ระบบคำถาม-คำตอบประยุกต์ใช้กับข้อความภาษาไทย

Question-Answering System applied in Thai Language

เกศสุดา ตุริยากรณ์ [1] และ ชุลีรัตน์ จรัสกุลชัย [1]

Kadesuda Turiyagron [1] and Chuleerat Jaruskulchai [1]

บทคัดย่อ

งานทางด้านการค้นคืนสารสนเทศ (Information Retrieval) จะได้ผลลัพธ์ที่แสดงรายการของเอกสารทั้งหมดที่เกี่ยวข้องกับคำขอนั้นๆ แต่ในความเป็นจริงผู้ใช้อาจต้องการเพียงแค่ข้อมูลบางอย่างในเอกสาร โดยต้องการทราบว่า นายชวน หลีกภัย คือใคร? หรือ นครโอซากาอยู่ที่ประเทศใด? ซึ่งการที่จะค้นหาคำตอบสำหรับคำถามเหล่านี้จะใช้เทคนิคที่เรียกว่า ระบบคำถาม-คำตอบ (Question Answering)

ในงานวิจัยฉบับนี้นำเสนอระบบคำถาม-คำตอบโดยประยุกต์ใช้กับข้อความภาษาไทย และได้นำวิธีการจัดหมวดหมู่เอกสารแบบนาอีฟ เบส์ (Naive Bayes) จำแนกเอกสารให้สอดคล้องกับคำตอบแต่ละประเภท และใช้วิธีการหาอัตราเปรียบเทียบการเกิดร่วมกัน (Co-occurrence ratio) ในการเลือกคำตอบจากเอกสารที่จัดหมวดหมู่แล้ว ซึ่งผลลัพธ์ของการวิจัยได้ประสิทธิผลในการตอบคำถามที่วัดด้วยค่า MRAR คือ 0.518 ถือว่าเป็นประสิทธิผลที่น่าพอใจ

ABSTRACT

Traditionally, the results of various Information Retrieval techniques are list of documents relevant to a user's query. However, in real world problems, user may want only some data in documents, for example to know "who is Chuan Leekpai?", "Where is Osaka?" and etc. Thus, Question Answering Systems now have many important roles in finding answers that match to a user's question.

In this paper, we present a Question Answering System applying in Thai language. A text classification technique, i.e. naive Bayes classifier is used to classify documents into three answer categories. We then utilize Co-occurrence ratio to select answers and list them to user. The result of research provides satisfactory efficiency in answering questions, the MRAR value given 0.518.

คำนำ

ในปัจจุบันข้อมูลข่าวสารมีความจำเป็นอย่างยิ่งต่อการดำเนินชีวิตมากขึ้น ไม่ว่าจะเป็นในเชิงธุรกิจ, การศึกษา หรือในการดำรงชีวิตประจำวัน ดังนั้นความต้องการใช้ข้อมูลจึงเป็นสิ่งที่มีความสำคัญและเมื่อความต้องการข้อมูลมีมากขึ้น ความรวดเร็วในการค้นหาข้อมูลและความถูกต้องของข้อมูลจึงเป็นสิ่งที่ผู้ใช้ต้องการ

---

1. ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ Department of Computer Science, Faculty of Science, Kasetsart University.

ระบบเสิร์จเอนจิน (Search Engine System) ในปัจจุบันจะเป็นการค้นคืนรายการของเอกสารที่สัมพันธ์กับคำขอ (query) ของผู้ใช้ และผู้ใช้จะต้องค้นหาคำตอบที่ต้องการจากรายการของเอกสารเหล่านั้น ซึ่งจะต้องใช้เวลาในการอ่านเอกสารและหาคำตอบที่ต้องการออกมา และเมื่อความต้องการข้อมูลของผู้ใช้ในบางครั้งต้องการขอบเขตในการค้นหาข้อมูลที่สั้นๆ เพียงเพื่อให้ได้คำตอบที่ต้องการจากคำถามที่เกิดขึ้น และเนื่องจากคำถามของผู้ใช้ที่จะใช้ในการค้นหาข้อมูลจะอยู่ในรูปแบบของภาษาธรรมชาติ ดังนั้นเพื่อให้การค้นหาคำตอบให้กับคำถามของผู้ใช้เป็นคำตอบที่ตรงกับคำถาม จะต้องมีการกำหนดประเภทของคำถามให้ตรงกับคำตอบที่ต้องการค้นหา เพื่อให้ได้คำตอบที่ต้องการมากที่สุด

