประโยคนี้ ตัวอย่างเช่น "ประเทศไทย/NPRP" ในบรรทัดที่ 2 จะบอกว่าคำว่า "ประเทศไทย" มีหน้าที่คำเป็น NPRP ซึ่ง NPRP เป็นตัวอักษรย่อที่ใช้บอกหน้าที่คำที่เป็นคำนามประเภทชื่อเฉพาะ โดยสัญลักษณ์ตัวย่อต่างๆ เหล่านี้ สามารถดูได้ในภาคผนวก ข

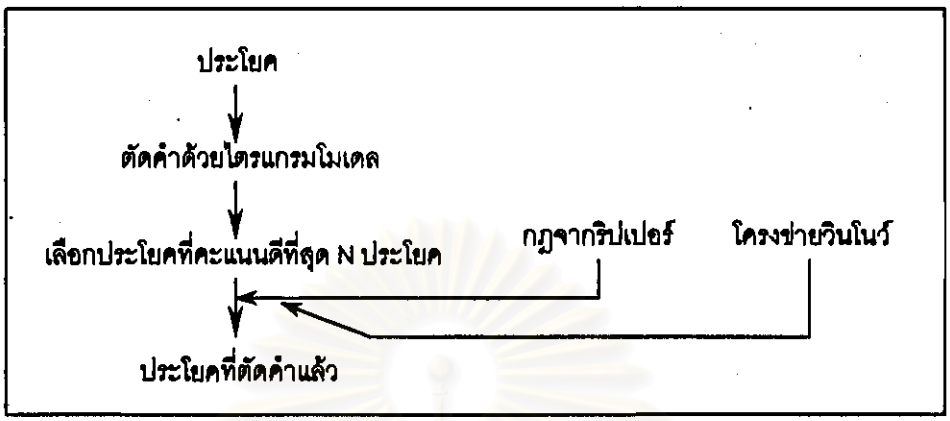
หน้าที่คำที่นำมาใช้ในคลังข้อความนี้ ได้ถูกแบ่งออกเป็น 47 หมวด (Virach Sornlertlamvanich et al., 1997) ซึ่งทำการแบ่งโดยนักภาษาศาสตร์ การที่ต้องแบ่งหน้าที่คำให้ละเอียดลงไปนั้น เนื่องจากถ้าแบ่งหน้าที่คำโดยแค่แบ่งเป็น คำนาม คำกริยา คุณศัพท์ ฯลฯ เท่านั้นจะไม่เพียงพอต่อการนำมาใช้ในการวิเคราะห์ทางภาษา จึงทำให้นักภาษาศาสตร์ได้มีการวิเคราะห์และแบ่งหน้าที่ของคำออกมาเป็นหมวดหมู่ต่างๆ



รูป 8-1 ตัวอย่างประโยคที่ทำการตัดคำและกำกับหน้าที่ของคำภายในคลังข้อความออร์คิด

สำหรับการทดลองวัดประสิทธิภาพของการเรียนรู้ของเครื่องวิปเปอร์กับวินโนว์ เพื่อที่จะนำมาใช้ในการเรียนรู้คุณลักษณะต่างๆ ในการแก้ไขปัญหาความกำกวม ในการทดลองนี้ได้เลือกข้อความที่กำกวมที่เกิดขึ้นบ่อยจากคลังข้อความออร์คิด แล้วทำการสร้างเซตลับสนและเซตข้อความส่วนหน้าสำหรับข้อความกำกวมที่เลือกขึ้นมา จากนั้นนำประโยคต่างๆ จากคลังข้อความออร์คิดมาใช้ในการเรียนรู้และทดสอบ โดยแบ่งประโยคที่นำมาใช้นั้นออกเป็น 2 ส่วน ส่วนแรกจะเป็นชุดสอน (Training Set) จำนวน 80% เพื่อที่จะนำไปให้ วิปเปอร์กับวินโนว์ ใช้ในการเรียนรู้ ส่วนที่สองจะเป็นชุดทดสอบ (Test Set) จำนวน 20% เพื่อใช้ทดสอบดูประสิทธิภาพของการเรียนรู้ของ วิปเปอร์ และวินโนว์ เมื่อนำมาใช้กับข้อมูลที่ไม่ได้มีการนำไปใช้ในการสร้างกฎ

