

บทที่ 5

ผลการทดลองและบทสรุป

ต่อไปนี้เป็นผลการทดลองที่ได้ทำการคัดแยกข้อความที่ต้องการจากบทความในสาขาสังคมวิทยาและสาขาไบโอเคมี โดยใช้หลักการของลิงค์แกรมมาพาสเซอร์ในการวิเคราะห์โครงสร้างของประโยคก่อนทำการคัดแยกข้อความจากโครงสร้างต้นไม้ของไวยากรณ์ และได้ทำการประเมินผลการคัดแยกข้อความด้วยค่าฟริชชันและค่ารีคอลซึ่งเป็นค่ามาตรฐานที่ใช้ในการประเมินประสิทธิภาพของระบบคัดแยกข้อมูล (Ralph Grishman, 1997) ได้ผลดังนี้

5.1 ผลการทดลอง

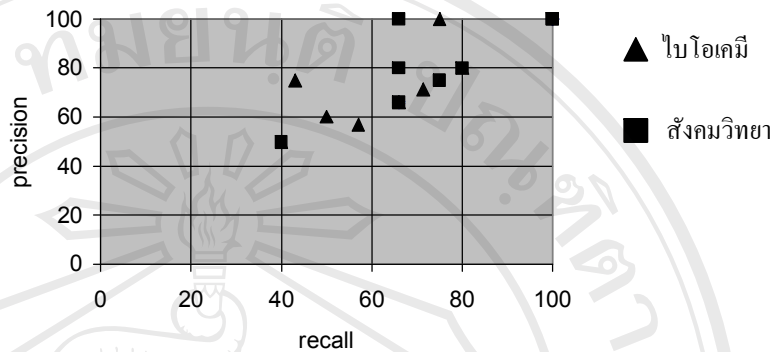
สำหรับชุดข้อมูลที่ใช้ในการทดสอบใช้ข้อมูล 2 ชุดด้วยกัน ประกอบด้วยบทคัดย่อของงานวิจัยในสาขาสังคมวิทยาเกี่ยวกับจริยธรรมทางธุรกิจและงานวิจัยในสาขาไบโอเคมี เป็นบทคัดย่อของงานวิจัยเกี่ยวกับจริยธรรมทางธุรกิจจำนวน 14 บทความ รวมทั้งหมด 109 ประโยคด้วยกัน พาสเซอร์ทำการวิเคราะห์ประโยคได้ 82 ประโยคด้วยกัน คิดเป็น 75.2 เปอร์เซ็นต์ของประโยคทั้งหมด เป็นประโยคที่ต้องการคัดแยกจำนวน 56 ประโยคจากประโยคทั้งหมดที่พาสเซอร์สามารถทำการวิเคราะห์ได้ สำหรับบทคัดย่อของงานวิจัยทางไบโอเคมีจำนวน 10 บทความ รวมทั้งหมด 87 ประโยค พาสเซอร์ทำการวิเคราะห์ประโยคได้ 62 ประโยค คิดเป็น 71.2 เปอร์เซ็นต์ของประโยคทั้งหมด เป็นประโยคที่ต้องการคัดแยกทั้งหมด 49 ประโยคจากประโยคทั้งหมดที่พาสเซอร์สามารถทำการวิเคราะห์ได้ ข้อความที่ต้องการคัดแยกนั้นมุ่งไปที่เนื้อหาสำคัญของงานวิจัยนั้น

สำหรับวิธีที่ใช้นี้ทำให้สามารถคัดแยกนามวลีที่ต้องการจากโครงสร้างของประโยคได้ โดยที่นามวลีที่ได้นั้นมีความหมายในตัวเองและสามารถคัดแยกส่วนขยายของประโยคได้ และเนื่องจากไม่มีชุดข้อมูลมาตรฐานที่มีคำตอบตรงตามจุดประสงค์ที่ต้องการดังนั้นในการทดลองจึงได้นำชุดข้อมูลที่ใช้ในการทดสอบนี้ไปให้กับผู้เชี่ยวชาญทางด้านภาษาช่วยในการตรวจสอบความถูกต้องของคำตอบที่ได้เพื่อใช้สำหรับการประเมินผลต่อไป ในการประเมินผลสำหรับการทดลองนั้นจะประเมินด้วยค่าฟริชชันและค่ารีคอล

$$\text{ค่าฟริชชัน} = \frac{\text{จำนวนข้อความที่ระบบสามารถคัดแยกมาได้อย่างถูกต้อง}}{\text{จำนวนข้อความที่เป็นคำตอบที่ถูกต้องทั้งหมด}} \times 100\%$$

$$\text{ค่ารีคอล} = \frac{\text{จำนวนข้อความที่ระบบสามารถคัดแยกมาได้อย่างถูกต้อง}}{\text{จำนวนข้อความที่ระบบสามารถทำการคัดแยกได้ทั้งหมด}} \times 100\%$$