ระบบคำถาม-คำตอบ (Question-Answering System: QA) ได้รับความสนใจจากนักวิจัยในการสนองตอบต่อความต้องการในการค้นหาข้อมูลแบบตอบคำถาม และเพิ่มประสิทธิภาพของระบบเสิร์จเอนจิน ให้มีความถูกต้องมากยิ่งขึ้น โดยที่ระบบคำถาม-คำตอบยังเป็นหนึ่งองค์ประกอบในการเพิ่มประสิทธิภาพของการค้นคืนสารสนเทศด้วย อย่างไรก็ตามจากกระบวนการทำงานของระบบการค้นคืนสารสนเทศ เพียงอย่างเดียวจะไม่สามารถตอบคำถามของผู้ใช้ได้ ระบบคำถาม-คำตอบจึงได้มีการประยุกต์งานด้านอื่นๆเข้ามาในระบบด้วย เช่น ระบบการสกัดสารสนเทศ (Information Extraction) และระบบการประมวลผลภาษาธรรมชาติ (Natural Language Processing) เพื่อช่วยให้การทำงานของระบบมีประสิทธิภาพและมีความถูกต้องมากยิ่งขึ้น โดยระบบการสกัดสารสนเทศจะเข้ามาช่วยในการสกัดสารสนเทศที่ต้องการจากเอกสาร เช่น การสกัดคำนามที่เป็นบุคคล จะดูจากคำนามที่มีคำขึ้นต้นด้วย Mr., Mrs., ฯ และการสกัดคำนามที่เกี่ยวกับองค์กรหรือบริษัท จะดูจากคำนามที่มี Inc., Corp., Ltd, ฯ ต่อท้าย เป็นต้น (Bailey, 2001) การนำระบบการประมวลผลภาษาธรรมชาติ เข้ามาช่วยในการวิเคราะห์โครงสร้างไวยากรณ์ (Syntactic Analysis) และวิเคราะห์ความหมายของคำ (Semantic Analysis) เช่น คำที่มีความหมายคลุมเครือ ตัวอย่าง คำว่า “bank” อาจมีได้หลายความหมาย ถ้าเป็นคำนามจะมีความหมายเกี่ยวกับสถาบันการเงิน หรือ ถ้าเป็นคำกริยาจะมีความหมายเกี่ยวกับการฝากเงิน เป็นต้น ดังนั้นจะต้องจัดเตรียมการแก้ปัญหาในคำที่มีความหมายคลุมเครือโดยการดูจากคำภายในประโยค (local) หรือ คำภายนอกประโยค (global) อย่างไรก็ตามการทำงานกับระบบภาษาธรรมชาติจะมีอุปสรรคในการทำงาน คือ ต้องการเวลาในการประมวลผล และทรัพยากรในการคำนวณ ดังนั้นจึงใช้การสร้างตัวแทนของคำถามและคำตอบและทำการเข้าคู่กันแทนการวิเคราะห์คำ เพราะจะทำให้ลดเวลาในการคำนวณ และโดยส่วนใหญ่ในการใช้ระบบภาษาธรรมชาติเข้ามาใช้ในงานของระบบคำถาม-คำตอบจะใช้ในการทำนายว่าผู้ใช้หมายความว่าอะไรในประโยคคำถาม