8.1 ขั้นตอนการนำคุณลักษณะเข้ามาใช้ในการแก้ปัญหาความกำกวม



รูปที่ 8-1 ขั้นตอนการแก้ไขปัญหาความกำกวมโดยใช้คุณลักษณะ

ขั้นตอนการตัดคำโดยใช้คุณลักษณะมาแก้ไขปัญหาความกำกวม ซึ่งขั้นตอนมีดังต่อไปนี้คือ (แสดงดังรูปที่ 8-1)

1. นำประโยคมาตัดคำโดยใช้ไตรแกรมโมเดล
2. เลือกประโยคที่คะแนนดีที่สุดจำนวน N ประโยค
3. นำกฎจากริปเปอร์ หรือโครงข่ายวินโนว์มาช่วยใช้ในการแก้ปัญหา

8.2 ผลการทดลองแก้ปัญหาความกำกวม

สำหรับตัวอย่างข้อความกำกวมต่างๆ ที่นำมาใช้ในการทดลองนั้นจะนำมาจากคลังข้อความออร์คิด โดยจะเลือกข้อความกำกวมที่เกิดขึ้นจำนวนมากในคลังข้อความ สำหรับรายละเอียดความถี่ของข้อความกำกวมต่างๆ สามารถแสดงดังในภาคผนวก ค.

การทดลองแก้ปัญหาความกำกวมนั้นได้แบ่งการทดลองออกเป็นดังนี้

1. การทดลองการแก้ไขปัญหาความกำกวม แบบที่ต้องใช้บริบท โดยใช้การแก้ปัญหาแบบเซตสับสน ซึ่งผลการทดลองแสดงในตารางที่ 8-1

2. การทดลองการแก้ไขปัญหาความกำกวม แบบที่ต้องใช้บริบท โดยใช้การแก้ปัญหาแบบเซตข้อความส่วนหน้า ซึ่งผลการทดลองแสดงในตารางที่ 8-2

3. การทดลองการแก้ไขปัญหาความกำกวม แบบที่ไม่ต้องใช้บริบท โดยใช้การแก้ปัญหาแบบเซตสับสน ซึ่งผลการทดลองแสดงในตารางที่ 8-3

4. การทดลองการแก้ไขปัญหาความกำกวมแบบที่ไม่ต้องใช้บริบท โดยใช้การแก้ปัญหาแบบเซตข้อความส่วนหน้า ซึ่งผลการทดลองแสดงในตารางที่ 8-4

8.3 สรุปผลการทดลองการแก้ปัญหาความกำกวม

จากการแก้ปัญหาความกำกวมตามตารางที่ 8-1, 8-2, 8-3 และ 8-4 แสดงให้เห็นว่าการนำคุณลักษณะเข้ามาใช้ในการแก้ปัญหาความกำกวมไม่ว่าจะเป็นความกำกวมแบบที่ต้องใช้บริบทหรือไม่ต้องใช้บริบทนั้นสามารถนำมาแก้ปัญหาได้ดีกว่าวิธีการเดิมคือวิธีการตัดคำโดยใช้โมเดลไตรแกรม และการตัดคำโดยเลือกแบบเหมือนมากที่สุด และคุณลักษณะที่ได้จากการเรียนรู้ของเครื่องวินโนว์จะให้ความถูกต้องมากกว่าริปเปอร์ เมื่อนำมาใช้ในการแก้ปัญหาความกำกวมทั้งสองประเภท ไม่ว่าจะเป็นการใช้เซตสับสนหรือเซตข้อความส่วนหน้า

การแก้ปัญหาสำหรับความกำกวมที่ต้องใช้บริบทนั้นจะมีความถูกต้องน้อยกว่าความกำกวมที่ไม่ต้องใช้บริบท โดยนำเซตสับสนมาประยุกต์ใช้ในการแก้ปัญหาความกำกวมทั้งสองแบบจะให้ผลความถูกต้องมากกว่าการนำเซตข้อความส่วนหน้ามาใช้

