การพัฒนาต้นแบบการให้คะแนนข้อสอบอัตนัยชนิดภาษาไทยด้วยระบบอัตโนมัติ กรณีศึกษา วิชาสารสนเทศในชีวิตประจำวัน มหาวิทยาลัยราชภัฏเชียงราย

เศรษฐชัย ใจฮึก[1*], สุรศักดิ์ มังสิงห์[2]
เทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม ถนนพหลโยธิน แขวงเสนานิคม เขตจตุจักร กรุงเทพมหานคร 10900
อีเมล์ Seatachai@gmail.com[1*], smungsing@gmail.com[2]

บทนำ

ข้อสอบอัตนัยเป็นแบบทดสอบที่ผู้สอบแสดงคำตอบด้วยการเขียนบรรยาย เป็นลักษณะการสอบที่เดาคำตอบได้ยาก เหมาะสำหรับการวัดความสามารถด้านวัดเจตคติ และส่งเสริมความคิดริเริ่มสร้างสรรค์ แต่ข้อสอบอัตนัยมีข้อจำกัดในด้านการตรวจให้คะแนน เนื่องจากผู้ตรวจข้อสอบต้องมีความรู้ความเชี่ยวชาญในเรื่องที่ทำการตรวจและมีความอดทนในการตรวจคำตอบเพื่อประเมินระดับคะแนนของผู้สอบจำนวนมาก งานวิจัยผ่านมาได้มีการพัฒนานำซอฟแวร์เพื่ออำนวยความสะดวกให้กับมนุษย์ด้วยเทคนิคต่างๆ เช่น เทคนิคการเปรียบเทียบคำสำคัญของเอกสารเฉลยและคำตอบของผู้สอบ เทคนิคการสืบค้นและเปรียบเทียบจำนวนความถี่ของคำสำคัญ และเทคนิคการสร้างโมเดลเพื่อจำแนกคำตอบที่อาจจะเกิดขึ้น ดังนั้นผู้วิจัยจึงเล็งเห็นประโยชน์ของการพัฒนาระบบข้อสอบอัตนัยออนไลน์และระบบการตรวจข้อสอบอัตนัยแบบอัตโนมัติเพื่อนำมาใช้พัฒนาการเรียนการสอนของมหาวิทยาลัยให้ดียิ่งขึ้น จึงมีแนวคิดที่จะศึกษาวิจัยเพื่อสร้างต้นแบบการตรวจข้อสอบอัตนัยชนิดภาษาไทยอัตโนมัติของมหาวิทยาลัยราชภัฏเชียงรายด้วยเทคนิคเปรียบเทียบความคล้ายกันระหว่างคำตอบเฉลยและคำตอบจากผู้สอบผ่านระบบออนไลน์และเทคนิคพิจารณาความถี่สะสมย้อนกลับของคำสำคัญที่เกิดขึ้น สำหรับการทดลองได้จัดทำระบบคลังข้อสอบอัตนัยออนไลน์วิชาสารสนเทศในชีวิตประจำวัน ผลการทดลองพบว่า ระบบต้นแบบของการตรวจข้อสอบอัตนัยภาษาไทยที่พัฒนาขึ้น เมื่อทดลองใช้กับรายวิชาสารสนเทศในชีวิตประจำวัน โดยทดสอบจากนักศึกษาที่ลงทะเบียนเรียนในรายวิชาสารสนเทศในชีวิตประจำวัน ปีการศึกษา 3/2559 จำนวน 300 คน ด้วยข้อสอบจำนวน 30 ข้อ ระดับคะแนนผลการตรวจด้วยระบบให้ระดับคะแนนแตกต่างจากผู้เชี่ยวชาญถึงร้อยละ XX และได้รับความพึงพอใจจากผู้ใช้งานคือผู้ตรวจอยู่ในระดับ และความพึงพอใจจากผู้สอบอยู่ในระดับดี (X=0.00, S.D.=0.00)

คำสำคัญ ข้อสอบ, อัตนัย, ตรวจสอบ, คำตอบ, ระบบอัตโนมัติ
Examination, Subjective, Checking, Answer, Automation

บทนำ

มหาวิทยาลัยราชภัฏเชียงรายเป็นสถาบันอุดมศึกษาเพื่อการพัฒนาท้องถิ่น มีภารกิจหลักที่ในการผลิตบัณฑิตและพัฒนาคนที่มีคุณภาพโดยให้โอกาสประชาชนทุกระดับได้มีโอกาสทางการศึกษา มหาวิทยาลัยราชภัฏเชียงรายได้มุ่งเน้นการนำเทคโนโลยีเข้ามามีส่วนร่วมในการพัฒนาศักยภาพการศึกษาให้เข้าถึงและครอบคลุมประชาชนทุกระดับ มีการจัดการเรียนการสอนและการวัดผลการเรียนอย่างเป็นระบบ มีการประเมินผลการเรียนด้วยการจัดสอบทั้งข้อสอบอัตนัย ปรนัย และการปฏิบัติงาน