ผลที่ได้จากการทดลองแสดงดังรูป 5.1 และสามารถสรุปค่าพรีซิชันและค่ารีคอลที่ได้จากการประเมินระบบการคัดแยกข้อมูลในงานวิจัยนี้ดังแสดงในตาราง 5.1



รูป 5.1 แสดงการประเมินผลลัพธ์ที่ได้จากการคัดแยก

ตาราง 5.1 แสดงค่าพรีซิชันและค่ารีคอล

ชุดข้อมูล	สังคัมวิทยา		ไบโอเคมี	
	Precision	Recall	Precision	Recall
ค่าประเมิน				
ค่าต่ำสุด	50 %	40 %	50 %	40 %
ค่าเฉลี่ย	80.3 %	76.1 %	69.7 %	61.2 %
ค่าสูงสุด	100 %	100 %	100 %	100 %

จากข้อมูลที่แสดงดังรูป 5.1 และตาราง 5.1 พบว่าชุดข้อมูลที่เป็นบทคัดย่อในสาขาสังคัมวิทยามีค่าพรีซิชันเฉลี่ย 80.3 เปอร์เซ็นต์ ซึ่งมากกว่าบทคัดย่อในสาขาไบโอเคมีที่มีค่าเฉลี่ย 69.7 เปอร์เซ็นต์ ในตาราง 5.1 นั้น แสดงค่าพรีซิชันเป็นซึ่งเป็นการวัดความถูกต้องของข้อความที่สามารถคัดแยกได้ ซึ่งจากค่าเฉลี่ยของค่าพรีซิชันดังกล่าวนี้แสดงให้เห็นว่าผลจากการคัดแยกข้อความจากบทคัดย่อในสาขาสังคัมวิทยานั้นมีความถูกต้องมากกว่าบทความในสาขาไบโอเคมี ซึ่งมีผลมาจากลักษณะของบทความที่มีความแตกต่างกัน สำหรับการประเมินข้อความที่คัดแยกได้ด้วยค่ารีคอลนั้นเป็นการวัดความสามารถในการสืบค้นคืนประโยคที่มีความเกี่ยวข้องกับคำสำคัญที่กำหนด ซึ่งพบว่าค่าเฉลี่ยของค่ารีคอลที่ได้จากการทดลองนั้นค่ารีคอลของข้อมูลในสาขาสังคัมวิทยามีค่าเฉลี่ย 76.1 เปอร์เซ็นต์ ซึ่งสูงกว่าข้อมูลทางไบโอเคมีเช่นกัน

เมื่อพิจารณาค่าเฉลี่ยของค่าพริชชันและค่ารีคอลที่ได้จากการประเมินระบบคัดแยกข้อความที่ได้นำเสนอในงานวิจัยครั้งนี้เปรียบเทียบกับค่าเฉลี่ยที่ได้จากงานวิจัยที่ได้กล่าวถึงในบทที่ 2 คือ ระบบฟาสต์ (Douglas E. Appelt, 1993) มีค่าพริชชันเฉลี่ย 52 เปอร์เซนต์ และค่ารีคอล 44 เปอร์เซนต์ และระบบการคัดแยกข้อความที่มีการจัดประเภทของคำ (Ellen Riloff, 2005) มีค่าพริชชันเฉลี่ย 56 เปอร์เซนต์ และค่ารีคอล 51 เปอร์เซนต์ ตามลำดับ ดังนั้นค่าพริชชันและค่ารีคอลที่ได้จากการประเมินในครั้งนี้ก็มีค่าอยู่ในเกณฑ์ที่น่าพอใจเมื่อเปรียบเทียบกับค่าพริชชันและค่ารีคอลที่ได้จากทั้งสองงานที่กล่าวถึงในบทที่ 2