ดังนั้นจากผลการทดลองสามารถสรุปผลได้ว่าการนำคุณลักษณะเข้ามาใช้ในการแก้ปัญหาความกำกวมนั้นสามารถจะแก้ปัญหาได้ดีกว่าวิธีการตัดคำโดยใช้โมเดลไตแกรม และการตัดคำโดยเลือกแบบเหมือนมากที่สุด และวินโนว์มีประสิทธิภาพในการแก้ปัญหาความกำกวมได้ดีกว่าริปเปอร์

ตารางที่ 8-1 ตารางแสดงประสิทธิภาพการแก้ไขปัญหาความกำกวม แบบที่ต้องใช้บริบท (Context Dependent)
โดยใช้การแก้ปัญหาแบบเซตสับสน

ข้อความที่กำกวม	เซตสับสน (Confusion Set)	วินโนว์		ริปเปอร์		โมเดลไตรแกรม	แบบเหมือนมากที่สุด
		ชุดสอน	ชุดทดสอบ	ชุดสอน	ชุดทดสอบ		
จัดการ	{ จัดการ , จัด การ }	100.00 %	98.03 %	99.61 %	92.06 %	84.55 %	91.82 %
หรือไม่	{ หรือไม่ , หรือ ไม่ }	100.00 %	92.07 %	96.20 %	79.26 %	83.54 %	68.35 %
ให้การ	{ ให้การ , ให้ การ }	100.00 %	94.37 %	95.54 %	89.29 %	64.29 %	10.71 %
ที่อยู่	{ ที่อยู่ , ที่ อยู่ }	100.00 %	90.09 %	95.12 %	84.21 %	74.04 %	18.27 %
พัฒนาการ	{ พัฒนาการ , พัฒนา การ }	100.00 %	93.10 %	95.45 %	69.75 %	67.47 %	18.07 %
ที่เกิด	{ ที่เกิด , ที่ เกิด }	100.00 %	99.71 %	94.29 %	93.75 %	46.59 %	6.82 %
ทางการ	{ ทางการ , ทาง การ }	100.00 %	84.06 %	98.15 %	69.23 %	43.48 %	30.43 %
คุ้มค่า	{ คุ้มค่า , คุ้ม ค่า }	100.00 %	95.36 %	93.75 %	100.00 %	95.00 %	95.00 %
ที่ตั้ง	{ ที่ตั้ง , ที่ ตั้ง }	100.00 %	96.38 %	100.00 %	80.00 %	53.85 %	42.31 %

ตารางที่ 8-2 ตารางแสดงประสิทธิภาพการแก้ไขปัญหาความกำกวม แบบที่ต้องใช้บริบท (Context Dependent)
โดยใช้การแก้ปัญหาแบบเซตข้อความส่วนหน้า

ข้อความที่กำกวม	เซตข้อความส่วนหน้า (Prefix Set)	วินโนว์		ริปเปอร์		โมเดลไตรแกรม	แบบเหมือนมากที่สุด
		ชุดสอน	ชุดทดสอบ	ชุดสอน	ชุดทดสอบ		
จัดการ	{ จัด , จัดการ }	100.00 %	94.52 %	83.33 %	69.28 %	84.55 %	91.82 %
หรือไม่	{ หรือ , หรือไม่ }	100.00 %	93.27 %	99.58 %	72.62 %	83.54 %	68.35 %
ให้การ	{ ให้ , ให้การ }	100.00 %	91.71 %	0.00 %	87.45 %	64.29 %	10.71 %
ที่อยู่	{ ที่ , ที่อยู่ }	100.00 %	90.97 %	99.97 %	56.98 %	74.04 %	18.27 %
พัฒนาการ	{ พัฒนา , พัฒนาการ }	100.00 %	89.11 %	99.83 %	65.94 %	67.47 %	18.07 %
ที่เกิด	{ ที่ , ที่เกิด }	100.00 %	91.41 %	0.00 %	75.01 %	46.59 %	6.82 %
ทางการ	{ ทาง , ทางการ }	100.00 %	89.68 %	99.80 %	89.19 %	43.48 %	30.43 %
คุ้มค่า	{ คุ้ม , คุ้มค่า }	100.00 %	93.03 %	90.62 %	94.76 %	95.00 %	95.00 %
ที่ตั้ง	{ ที่ , ที่ตั้ง }	100.00 %	88.79 %	91.04 %	99.98 %	83.26 %	42.31 %