มหาวิทยาลัยราชภัฏเชียงรายได้เลือกข้อสอบอัตนัย (Subjective Exams) เป็นเครื่องมือหนึ่งที่ใช้ในการวัดและประเมินผลการเรียน เพราะข้อสอบอัตนัยมีจุดเด่นที่ผู้สอบคาดเดาคำตอบได้ยาก โดยผู้สอบจะต้องตอบคำถามด้วยวิธีการเขียนบรรยาย เพราะไม่มีตัวเลือกคำตอบเหมือนเช่นข้อสอบปรนัย ดังนั้นข้อสอบอัตนัยจึงเหมาะสมแก่การนำมาใช้วัดรู้ด้านเจตคติและวัดความคิดริเริ่มสร้างสรรค์ของผู้สอบ แต่การเลือกใช้ข้อสอบอัตนัยมีข้อจำกัดด้านการตรวจประเมินผลระดับคะแนน เนื่องจากผู้ตรวจข้อสอบที่มีความรู้ความเชี่ยวชาญในเรื่องที่ทำการตรวจมีจำนวนจำกัด ผู้ตรวจข้อสอบต้องมีความอดทนในการพิจารณาประเมินระดับคะแนนความรู้จากคำตอบของผู้สอบจำนวนมาก นอกจากนั้นอารมณ์และความรู้สึกของผู้ตรวจยังส่งผลต่อระดับคะแนนของผู้สอบแต่ละคนได้เช่นกัน ซึ่งปัจจุบันมหาวิทยาลัยราชภัฏเชียงรายยังไม่มีซอฟแวร์อำนวยความสะดวกในการตรวจข้อสอบอัตนัยแบบอัตโนมัติและยังคงใช้วิธีการตรวจด้วยมนุษย์

เมื่อศึกษาถึงงานวิจัยการตรวจข้อสอบอัตนัยที่ผ่านมา พบว่าได้มีการนำระบบคอมพิวเตอร์และซอฟแวร์มาใช้เป็นเครื่องมือในการลดข้อจำกัดของการตรวจข้อสอบอัตนัย นักวิจัยได้นำเสนอรูปแบบการตรวจข้อสอบอัตนัยด้วยวิธีการต่างๆ เช่น Anette Hulth (2000) ได้ใช้เทคนิคการเปรียบเทียบคำสำคัญ (Keyword) ของเอกสารเฉลยและคำตอบของผู้สอบ ถัดมา Jill Burstein, Claudia Leacock และ Richard Swartz (2001) ด้วยเทคนิค E-Rater ใช้เทคนิคการสืบค้นและเปรียบเทียบจำนวนความถี่ของคำสำคัญ (Text-Based/Keyword Based Information Retrieval) ถัดมา Lawrence M, Rudner และ Tahung Liang (2002) ได้เลือกใช้อัลกอริทึมแบบ K-NN (K-Nearest Neighbor Algorithm) ซึ่งเป็นการสร้างโมเดลเพื่อจำแนกคำตอบ (Classification) ที่อาจจะเกิดขึ้นในการเทียบคำสำคัญของของของเอกสารเฉลยและคำตอบของผู้สอบ

เมื่อนำเทคนิควิธีการตรวจข้อสอบอัตนัยแบบอัตโนมัติมาประยุกต์ใช้ให้เหมาะกับคำถามคำตอบภาษาไทยจำเป็นต้องมีการคัดแยกคำออกจากประโยค (Word Segmentation) เข้ามาช่วย เนื่องจากคุณลักษณะเฉพาะของภาษาไทยที่เขียนต่อกันเป็นประโยคยาว การเชื่อมต่อคำไม่เว้นวรรค และไวยากรย์สามารถสลับตำแหน่งได้ ปัจจุบันวิธีคัดแยกคำออกจากประโยคภาษาไทยที่ได้รับความการยอมรับอยู่สามวิธีคือ การใช้กฎการสร้างพยางค์ภาษาไทย (Rule Base Approach) การใช้พจนานุกรมคำศัพท์ (Dictionary Approach) และใช้คลังข้อความ (Corpus Approach) เช่น งานวิจัยของ Payothorn Urathamakun และ Kanda Runapongsa (2005) โดยการพิจารณาจากกฎและทำการปรับปรุงกฎที่ยังไม่ครอบคลุม เพื่อนำมาใช้พิจารณาคำที่เกิดขึ้นใหม่ ถัดมา Kessaraporn Suesatpanit และ Atiwong Suchato (2008) เลือกใช้วิธีเทคนิคไตรแกรมโดยใช้คลังข้อความที่ถูก

แบ่งคำแล้ว BEST Corpus จากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) เพื่อหาความน่าจะเป็นของคำที่เกิดขึ้น ถัดมาห้องปฏิบัติการวิจัยเทคโนโลยีเสียง (SPT) ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (2009) ได้พัฒนา TLexPlus โดยใช้คลังข้อมูลของ BEST2009 จำนวน 9 ล้านคำร่วมกับวิธีการ Conditional Random Fields(CRFs) โดยใช้เทคนิคการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) ในการพิจารณาคัดแยกคำออกจากประโยคที่เป็นคำศัพท์ภาษาไทยที่เกิดขึ้นใหม่ คำในภาษาต่างประเทศ หรือคำแสลงได้