งานวิจัยฉบับนี้นำเสนอการพัฒนาระบบคำถาม-คำตอบประยุกต์ใช้กับข้อความภาษาไทย โดยในการทำวิจัยจะเก็บรวบรวมเอกสารภาษาไทยและค้นคืนให้ได้คำตอบที่ต้องการ และเนื่องจากข้อความภาษาไทยมีลักษณะพิเศษทางโครงสร้างของภาษาที่แตกต่างจากภาษาอังกฤษ คือ การมีวรรณยุกต์และสระเข้ามามีส่วนสำคัญในคำ รวมทั้งการเว้นเครื่องหมายวรรคตอน และการที่ข้อความภาษาไทยเป็นภาษาที่ไม่มีรูปแบบของภาษาที่แน่นอน ดังนั้นจึงได้นำงานทางด้านความน่าจะเป็นเชิงสถิติเข้ามาประยุกต์ใช้ในระบบคำถาม-คำตอบ เช่น การจัดหมวดหมู่เอกสาร (Text Classification) เพื่อลดความยุ่งยากในการวิเคราะห์เชิงภาษาศาสตร์ และเพื่อความสะดวกในการจัดเอกสารให้ตรงกับประเภทของคำตอบที่ต้องการ ในการทดลองจะทำการทดลองกับคำถาม 3 ประเภท คือ (1) คำถามเกี่ยวกับบุคคล: Who Question (2) คำถามเกี่ยวกับสถานที่: Where Question (3) คำถามเกี่ยวกับเวลา: When Question จากนั้นจะนำเอกสารที่มีหมวดหมู่ตรงกับคำถามนำมาเลือกคำตอบ

(Answer Selection) โดยใช้อัตราเปรียบเทียบการเกิดร่วมกัน (Co-occurrence ratio) เพื่อให้ได้เอกสารที่มีคำตอบที่มีความถูกต้องและใกล้เคียงกับคำถามมากที่สุด

อุปกรณ์และวิธีการ

ในงานวิจัยนี้จะแบ่งขั้นตอนการทำงานออกเป็น 6 ขั้นตอน โดยจะกล่าวถึงวิธีการและทฤษฎีที่ใช้ในการทำวิจัย คือ

1. **การเตรียมข้อมูลของคำถามและเอกสาร** (Questions and Documents Pre-Processing)

เป็นการจัดเตรียมเอกสารเพื่อให้อยู่ในรูปแบบที่สามารถนำไปใช้ประโยชน์ได้ง่ายและสะดวก ดังนั้นเอกสารจะถูกประมวลผลก่อนที่จะถูกนำไปทำงานในขั้นตอนต่อไป โดยนำเอกสารทั้งหมดมาตัดคำโดยใช้อัลกอริทึม Longest Matching (Sornlertlamvanich V., 1993) และใช้โปรแกรมตัดคำ (Kruengkrai and Jaruskulchai, 2001) จากนั้นจัดเก็บข้อมูลให้อยู่ในรูปแบบเวกเตอร์ของคำ

การเตรียมข้อมูลของคำถาม ทำเช่นเดียวกับขั้นตอนการเตรียมเอกสาร เช่น การตัดคำ, และการสร้างเวกเตอร์ของคำในประโยคคำถาม

2. **การแบ่งประเภทคำถาม** (Question Classification)

การแบ่งประเภทของคำถามเพื่อเป็นแนวทางในการค้นหาคำตอบ โดยจะแบ่งให้สอดคล้องกับคำตอบที่ต้องการ (Kim et al., 2000) การแบ่งประเภทคำถามจะกำหนดตามการวิเคราะห์สารสนเทศในประโยคคำถาม ในงานวิจัยนี้จะทำการทดลองแบ่งประเภทคำถามเป็น 3 ประเภท ประกอบด้วย

2.1 คำถามเกี่ยวกับสถานที่ (Where Question) จะมีคำที่เกี่ยวกับสถานที่ปรากฏอยู่ในคำถาม เช่น อำเภอใด
จังหวัดใด, ประเทศใด, ที่ไหนและที่ใด เป็นต้น
ตัวอย่างคำถามเกี่ยวกับสถานที่ เช่น

- อำเภอฝางอยู่ที่จังหวัดใด?

2.2 คำถามเกี่ยวกับบุคคล (Who Question) จะมีคำที่เกี่ยวกับบุคคลปรากฏอยู่ในคำถาม เช่น ใคร เป็นต้น
ตัวอย่างคำถามเกี่ยวกับบุคคล เช่น

- นายชวน หลีกภัยคือใคร?