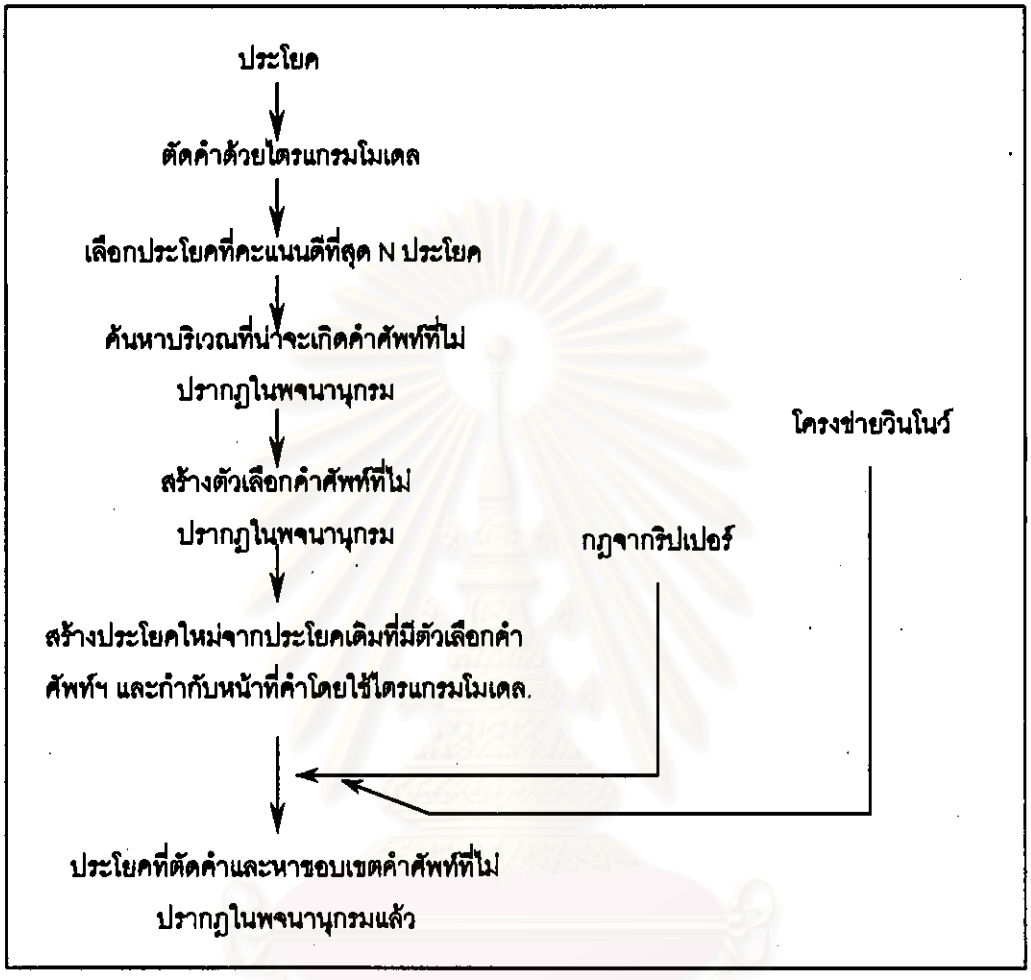
ตารางที่ 8-3 ตารางแสดงประสิทธิภาพการแก้ไขปัญหาความกำกวม แบบที่ไม่ต้องใช้บริบท (Context Independent) โดยใช้การแก้ปัญหาแบบเซตสับสน