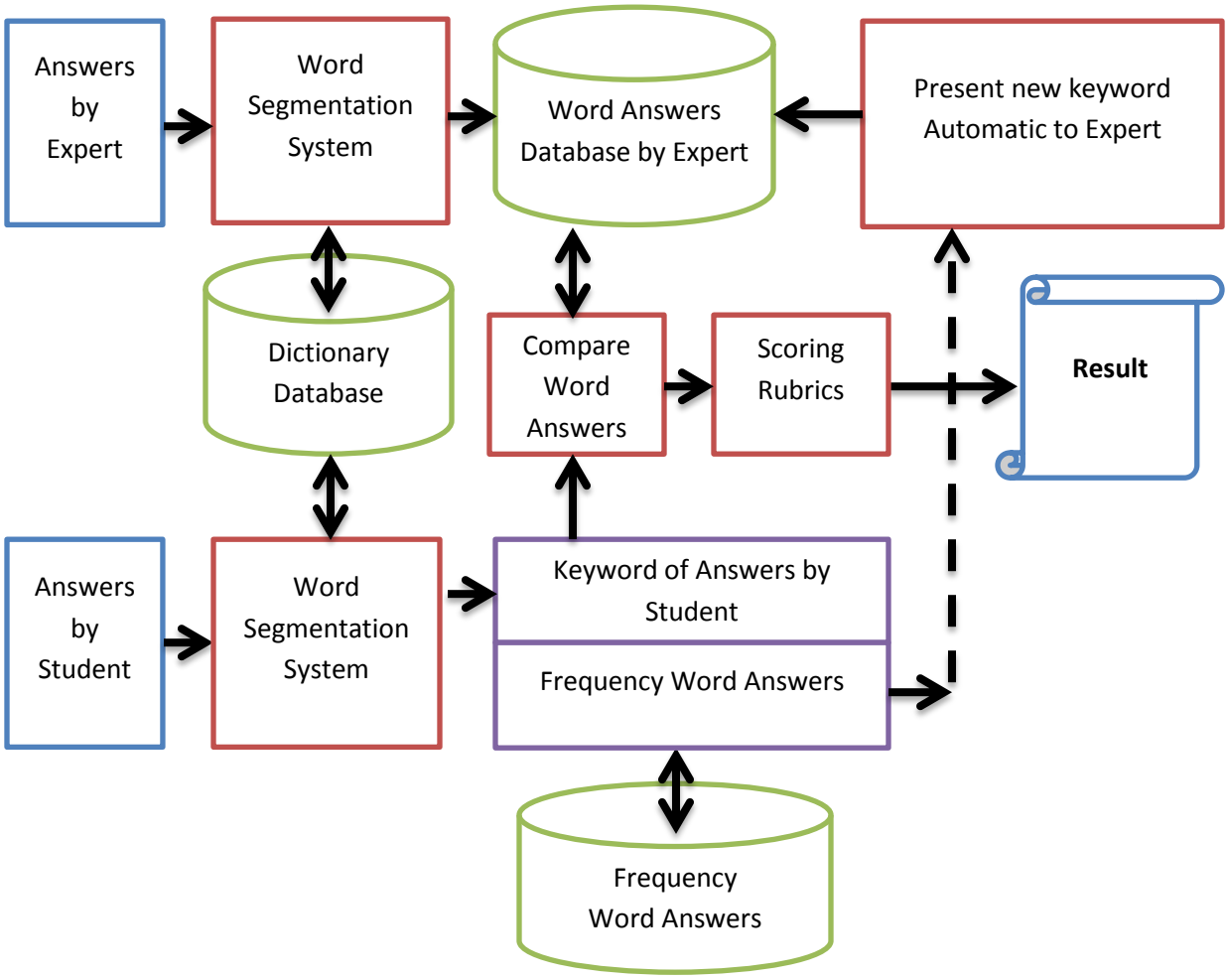
จากการสืบค้นผลวิจัยการตรวจข้อสอบอัตนัยชนิดภาษาไทยอัตโนมัติ เช่น ผลงานวิจัยของ Lawrence M และ Rudner and Tahung Liang (2002) ได้นำเสนอเทคนิคการเปรียบเทียบความเหมือนกันระหว่างคำตอบเฉลยและคำตอบจากผู้สอบด้วยวิธีการการพิจารณาค่าเวกเตอร์เส้นทางของคำตอบ ถัดมา Chanunya Loraksa และ Ratchata Peachavanish (2007) ได้นำเสนอเทคนิคการพิจารณาคะแนนของคำตอบจากผู้สอบด้วยโครงข่ายประสาทเทียมจากคลังข้อมูลคำตอบที่มีอยู่ โดยกำหนดค่าน้ำหนักของคำตอบในโครงข่ายประสาทเทียมในแต่ละนิวรอล (Node) เพื่อหาค่าผลต่างระหว่างตอบเฉลยกับคำตอบของผู้สอบและปรับค่าของโมเดลทุกรอบสำหรับใช้เป็นโมเดลพิจารณาคะแนนของคำตอบจากผู้สอบในกลุ่มข้อมูลถัดไป ถัดมา Sammart Aungkaseraneekul และ Chuleerat Jaruskulchai (2010) ได้นำเสนอการแก้ปัญหาโดยใช้วิธี K-NN (K-Nearest Neighbor Algorithm) เพื่อจำแนกกลุ่มคำตอบสำหรับการนำมาใช้พิจารณาการตรวจประเมินคะแนนเป็นต้น