2.3 คำถามเกี่ยวกับเวลา (When Question) จะมีคำที่แสดงถึงเวลาปรากฏอยู่ในคำถาม เช่น วันที่เท่าใด, เดือนใด, ปีใด, เวลาใดและเมื่อไหร่ เป็นต้น
ตัวอย่างคำถามเกี่ยวกับเวลา เช่น

- แผนพัฒนาเศรษฐกิจและสังคมแห่งชาติ ฉบับที่ 8 เริ่มในปีใด?

3. **การค้นคืนสารสนเทศ** (Information Retrieval)

การค้นคืนสารสนเทศจะเป็นขั้นตอนในการค้นคืนเอกสารที่สัมพันธ์กับคำถาม โดยในงานวิจัยฉบับนี้จะใช้วิธีการทำงานของการค้นคืนสารสนเทศจากวิธีการที่เรียกว่า Sparse-Matrix (Goharian et al., 2000) โดยในวิธีการทำงานของระบบการค้นคืนสารสนเทศจะมีใช้ใน 2 ขั้นตอนของการทำวิจัย คือ ขั้นตอนการค้นคืนเอกสารที่สัมพันธ์กับคำถามโดยจะค้นคืนเอกสารทั้งฉบับ และใช้อีกครั้งหลังจากที่มีการแบ่งเอกสารให้เป็นประโยค (ในขั้นตอนที่ 4)

4. **การแบ่งเอกสารให้เป็นประโยค** (Separate Documents)

เอกสารที่ได้มาจากขั้นตอนการค้นคืนสารสนเทศจะถูกนำมาแบ่งเอกสารให้เป็นเอกสารย่อยๆ โดยการแบ่งเอกสารในงานวิจัยฉบับนี้จะตัดประโยคในเอกสารซึ่งจะเอาเฉพาะข้อความที่สำคัญโดยอ้างอิงจากรูปแบบของ SGML (Standard Generalized Markup Language) จากนั้นจะนับจำนวนคำในแต่ละเอกสารให้มีจำนวน 40 คำเพื่อให้ขอบเขตในการค้นหาคำตอบแคบลง และเอกสารย่อยๆที่ได้มาจะต้องนำมาทำการค้นคืนข้อมูลสารสนเทศอีกครั้งเพื่อหาความสัมพันธ์ระหว่างข้อความที่ได้ในขั้นต้นกับคำถาม

5. การจัดหมวดหมู่เอกสาร *(Text Classification)*

การจัดหมวดหมู่เอกสารจะแยกตามประเภทของคำตอบโดยแบ่งเป็น 3 หมวดหมู่ คือ หมวดหมู่บุคคล, หมวดหมู่สถานที่ และหมวดหมู่เวลา โดยการจัดหมวดหมู่เอกสารจะใช้วิธีการของ Naive Bayes (Michell, 2001)

ในขั้นตอนการเรียนรู้จากเอกสารตัวแบบของ Naive Bayes จะต้องมีการจัดเตรียมเอกสารสำหรับการเรียนรู้ โดยจะแบ่งเอกสารเป็น 3 ตัวแบบการเรียนรู้ตามการจัดหมวดหมู่เอกสารและแต่ละหมวดหมู่จะใช้เอกสารจำนวน 30 เอกสารในการเรียนรู้ โดยเอกสารที่จะนำมาทำการเรียนรู้จะต้องมีคำที่บ่งบอกถึงหมวดหมู่ที่ชัดเจน ตัวอย่างเอกสารที่ใช้ในการเรียนรู้การจัดหมวดหมู่แสดงได้ดังตารางที่ 1