ข้อความที่กำกวม	เซตสับสน (Confusion Set)	วินโนว์		วิปเปอร์		โมเดลไตรแกรม	แบบเหมือนมากที่สุด
		ชุดสอน	ชุดทดสอบ	ชุดสอน	ชุดทดสอบ		
ข้อมูล	{ ข้อมูล , ข้อ มูล }	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
อบรม	{ อบรม , อบ รม }	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
หน้าที่	{ หน้าที่ , หน้า ที่ }	100.00 %	97.54 %	100.00 %	100.00 %	95.84 %	50.14 %
คุณภาพ	{ คุณภาพ , คุณ ภาพ }	100.00 %	100.00 %	100.00 %	100.00 %	98.67 %	74.27 %
กำลัง	{ กำลัง , กำ ลัง }	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
กำลังคน	{ กำลังคน , กำลัง คน , กำ ลัง คน }	100.00 %	98.37 %	98.37 %	94.23 %	100.00 %	46.29 %
ลงทุน	{ ลงทุน , ลง ทุน }	100.00 %	100.00 %	100.00 %	100.00 %	92.07 %	60.97 %
รายได้	{ รายได้ , ราย ได้ }	100.00 %	96.78 %	100.00 %	100.00 %	94.54 %	74.50 %
ความรู้	{ ความรู้ , ความ รู้ }	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	42.31 %
เพื่อให้	{ เพื่อให้ , เพื่อ ให้ }	100.00 %	97.06 %	100.00 %	98.78 %	53.85 %	54.97 %

ตารางที่ 8-4 ตารางแสดงประสิทธิภาพการแก้ไขปัญหาความกำกวม แบบที่ไม่ต้องใช้บริบท (Context Independent)
โดยใช้การแก้ปัญหาแบบเซตข้อความส่วนหน้า

ข้อความที่กำกวม	เซตข้อความส่วนหน้า (Prefix Set)	วินโนว์		ริปเปอร์		โมเดลไตรแกรม	แบบเหมือนมากที่สุด
		ชุดสอน	ชุดทดสอบ	ชุดสอน	ชุดทดสอบ		
ข้อมูล	{ ข้อ , ข้อมูล }	100.00 %	100.00 %	98.58 %	91.23 %	100.00 %	100.00 %
อบรม	{ อบ , อบรม }	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
หน้าที่	{ หน้า , หน้าที่ }	97.24 %	94.38 %	97.68 %	92.00 %	95.84 %	50.14 %
คุณภาพ	{ คุณ , คุณภาพ }	100.00 %	100.00 %	98.33 %	91.01 %	98.67 %	74.27 %
กำลัง	{ กำ , กำลัง }	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %	100.00 %
กำลังคน	{ กำ , กำลัง , กำลังคน }	99.54 %	90.74 %	98.37 %	90.46 %	100.00 %	46.29 %
ลงทุน	{ ลง , ลงทุน }	100.00 %	100.00 %	100.00 %	100.00 %	92.07 %	60.97 %
รายได้	{ ราย , รายได้ }	100.00 %	99.78 %	100.00 %	100.00 %	94.54 %	74.50 %
ความรู้	{ ความ , ความรู้ }	100.00 %	99.67 %	100.00 %	99.36 %	100.00 %	42.31 %
เพื่อให้	{ เพื่อ , เพื่อให้ }	100.00 %	97.46 %	100.00 %	94.17 %	53.85 %	54.97 %

8.4 ขั้นตอนการนำคุณลักษณะเข้ามาใช้ในการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม



รูปที่ 8-2 ขั้นตอนการแก้ไขปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมโดยใช้คุณลักษณะ

จากรูปที่ 8-2 แสดงขั้นตอนการทำงานของการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจานุกรมโดยใช้คุณลักษณะ ซึ่งจะมีขั้นตอนดังนี้คือ