ดังนั้นผู้วิจัยจึงมีแนวคิดในการพัฒนาต้นแบบสำหรับการตรวจข้อสอบอัตนัยชนิดภาษาไทยอัตโนมัติของมหาวิทยาลัยราชภัฏเชียงราย ด้วยเทคนิคเปรียบเทียบความเหมือนกันระหว่างคำตอบเฉลยและคำตอบจากผู้สอบร่วมกับเทคนิคพิจารณาความถี่สะสมย้อนกลับของคำสำคัญที่เกิดขึ้น การทดสอบสมมุติฐานในครั้งนี้ได้จัดทำเว็บไซต์การวัดผลและประเมินความรู้ด้วยข้อสอบอัตนัยรายวิชา Gen1102 สารสนเทศในชีวิตประจำวัน ของมหาวิทยาลัยราชภัฏเชียงราย มีการจัดทำระบบคลังข้อมูลคำถามและคำตอบเพื่อใช้เก็บสถิติความถี่ของคำตอบที่ถูกต้อง และระดับความพึงพอใจในคำตอบจากผู้เชี่ยวชาญ ทั้งนี้ผู้วิจัยคาดหวังว่าระบบจะลดภาระการทำงานของบุคลากรสายการสอน และลดข้อจำกัดของการเลือกใช้ข้อสอบอัตนัยในการจัดการเรียนการสอนในปัจจุบัน นอกจากนั้นจะสามารถเกิดประโยชน์ในการจัดเก็บองค์ความรู้แทนผู้สอนในอนาคตได้

กระบวนการและวิธีการ

จากการศึกษางานวิจัยด้านการตรวจข้อสอบอัตนัยชนิดภาษาไทยอัตโนมัติ ทำให้ผู้วิจัยได้ออกแบบเทคนิค ชื่อว่า Grading exams Subjective Thai automatic Model V 1.0 (GSTM-V1) ซึ่งจะเป็นต้นแบบเริ่มต้นสำหรับการต่อยอดในอนาคต

Grading exams Subjective Thai automatic Model V 1.0 (GSTM-V1)



รูปที่ 1 แนวคิดโครงสร้างการทำงานของการตรวจข้อสอบอัตนัยชนิดภาษาไทยอัตโนมัติ

1. Dictionary Database คือ ฐานข้อมูลคำศัพท์อิเล็กทรนิกส์ ที่อ้างอิงข้อมูลจากพจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ.2554
2. Word Answers Database by Expert คือ ฐานข้อมูลที่เก็บคำตอบและคำศัพท์ของผู้เชี่ยวชาญ
3. Frequency Word Answers คือ ฐานข้อมูลที่เก็บคำตอบและคำศัพท์ของผู้สอบ
4. Answers by Expert คือ โมดูลการรับคำตอบเฉลยด้วยภาษาไทยจากผู้เชี่ยวชาญ จำนวนไม่เกิน 200 อักขระ และทำสถิติความถี่คำศัพท์ โดยนำเก็บไว้ในฐานข้อมูลคำตอบเฉลยของแต่ละข้อของผู้เชี่ยวชาญ

(Word Answers Database by Expert) ที่จะนำมาใช้เป็นคำตอบหลักในการเปรียบเทียบความถูกต้องของคำตอบจากนักศึกษาหรือผู้สอบต่อไป

5. Answers by Student คือ โมดูลการรับคำตอบจากผู้นักศึกษาหรือผู้ทำการสอบ โดยมีข้อจำกัดของคำตอบที่พิมพ์ลงไป เป็นภาษาไทยที่มีจำนวนไม่เกิน 200 อักขระ
6. Word Segmentation System คือ โมดูลการคัดแยกคำออกจากประโยคโดยวิธีการเทียบคำจากพจนานุกรมคำศัพท์ (Dictionary Database) ด้วยกระบวนการตรวจสอบคำแบบย้อนกลับประโยค และจะพิจารณาเปรียบเทียบคำศัพท์กลุ่มคอมพิวเตอร์ก่อนเพื่อลดข้อจำกัดด้านคำพ้องรูป
7. Compare Word Answers คือ โมดูลการเปรียบเทียบคำศัพท์ (Keyword) คำตอบของนักศึกษาและคำตอบจากผู้เชี่ยวชาญจากกระบวนการตารางความจริง (True Table) โดยจะพิจารณาเฉพาะความสถิติถี่ของคำตอบเฉลยที่ถูกต้องที่สุดเพียง 20% แรกของคำตอบจากผู้เชี่ยวชาญ (สามาถปรับระดับความการพิจารณาได้)
8. Scoring Rubrics คือ โมดูลการพิจารณาระดับของคะแนนด้วยวิธีเกณฑ์ การประเมินตามคุณภาพของงานหรือคำตอบในภาพรวม (Holistic Rubric) ดังตาราง

ตารางการเปรียบเทียบผลคะแนน		
ระดับคะแนน	**คำตอบที่พบ**	**ความหมาย**
3	ร้อยละ 91 - 100	คำตอบมีความถูกต้องระดับมากและตอบครบถ้วน มีแนวโน้มว่าเป็นตอบที่ถูกต้อง
2	ร้อยละ 71 - 90	คำตอบมีความถูกต้องระดับปานกลาง ตอบไม่ครบถ้วน มีแนวโน้มว่าเป็นตอบที่ถูก
1	ร้อยละ 61 - 70	คำตอบมีความถูกต้องระดับน้อย ตอบไม่ครบถ้วน ไม่สมบูรณ์ มีแนวโน้มว่าอาจจะเป็นตอบที่ถูก
0	น้อยกว่าร้อยละ 60	สอบไม่ได้เลย

* ร้อยละของการตรวจพบคำตอบที่เหมือนกัน สามารถปรับแต่งได้ตามแต่ละข้อจากการพิจารณาจากผู้เชี่ยวชาญ และผู้ดูและระบบ