สำหรับความแตกต่างของผลการประเมินที่ได้ทำการทดสอบกับบทคัดย่อทั้งสองประเภทในงานวิจัยนี้สามารถอธิบายได้ว่า ความผิดพลาดส่วนใหญ่ที่เกิดขึ้นกับการทดสอบโดยบทคัดย่อในสาขาไบโอเคมีนั้นมีสาเหตุเนื่องมาจากบทคัดยอดังกล่าวมีคำศัพท์ที่เป็นคำเฉพาะจำนวนมาก รวมทั้งมีการใช้สัญลักษณ์ทางคณิตศาสตร์ และสูตรเคมีต่าง ๆ ที่เป็นสัญลักษณ์และตัวเลข มีลักษณะของประโยคเป็นประโยคที่มีความยาวมากและมีความซับซ้อน ซึ่งลิงก์แกรมมาฟาสเซอร์ก็ยังมีข้อจำกัดเกี่ยวกับคำศัพท์เฉพาะต่าง ๆ และบทความที่มีการใช้สัญลักษณ์มาก ๆ สัญลักษณ์นั้นอาจมีผลในการวิเคราะห์แบ่งวลี ซึ่งอาจทำให้เกิดการวิเคราะห์ที่ผิดพลาดอันเนื่องมาจากการใช้เครื่องหมายได้ ทำให้ผลจากการคัดแยกนั้นผิดพลาด สำหรับบทคัดย่อในสาขาสังคมวิทยานั้นไม่มีการใช้คำศัพท์เฉพาะมากเท่าบทคัดย่อทางไบโอเคมี สำหรับบางส่วนที่ไม่ถูกต้องนั้นพบว่าเป็นประโยคที่มีความยาวมาก มีลักษณะเป็นประโยคความซ้อนหรือความรวมในการอธิบายความในบทคัดย่อ ซึ่งมีความซับซ้อนและมีบางครั้งมีการใช้เครื่องหมายต่าง ๆ เช่น จุลภาค อัฒภาค ทวิภาค ซึ่งจะมีผลต่อการวิเคราะห์โครงสร้างในส่วนของกรแบ่งกลุ่มของวลีทำให้การแบ่งกลุ่มผิดพลาดได้ หรือในบางประโยคที่มีการใช้คำเชื่อมต่าง ๆ เช่น “and” “or” นั้นมีผลทำให้พาสเซอร์ทำการวิเคราะห์โครงสร้างของประโยคนั้นผิดพลาดได้เช่นกัน

เนื่องจากการวิจัยในครั้งนี้มีจุดมุ่งหมายในการคัดแยกข้อความที่มีความหมายและทำหน้าที่ในประโยคได้อย่างถูกต้องตามที่กำหนดมากกว่าจะมุ่งเน้นไปที่จำนวนประโยคที่สามารถค้นคืนมาได้ ดังนั้นจึงให้ความสำคัญต่อค่าพริชชันมากกว่าค่ารีคอล การที่ค่ารีคอลมีค่าต่ำและค่าพริชชันมีค่าสูงนั้นแสดงให้เห็นว่า ถึงแม้ความสามารถในการค้นคืนประโยคที่เกี่ยวข้องกับคำสำคัญที่กำหนดนั้นจะได้ประโยคที่ต้องการมาเป็นจำนวนไม่มากนัก เช่น พบประโยคที่มีคำสำคัญ “produced” แต่ไม่พบประโยคที่มีคำสำคัญ “produces” เนื่องจากคำสำคัญดังกล่าวไม่ได้ผ่านกระบวนการจัดการกับคำมาก่อน แต่จากประโยคทั้งหมดที่ค้นคืนมาได้จากคำสำคัญ “produced” นั้นส่วนใหญ่จะสามารถทำการคัดแยกข้อความที่ต้องการได้อย่างถูกต้องและมีความหมาย