1. นำประโยคมาทำการตัดคำโดยใช้ไตรแกรมโมเดล
2. เลือกประโยคที่ดีที่สุด N ประโยค
3. ทำการค้นหาบริเวณที่น่าจะเกิดคำศัพท์ที่ไม่ปรากฏในพจนานุกรม
4. สร้างตัวเลือกคำศัพท์ที่ไม่ปรากฏในพจนานุกรม
5. สร้างประโยคใหม่จากประโยคเดิมโดยนำตัวเลือกฯ ไปแทนที่
6. กำกับหน้าที่คำโดยไตรแกรมโมเดล
7. นำกฎจากริปเปอร์หรือโครงข่ายวินโนว์ เข้ามาใช้ในการเลือกตัวเลือกฯ ที่มีคะแนนมากที่สุด

8.5 ผลการทดลองการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

สำหรับการทดลองแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม จะแสดงดังตารางที่ 8-5 ซึ่งในตารางนี้จะวัดประสิทธิภาพในการเรียนรู้คุณลักษณะที่จะนำมาใช้ในการแก้ปัญหาระหว่างวินโนว์กับริปเปอร์ โดยจำนวนตัวอย่างที่ใช้ในการเรียนรู้ทั้งจะใช้ทั้งตัวอย่างที่ถูกและตัวอย่างที่ผิดจำนวน 1509 ตัวอย่างและ 9357 ตัวอย่างตามลำดับ สำหรับตัวอย่างที่ให้ทดสอบก็มีจำนวน 377 ตัวอย่างสำหรับตัวอย่างจริงและ 2337 ตัวอย่างสำหรับตัวอย่างเท็จ และสำหรับจำนวนคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างชัดเจนและอย่างซ่อนเร้นบางส่วนนั้นจะมีจำนวน 1235 ตัวอย่าง และจำนวนคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นทุกส่วนมีจำนวน 651 ตัวอย่าง โดยแบ่งออกเป็น 2 ส่วนคือ ส่วนแรก 80% สำหรับการสอน และอีก 20% สำหรับการทดสอบ

ตารางที่ 8-5 ตารางแสดงผลการทดลองการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

	คำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างชัดเจนและอย่างซ่อนเร้นบางส่วน		คำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นทุกส่วน	
	ชุดสอน	ชุดทดสอบ	ชุดสอน	ชุดทดสอบ
ค้นหาพบ	100.00 %	100.00 %	87.82 %	83.25 %
วินโนว์	95.26 %	92.75 %	72.87 %	68.21 %
ริปเปอร์	93.25 %	89.75 %	69.25%	65.03 %

จากผลการทดลองในตารางที่ 8-1 แสดงให้เห็นว่าการค้นหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างชัดเจนและอย่างซ่อนเร้นบางส่วนนั้น จะสามารถค้นหาได้ถูกต้องทั้งหมด ในขณะที่การค้นหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นทุกส่วนนั้นจะถูกค้นพบเพียงประมาณ 87% และสำหรับการเรียนรู้คุณลักษณะที่นำมาใช้ในการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้น วินโนว์จะให้ความถูกต้องสูงกว่าริปเปอร์สำหรับคำศัพท์ที่ไม่ปรากฏในพจนานุกรมทุกประเภท

8.6 สรุปผลการทดลองการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

วิธีการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม ภายในวิทยานิพนธ์นี้ จะสามารถแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมได้ทุกรูปแบบ โดยที่การค้นหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างชัดเจนและอย่างซ่อนเร้นบางส่วนจะสามารถค้นพบได้ทั้งหมด แต่สำหรับคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นทุกส่วนนั้นจะไม่สามารถค้นพบได้ทั้งหมด ส่วนการนำเอาวินโนว์กับริปเปอร์เข้ามาใช้ในการแก้ปัญหานั้นจะเห็นว่าวินโนว์สามารถแก้ปัญหาได้ดีกว่าริปเปอร์ทั้งในการแก้ปัญหาของคำศัพท์ที่ไม่ปรากฏใน

พจนานุกรมแบบอย่างชัดเจนและซ่อนเร้นบางส่วน กับคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นทุกส่วน