9. Frequency Word Answers คือ โมดูลการพิจารณาความถี่คำศัพท์ของผู้ตอบคำถามเปรียบเทียบกับคำตอบที่มีอยู่ในฐานข้อมูล Word Answers Database by Expert หากพบปริมาณความถี่ที่มากกว่าคำศัพท์ของผู้เชี่ยวชาญ โมดูลส่งต่อการทำงานไปยังโมดูล Present new keyword Automatic to Expert
10. Present new keyword Automatic to Expert คือ โมดูลการจัดเก็บคำตอบ และคำศัพท์ของผู้สอบได้จากโมดูล Word Segmentation System ทำงานประสานกับโมดูล Frequency Word Answers หาก

พบปริมาณความถี่ที่มากกว่าคำศัพท์ของผู้เชียวชาญ โมดูลจะทำการแจ้งเตือนให้แก่ผู้ดูแลระบบในการพิจารณาเพิ่มคำตอบในฐานข้อมูล Word Answers Database by Expert ต่อไป

การทดลอง

การทดลองเมื่อได้จัดทำระบบคลังข้อสอบอัตนัยออนไลน์วิชาสารสนเทศในชีวิตประจำวันเสร็จสิ้น และนำมาทดสอบนักศึกษาที่ลงทะเบียนเรียนในรายวิชาสารสนเทศในชีวิตประจำวัน ปีการศึกษา 3/2559 จำนวน 300 คน ด้วยข้อสอบจำนวน 10 ข้อ และทำการสุ่มการตอบของนักศึกษาโดยวิธีการแบบไม่เจาะจง จำนวน 169 คน (Krejcie and Morgan, 1970) นำข้อมูลมาใช้ในการเปรียบเทียบ ดังนี้ (1) การทดสอบวัดประสิทธิภาพด้วยค่าเอฟเมเชอร์ซึ่งทำการเปรียบเทียบระหว่างการตรวจด้วยผู้เชี่ยวชาญ และตรวจด้วยระบบเมื่อมีการ Feedback คำตอบ (2) การเปรียบเทียบระหว่างการตรวจให้คะแนนจากมนุษย์และการตรวจจากระบบ และ (3) การประเมินผลความพึงพอใจจากผู้ใช้งาน (ประเมินตามรูปแบบสารสนเทศ)

ตารางที่ 1 ทำการทดสอบโดยการสุ่มคำถามจำนวน 10 ข้อ จากทั้งสิ้น 30 ข้อ

ข้อที่	การตรวจด้วยการพิจารณาคำศัพท์ในคำตอบ(Keyword)		
	การตรวจด้วย ผู้เชี่ยวชาญ (มนุษย์)	การตรวจด้วยระบบ รอบที่ 1	การตรวจด้วยระบบรอบที่ 2 เมื่อมีการ Feedback คำตอบที่ได้จากนักศึกษา โดยผ่านการพิจารณาจากผู้เชี่ยวชาญ
1	12	7	8
2	5	3	3
3	8	5	5
4	5	4	4
5	4	4	4
6	5	2	3
7	9	5	5
8	6	2	4
9	4	4	4
10	6	3	5
รวม	**64**	**39**	**45**

ผลการทดสอบวัดประสิทธิภาพด้วยค่าเอฟเมเชอร์ซึ่งทำการเปรียบเทียบระหว่างการตรวจด้วยผู้เชี่ยวชาญ (มนุษย์) ที่ตั้งสมมุติฐานว่ามีความถูกต้องมากกว่าการตรวจด้วยระบบรอบที่ 2 เมื่อมีการ Feedback คำตอบที่ได้จากนักศึกษาโดยผ่านการพิจารณาจากผู้เชี่ยวชาญ ซึ่งค่าเอฟเมเชอร์เป็นการวัดประสิทธิภาพพื้นฐานในการจัดกลุ่มโดยคำนวณได้จากสมการดังนี้

$$F - Measure = \frac{2RP}{R+P} \quad (1)$$

$$P = \frac{A}{A+B}$$

$$R = \frac{A}{A+C}$$

เมื่อ P คือ ค่าความถูกต้อง (Precision)

R คือ ค่าความครบถ้วน (Recall)

A คือ จำนวนเอกสารที่สามารถเลือกได้ถูกต้อง

B คือ จำนวนเอกสารที่เลือกมาไม่ถูกต้อง

C คือ จำนวนเอกสารที่ถูกต้องแต่ไม่ถูกเลือก

จากการวิจัยนี้พบว่า ค่า Precision เท่ากับ 7.35%, ค่า Recall เท่ากับ 4.19% และค่า F-measure เท่ากับ 5.33%

ตารางที่ 2 การเปรียบเทียบระหว่างการตรวจให้คะแนนจากมนุษย์และการตรวจจากระบบ ด้วยตัวอย่างการสุ่มจากนักศึกษาจำนวน 169 คน จำนวนตัวอย่าง 5 ข้อ

คนที่	ข้อที่ 1		ข้อที่ 2		ข้อที่ 3		ข้อที่ 4		ข้อที่ 5	
คนที่	มนุษย์	ระบบ	มนุษย์	ระบบ	มนุษย์	ระบบ	มนุษย์	ระบบ	มนุษย์	ระบบ
1	3	3	3	3	2	2	1	0	2	2
2	3	2	3	3	2	1	2	2	3	1
3	2	3	3	2	3	2	2	1	3	2
4	2	0	2	1	1	0	3	1	2	1
5	3	2	0	0	0	0	3	1	2	2
...										
169	3	3	1	1	0	0	2	1	2	2