Table 1 Example training documents

หมวดหมู่เอกสารที่ใช้ในการเรียนรู้	**ตัวอย่างเอกสาร**
● หมวดหมู่ของบุคคลจะต้องมีชื่อบุคคลปรากฏอยู่ในเอกสาร	- <DOC> วาน นี้ ว่า ประธานาธิบดี บิล ค ลิ น ตัน แห่ง สหรัฐ มี ความ วิตก กังวล เกี่ยว กับ เหตุการณ์ ต่าง ๆ ที่ แวดล้อม การ พิจารณา คดี โอ . เจ . ซิมพ์ สัน อดีต นัก อเมริกัน ฟุตบอล ชื่อ ดัง ของ สหรัฐ ที่ ตก เป็น </DOC> - <DOC> ที่ สังหาร นายกรัฐมนตรี ยิตซัค รา บิน แห่ง อิสราเอล เปิดเผย ว่า คณะ กรรมการ วินัย ของ มหาวิทยาลัย เปิด ประชุม กัน ใน วัน พฤหัสบดี เพื่อ พิจารณา ข้อ เสนอ ของ อธิการบดี ที่ แนะนำ ให้ ไล่ นาย ยิตซัค นิ ว แมน นัก ศึกษา ชั้น ปี </DOC>
● หมวดหมู่ของสถานที่จะต้องมีชื่อสถานที่ปรากฏอยู่ เช่น ชื่อประเทศ, ชื่ออำเภอ หรือ ชื่อจังหวัด เป็นต้น	- <DOC> หรือ เอ เปค ที่ นคร โอ ซา กา ของ ญี่ปุ่น ใน สัปดาห์ หน้า หาก ที่ ประชุม ยกเว้น ภาค เกษตร ออก จาก แผนการ ค้า เสรี เอ เปค ก่อน หน้า นี้ สมาชิก เอ เปค บาง ประเทศ เตือน ว่า เอ เปค อาจ เกิด </DOC> - <DOC> เวลา ๒ วัน ที่ เมือง โอ ปอร์ โต ของ โปรตุเกส ระบุ ว่า ยู ฟ่า ได้ ออก แผนการ ชื่อว่า " วิ ชั่น อี เลฟเว่น " เพื่อ เรียก ร้อง ให้ ฟีฟ่า ปฏิรูป องค์กร ใหม่ ให้ อยู่ ภาย ใต้ การ ดูแล ของ คณะ กรรมการ </DOC>
● หมวดหมู่ของเวลาจะต้องมีคำที่แสดงเวลาปรากฏอยู่ เช่น วันที่, เดือน หรือ ปี เป็นต้น	- <DOC> สถานี บริการ น้ำมัน ปตท . ประจำ วัน ที่ ๑ กันยายน ๒๕๓๘ หน่วย : บาท / ลิตร กรุงเทพ ฯ เชียงใหม่ เบนซิน พิเศษ (เพอร์ฟอร์ มา ๙๘) ๘ . ๘๑ ๙ . ๒๖ เบนซิน พิเศษ (ซูเปอร์ ๙๗) ๘ . ๙๑ </DOC> - <DOC> นิ วฟาวน์แลนด์ ใน คืน วัน ที่ ๑๔ เมษายน ๒๔๕๕ ใน การ เดิน ทาง เที่ยว แรก จาก อังกฤษ สู่ สหรัฐ ทำให้ ผู้ ที่ อยู่ บน เรือ เสีย ชีวิต ประมาณ ๑ , ๕๐๐ คน จาก ทั้งหมด ๒ , ๒๐๐ คน ต่อ มา เมื่อ ปี ๒๕๒๘ คณะ นัก </DOC>

6. **การเลือกคำตอบ** (Answer Selection)

ในงานวิจัยฉบับนี้จะทำการเลือกคำตอบโดยจะเลือกจากเอกสารที่มีหมวดหมู่เดียวกับหมวดหมู่ของคำตอบ นำมาคำนวณความสัมพันธ์ระหว่างคำในเอกสารและคำในประโยคคำถาม เพื่อเลือกเอกสารที่มีความสัมพันธ์กับคำถามมากที่สุดเป็นคำตอบ ซึ่งการเลือกคำตอบจะใช้*วิธีการหาอัตราเปรียบเทียบการเกิดร่วมกัน* (Co-occurrence ratio) (Kim et al., 2000) โดยมีแนวคิดที่ว่าในแต่ละเอกสารจะมีคำที่เหมือนกับในคำถามที่แตกต่างกัน ถ้าเอกสารใดมีคำที่เหมือนกับในประโยคคำถามมากกว่าเอกสารนั้นจะมีอัตราเปรียบเทียบมากกว่า และคาดว่าน่าจะเป็นคำตอบที่ถูกต้อง โดยที่ R_i คือ อัตราเปรียบเทียบการเกิดร่วมกัน

$$R_i = \frac{Number\ of\ question\ words\ appeared\ in\ the\ passage}{Total\ number\ of\ question\ words} \quad (1)$$