5.2 บทสรุป

งานวิจัยนี้ได้นำเสนอการคัดแยกข้อความที่ต้องการจากประโยค โดยใช้หลักการของลิงค์แกรมมาพาสเซอร์ในการวิเคราะห์โครงสร้างของประโยคก่อนที่จะทำการคัดแยกข้อความที่ต้องการเพื่อให้ได้ข้อมูลที่มีความหมายมากขึ้น ซึ่งการวิเคราะห์โครงสร้างของประโยคด้วยลิงค์แกรมมาพาสเซอร์นั้นจะต้องใช้เวลาพอสมควรในการวิเคราะห์โครงสร้างของแต่ละประโยคและในบางประโยคพาสเซอร์จะไม่สามารถทำการวิเคราะห์ได้ด้วยเหตุผลดังที่ได้กล่าวในบทที่ 3 ดังนั้นในงานวิจัยนี้จึงให้ความสำคัญกับความถูกต้องของข้อมูลที่ได้จากการคัดแยกมากกว่าความรวดเร็วในกระบวนการการคัดแยก ดังนั้นเมื่อประโยคถูกวิเคราะห์โครงสร้างด้วยลิงค์แกรมมาพาสเซอร์แล้วจึงนำโครงสร้างของประโยคที่ได้จากการวิเคราะห์ดังกล่าวมาจัดให้อยู่ในโครงสร้างต้นไม้ที่เรียกว่าต้นไม้ไวยากรณ์ เพื่อใช้ในการคัดแยกข้อความที่ต้องการจากโครงสร้างต้นไม้ นั้น โดยกำหนดว่าข้อมูลที่คัดแยกมาได้นั้นจะต้องเป็นกลุ่มคำที่เป็นนามวลีที่มีความสัมพันธ์กับคำสำคัญที่กำหนดและคำสำคัญนั้นจะต้องเป็นคำที่อยู่ในกลุ่มของกิริยาวิลี งานวิจัยนี้ได้ทำการทดสอบกับบทความที่เป็นบทความย่อภาษาอังกฤษ ซึ่งมีเนื้อหาและการเขียนที่มีรายละเอียดแตกต่างกันไป 2 ประเภทด้วยกัน คือ บทความย่อที่มีเนื้อหาในสาขาสังคมวิทยาและบทความย่อในสาขาไบโอเคมีซึ่งมีศัพท์เฉพาะมากกว่า พบว่าสามารถนำหลักการนี้มาใช้ในการคัดแยกข้อความที่ต้องการได้และได้ผลลัพธ์เป็นข้อความที่มีความหมายสมบูรณ์มากขึ้น

จากผลของคำพรีซิชั่นที่ได้จากการประเมินนั้น พบว่าการหลักการของลิงค์แกรมมาพาสเซอร์ที่ได้นำมาใช้ในการคัดแยกข้อความในครั้งนี้มีความเหมาะสมสำหรับบทความที่มีลักษณะในการเขียนที่มีความถูกต้องตรงตามหลักไวยากรณ์และไม่มีการใช้คำศัพท์ที่เป็นคำเฉพาะพิเศษที่ไม่ปรากฏในพจนานุกรมทั่วไป เช่น สุตรงทางเคมี มากกว่าการนำไปใช้กับบทความที่มีคำศัพท์เฉพาะหรือใช้เครื่องหมายต่าง ๆ จำนวนมาก เช่น บทความทางวิทยาศาสตร์ ซึ่งลักษณะดังกล่าวจะมีผลต่อความถูกต้องในการวิเคราะห์ประโยค เมื่อการวิเคราะห์ประโยคผิดพลาดจึงมีผลทำให้การคัดแยกมีความผิดพลาดหรือไม่สมบูรณ์ได้เช่นกัน

สำหรับความสามารถในการค้นคืนที่แสดงโดยผลของคำรีคอลที่ประเมินได้นั้นจะสามารถปรับปรุงให้มีค่าที่ดีขึ้นได้ เนื่องจากคำสำคัญที่ใช้เป็นทริกเกอร์เวิร์ดในการคัดแยกสำหรับงานวิจัยในครั้งนี้ยังไม่ได้มีการผ่านกระบวนการในการจัดการคำก่อนที่จะนำคำดังกล่าวมาใช้ในการสืบค้นเพื่อทำการคัดแยกข้อความที่ต้องการ เช่น กระบวนการในการจัดกลุ่มคำที่มีความหมายเหมือนกันหรือกระบวนการในการทำให้คำสำคัญนั้นเป็นรากศัพท์ (Stemming) ก่อนที่จะทำการสืบค้นใน

ประโยค ซึ่งกระบวนการดังกล่าวจะทำให้ความสามารถในการค้นคว้า ประโยคที่เกี่ยวข้องกับคำ
สำคัญนั้นมีประสิทธิภาพเพิ่มขึ้นได้



ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright © by Chiang Mai University
All rights reserved