ตารางที่ 3 การประเมินผลความพึงพอใจจากผู้ใช้งาน (ประเมินตามรูปแบบสารสนเทศ) โดยให้คะแนน 5 ระดับ เป็นแบบสอบถามเกี่ยวกับการประเมินผลความพึงพอใจจากผู้ใช้งาน มีลักษณะเป็นแบบมาตราส่วนประมาณค่า (Rating Scale) 5 ระดับ วิเคราะห์ข้อมูลโดยใช้วิธีหาค่าเฉลี่ย(Average) และค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) โดยกำหนดเกณฑ์การแปลความหมายเพื่อจัดระดับค่าเฉลี่ยออกเป็นช่วงดังต่อไปนี้

ค่าเฉลี่ย 4.50 – 5.00 หมายความว่า พึงพอใจมากที่สุด

ค่าเฉลี่ย 3.50 – 4.49 หมายความว่า พึงพอใจมาก

ค่าเฉลี่ย 2.50 – 3.49 หมายความว่า พึงพอใจปานกลาง

ค่าเฉลี่ย 1.50 – 2.49 หมายความว่า พึงพอใจน้อย

ค่าเฉลี่ย 1.00 – 1.49 หมายความว่า พึงพอใจน้อยที่สุด

ด้าน/ผู้เชี่ยวชาญคนที่	1	2	3	4	5	ผลรวม	X	S.D	พึงพอใจ
1. ด้านกระบวนการ/ขั้นตอนการใช้งานระบบ									
1.1. รูปแบบการใช้งานระบบ ความยาก - ง่าย	0	0	0	2	3	5	4.60	0.49	มากที่สุด
1.2. กระบวนการทำงานของระบบ	0	0	1	3	1	5	4.00	0.63	มากที่สุด
ด้าน/ผู้เชี่ยวชาญคนที่	**1**	**2**	**3**	**4**	**5**	**ผลรวม**	**X**	**S.D**	พึงพอใจ
2. ด้านประสิทธิภาพของระบบ									
2.1. ความถูกต้อง แม่นยำของระบบ	0	0	2	2	1	5	3.80	0.75	มากที่สุด
2.2. ตรงตามวัตถุประสงค์ที่ต้องการ	0	0	0	2	3	5	4.60	0.49	มากที่สุด
2.3. การออกแบบให้ใช้งานง่าย เมนูไม่ซับซ้อน	0	0	1	2	2	5	4.20	0.75	มากที่สุด
2.4. ความเป็นปัจจุบันของข้อมูล	0	0	0	1	4	5	4.80	0.40	มาก
ด้าน/ผู้เชี่ยวชาญคนที่	**1**	**2**	**3**	**4**	**5**	**ผลรวม**	**X**	**S.D**	พึงพอใจ
3. ด้านความสะดวก สวยงาม									
3.1. ความสะดวกในการใช้งานโปรแกรม	0	0	0	1	4	5	4.80	0.40	มาก
3.2. ความเหมาะสมในการใช้งานโปรแกรม	0	0	0	2	3	5	4.60	0.49	มาก
ด้าน/ผู้เชี่ยวชาญคนที่	**1**	**2**	**3**	**4**	**5**	**ผลรวม**	**X**	**S.D**	พึงพอใจ
4. ด้านคุณภาพของระบบ									
4.1. ความพึงพอใจในการใช้งาน	0	0	0	3	2	5	4.40	0.49	มาก
4.2. ความสามารถของระบบ ในการนำไปใช้ประโยชน์	0	0	0	1	4	5	4.80	0.40	มาก

ผลการทดลอง

ผลเปรียบเทียบผลพบว่า (1) การทดสอบประสิทธิภาพของการค้นหาคำตอบเปรียบเทียบระหว่างการตรวจด้วยผู้เชียวชาญและระบบโดยการสุ่มคำถามจำนวน 10 ข้อ จากทั้งสิ้น 30 ข้อ ในการทดสอบกับนักศึกษาในเทอมที่ 3/2559 พบว่าการตรวจด้วยผู้เชี่ยวชาญ(มนุษย์)ที่ตั้งสมมุติฐานว่ามีความถูกต้องมากกว่าการตรวจด้วยระบบพบว่าการตรวจด้วยระบบในครั้งแรกให้ค่าการตรวจที่ใกล้เคียงกับมนุษย์ถึง 64.93% เมื่อทำการการตรวจด้วยระบบรอบที่ 2 เมื่อมีการ Feedback คำตอบที่ได้จากนักศึกษาโดยผ่านการพิจารณาจากผู้เชี่ยวชาญ ทำให้มีความถูกต้องใกล้เคียงกับการตรวจด้วยมนุษย์ถึง 70.31% ซึ่งมากกว่าครั้งที่ 1 ถึง 5.38% มีค่า Precision เท่ากับ 7.35%, ค่า Recall เท่ากับ 4.19% และค่า F-measure เท่ากับ 5.33% (2) การเปรียบเทียบระหว่างการตรวจให้คะแนนจาก