จากขั้นตอนการทำงานข้างต้น สามารถแสดงได้ดังรูปที่ 1

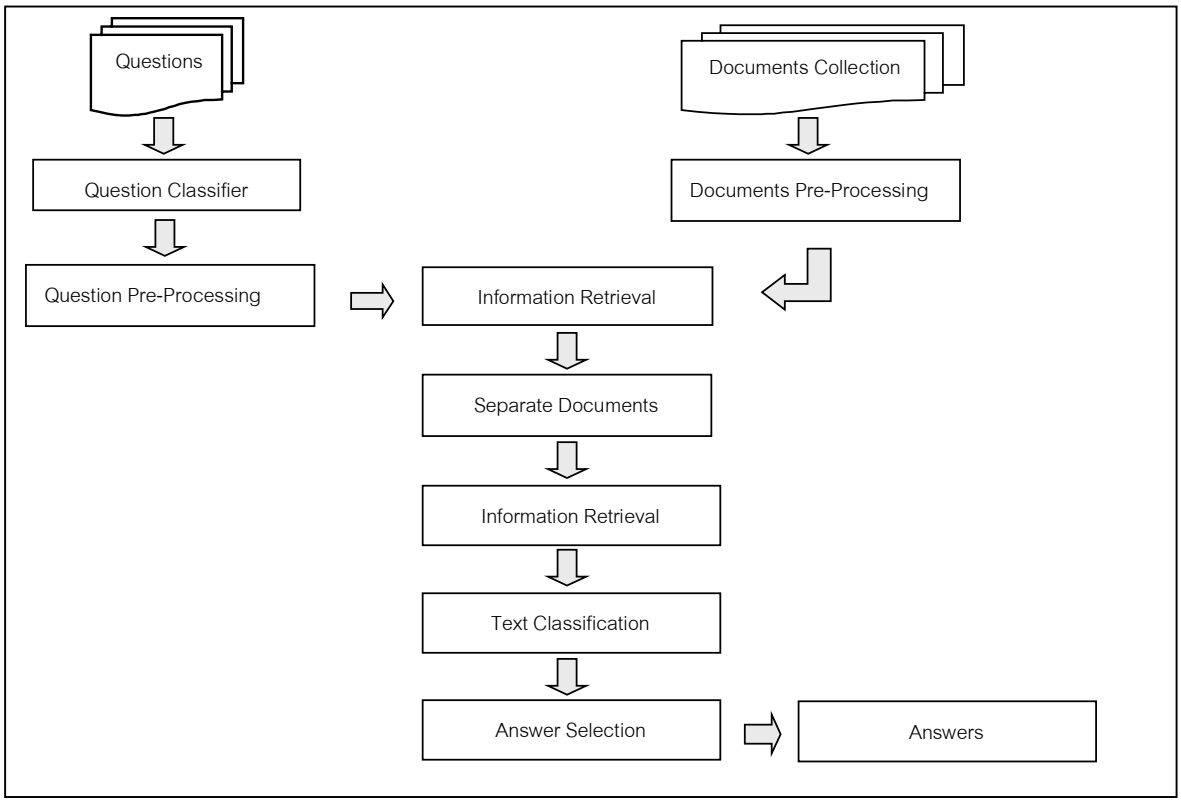

Figure 1 Stages of question answering system

ผลการทดลองและวิจารณ์

เอกสารที่ใช้ในงานวิจัยได้มาจากการเก็บรวบรวมข่าวภาษาไทยจากหนังสือพิมพ์เดลินิวส์ระหว่างวันที่ 1 กันยายน ถึง วันที่ 22 พฤศจิกายน พ.ศ. 2538 จำนวน 4,800 เอกสาร และนำคำถามที่ต้องการค้นหาซึ่งสร้างมาจากข้อความในเอกสารป้อนให้กับระบบคำถาม-คำตอบ ผลลัพธ์ที่ได้จะเป็นรายการของคำตอบดังที่แสดงในตารางที่ 2