มนุษย์และการตรวจจากระบบ ด้วยตัวอย่างการสุ่มจากนักศึกษาจำนวน 169 คน จำนวนตัวอย่าง 5 ข้อ พบว่า การตรวจของระบบมีคะแนนที่ใกล้เคียงกับการตรวจของผู้เชี่ยวชาญ (3) การประเมินผลความพึงพอใจจากผู้ใช้งาน (ประเมินตามรูปแบบสารสนเทศ) พบว่า 1.ด้านกระบวนการ/ขั้นตอนการใช้งานระบบของรูปแบบการใช้งานระบบความยาก – ง่ายมีความพึงพอใจในระดับมากที่สุด(X=0.49) และกระบวนการทำงานของระบบมีความพึงพอใจในระดับมากที่สุด(X=0.63) 2. ด้านประสิทธิภาพของระบบของความถูกต้อง แม่นยำของระบบมีความพึงพอใจในระดับมากที่สุด(X=0.75), ตรงตามวัตถุประสงค์ที่ต้องการมีความพึงพอใจในระดับมากที่สุด(X=0.49), การออกแบบให้ใช้งานง่าย เมนูไม่ซับซ้อนมีความพึงพอใจในระดับมากที่สุด(X=0.75)และความเป็นปัจจุบันของข้อมูลมีความพึงพอใจในระดับมาก(X=0.40) 3.ด้านความสะดวก สวยงาม ของความสะดวกในการใช้งานโปรแกรมมีความพึงพอใจในระดับมาก (X=0.40)และความเหมาะสมในการใช้งานโปรแกรมมีความพึงพอใจในระดับมาก(X=0.49) และสุดท้าย 4.ด้านคุณภาพของระบบของความพึงพอใจในการใช้งานมีความพึงพอใจในระดับมาก(X=0.49) และความสามารถของระบบ ในการนำไปใช้ประโยชน์มีความพึงพอใจในระดับมาก(X=0.40)

การอภิปราย

จากการวิจัยทำให้ทราบถึงความสามารถของระบบในการประเมินคะแนนที่มีความใกล้เคียงกับการตรวจคำตอบด้วยมนุษย์ ถึงแม้ว่ากระบวนการทั้งหมดอาจจะไม่ได้ผลลัพธ์คล้ายกับการตรวจโดยมนุษย์โดยทั้งหมด แต่ได้เป็นจุดเริ่มต้นการสร้างต้นแบบสำหรับใช้งานระบบตรวจข้อสอบอัตนัยภาษาไทย ของมหาวิทยาลัยราชภัฏเชียงรายและเป็นเครื่องมือในการอำนวยความสะดวกแก่การจัดการเรียนการสอน ซึ่งผลการวิจัยนั้นสอดคล้องกับผลการวิจัยของ Anette Hulth (2000) ที่ได้ใช้เทคนิคการเปรียบเทียบคำสำคัญ (Keyword) ของเอกสารเฉลยและคำตอบของผู้สอบ ผลงานของ Jill Burstein, Claudia Leacock และ Richard Swartz (2001) ด้วยเทคนิค E-Rater ใช้เทคนิคการสืบค้นและเปรียบเทียบจำนวนความถี่ของคำสำคัญ (Text-Based/Keyword Based Information Retrieval) และผลงานวิจัยของ Lawrence M และ Rudner and Tahung Liang (2002) ที่ได้นำเสนอเทคนิคการเปรียบเทียบความเหมือนกันระหว่างคำตอบเฉลยและคำตอบจากผู้สอบด้วยวิธีการการพิจารณาค่าเวกเตอร์เส้นทางของคำตอบ ที่พบว่าเมื่อเขียนคำที่สะกดผิดทำให้ระบบไม่สามารถค้นหาและเปรียบเทียบคำที่ถูกต้องได้ นอกจากนั้นคำถามก็ยังมีผลต่อคะแนนของคำตอบ กล่าวคือเมื่อคำถามเป็นคำถามที่เปิดกว้าง และอิสระในการตอบก็ทำให้การประเมินผลคะแนนมีความต่างกัน เนื่องจากคำตอบอาจไม่ครอบคลุมในการตรวจสอบคำตอบได้ทั้งหมด ถึงแม้ว่าโมเดลการพัฒนาต้นแบบการให้คะแนนข้อสอบอัตนัยชนิดภาษาไทยด้วยระบบอัตโนมัติ กรณีศึกษา วิชาสารสนเทศในชีวิตประจำวัน มหาวิทยาลัยราชภัฏเชียงราย จะมีกระบวนการแก้ไขคำศัพท์ที่เกิดขึ้นใหม่ แต่ก็ยังไม่สามารถกระทำได้ทันที แต่ยังต้องอาศัยระยะเวลาในการเก็บข้อมูลการสอบหลายๆ ครั้งเพื่อให้คอมพิวเตอร์สร้างคำศัพท์ที่ครอบคลุมสำหรับการตรวจคำตอบ

บทสรุปและข้อเสนอแนะ

การพัฒนาต้นแบบการให้คะแนนข้อสอบอัตนัยชนิดภาษาไทยด้วยระบบอัตโนมัติ กรณีศึกษา วิชาสารสนเทศในชีวิตประจำวัน มหาวิทยาลัยราชภัฏเชียงราย ยังเป็นเพียงกระบวนการเบื้องต้นสำหรับใช้พัฒนาต่อยอด ซึ่งได้ทำการทดลองเฉพาะรายวิชาที่มีการเขียนตอบด้วยการบรรยายเชิงพรรณา ยังมีข้อจำกัดในการตอบคำถามด้านการคำนวณคณิตศาสตร์ เช่นสูตรสมการหรือค่ามากกว่าน้อยกว่า เป็นต้น นอกจากนั้นหากมีโมเดลที่สามารถวิเคราะห์คำศัพท์ที่ผู้ตอบเขียนสะกดผิดก็จะทำให้ระบบค้นหาคำและเปรียบเทียบให้ใกล้เคียงกับการตรวจโดยผู้เชี่ยวชาญได้มากขึ้น การพัฒนาต่อยอดให้ต้นแบบโมเดลมีความชาญฉลาดในการพิจารณาถึงกลุ่มคำตอบได้อย่างครอบคลุมก็เป็นสิ่งสำคัญที่ผู้ทำการวิจัยจะนำมาพัฒนาต่อ ผู้วิจัยจะนำระบบปัญญาประดิษฐ์เข้ามาผสานกับระบบสำหรับใช้ในการประเมินเปรียบเทียบคำตอบที่เขียนต่างกันแต่มีความหมายกัน เพื่อลดการซ้ำซ้อนของการประเมินผลการให้คะแนนให้มีความใกล้เคียงกับการตรวจด้วยผู้เชี่ยวชาญต่อไป

ประโยชน์ความรู้ที่ได้รับ

จากวิจัยการพัฒนาต้นแบบการให้คะแนนข้อสอบอัตนัยชนิดภาษาไทยด้วยระบบอัตโนมัติ กรณีศึกษา วิชาสารสนเทศในชีวิตประจำวัน มหาวิทยาลัยราชภัฏเชียงราย ทำให้ได้ประโยชน์หลัก 2 ด้าน คือ (1) ด้านการวิจัยสารสนเทศที่จะได้ต้นแบบสำหรับการพัฒนาต่อยอดการให้คะแนนข้อสอบอัตนัยชนิดภาษาไทยด้วยระบบอัตโนมัติของมหาวิทยาลัยราชภัฏเชียงรายต่อไป และ (2) ด้านนวัตกรรมที่ช่วยสนับสนุนทางการศึกษาที่อำนวยความสะดวกให้กับหน่วยจัดการทดสอบความรู้และทักษะในใช้ตรวจคำตอบข้อสอบอัตนัย ช่วยลดระยะเวลาการประเมินผลคะแนน ลดความผิดพลาดในการตรวจข้อสอบ และเพิ่มความแม่นยำของการให้คะแนน นอกจากนั้นผู้เรียนสามารถนำระบบ มาใช้งานเพื่อการวิเคราะห์ระดับความสามารถและศักยภาพของการเรียนรู้ด้วยตนเอง

อ้างอิง

[1] Chandrasekaran, B., Josephson, J.R. & Benjamins, V.R. (1999), What are ontologies And Why Do We Need Them, IEEE Intelligent System, 14(1), 20-26.

[2] Elbassuoni, S., Ramanath, M., Schenkel, R. & Weikum, G. (2010). Searching RDF Graphs with SPARQL and Keywords, IEEE.

[3] Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specification.Knowledge Acquisition, 5(2), 199-220.

[4] I. Murua, E. Llado and B. Llodra, (2006), “The semantic web for improving dynamic tourist packages commercialisation”, URL : http://www.ibit.org/dades/doc/1108_ca.pdf, access on 20/12/2009

[5] Hartig, O., Bizer, C. & Freytag, J. (2009). Executing SPARQL Queries over the Web of Linked Data, In International Semantic Web Conference, Vol. 5823, 293-309.

[6] Kolas, D. (2008). Supporting Spatial Semantics with SPARQL, Terra Cognita Work-shop.
[7] McGuinness, D.L. and Harmelen, F.V. (2004), Owl Web Ontology Language Overview, World Wide Web Consortium (W3C) Recommendation, URL : http://www.w3.org/ TR/owl-features, access on 25/11/2010.
[8] Quilitz, B. and Leser, U. (2008). Querying distributed RDF data sources with SPARQL, Proceedings of the 5th European Semantic Web Conference (ESWC), Volume 5021 of Lecture Notes in Computer Science, Springer Verlag, 524–538.
[9] Sbodio, M.L., Martin, D. & Moulin, C. (2010), Discovering Semantic Web services using SPARQL and intelligent agents, Web Semantics: Science, Services and Agents on the World Wide Web, Volume 8, Issue 4, November, 310-328.
[10] จุฑาวรรณ สิทธิโชคสถาพร. (2555). ต้นแบบออนโทโลยีสำหรับการสืบค้นสารสนเทศเชิงความหมาย สำหรับงานสารบัญอิเล็กทรนิกส์ กรณีศึกษางานบริหารและธุรการ คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์, มหาวิทยาลัยสงขลานครินทร์ : The Thesis of Songkla University.
[11] นฤพนธ์ พนาวงศ และ จักรกฤษณ์ เสน่ห์, (2553), ระบบค้นหาสถานที่ท่องเที่ยวในประเทศไทยด้วยหลักการออนโทโลยีและเนมแมทชิง, Journal of Information Science and Technology, Page 60-69.
[12] ราชกิจจานุเบกษา. (2553). กฎกระทรวงว่าด้วยระบบ หลักเกณฑ์ และวิธีการประกันคุณภาพการศึกษา พ.ศ. 2553. เล่ม 127 ตอนที่ 23 ก.
[13] สำนักงานคณะกรรมการการอุดมศึกษา. (2553). คู่มือการประกันคุณภาพการศึกษาภายในสถานศึกษาระดับอุดมศึกษา พ.ศ. 2553, กรุงเทพฯ: ภาพพิมพ์.