Table 2 Results on processing

คำถาม	ผลลัพธ์ที่ได้จากการทดลอง
● ตัวอย่าง Q1: พ.ต.ท. ทักษิณ ชินวัตร คือใคร?	- **เพื่อประโยชน์ของตัวเองหรือทำลายใครพ.ต.ท.ทักษิณชินวัตรรองนายกรัฐมนตรีและหัวหน้าพรรคพลังธรรมกล่าวภายหลังการประชุมระหว่างหัวหน้าพรรคร่วมรัฐบาล๗พรรคในวันนี้(๖** - ว่ารัฐบาลยังมีผลงานเป็นที่น่าพอใจแต่หากเห็นว่าใครไม่เหมาะสมควรจะระบุชื่อเป็นรายคนพ.ต.ท.ทักษิณชินวัตรรองนายกรัฐมนตรีและหัวหน้าพรรคพลังธรรม - ต.ท.ทักษิณชินวัตรรองนายกรัฐมนตรีและหัวหน้าพรรคพลังธรรมในการจะเสนอให้คณะรัฐมนตรีพิจารณาอนุมัติให้บริษัททางด่วนกรุงเทพจำกัดหรือบีอีซีแอลได้รับสิทธิ์-ในการ - .ท.ทักษิณชินวัตรรองนายกรัฐมนตรีหัวหน้าพรรคพลังธรรมเป็นข่าวโคมลอยซึ่งไม่น่าจะเป็นไปได้นายถวิลกล่าวว่าขณะนี้พ.ต.ท.ทักษิณเป็นรองนายกรัฐมนตรีที่รับผิดชอบใน - ที่จะถอนตัวจากรัฐบาลพ.ต.ท.ทักษิณชินวัตรรองนายกรัฐมนตรีหัวหน้าพรรคพลังธรรมกล่าวว่าการที่รัฐบาลจะให้กรมประชาสัมพันธ์สำรวจความคิดเห็นของประชาชนที่มีต่อรัฐบาลหลังจาก
● ตัวอย่าง Q2: แผนพัฒนาเศรษฐกิจและสังคมแห่งชาติฉบับที่ 8 เริ่มในปีใด?	- **นโยบายเศรษฐกิจรวมทั้งการทำงานด้านเศรษฐกิจและด้านอื่น ๆคือการจัดทำแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติฉบับที่๘ซึ่งจะเริ่มใช้ตั้งแต่ปี๒๕๔๐คงจะทำให้เศรษฐกิจไทยไม่** - ได้เพิ่มงบประมาณเพื่อแก้ไขปัญหาจราจรไว้มากกว่าปีที่ผ่านมาถึงร้อยละ๙๐สำหรับแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติฉบับที่๘นั้นจะมีการสรุปโครงการจากแผนงานเดิม - ในแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติฉบับที่๘เน้นการพัฒนาคนซึ่งแบ่งเป็น๒ส่วนคือต้องพัฒนาคนให้เป็นเลิศทางปัญญาโดยพัฒนาคนให้มีพื้นฐานทางปัญญาที่แข็งแกร่ง - สถาบันวิจัยเพื่อการพัฒนา-ประเทศไทย(ทีดีอาร์ไอ)และคณะมาหารือกำหนดนโยบายวางแผนด้านการคมนาคมเพื่อบรรจุในแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติฉบับที่๘โดยวางเป้าหมาย
● ตัวอย่าง Q3: กรุงพนมเปญอยู่ที่ประเทศ ใด?	- กล่าวต่อสโมสรผู้สื่อข่าวต่างประเทศในกรุงพนมเปญเสนอแนะว่ารัฐบาลกัมพูชาจะสามารถจัดการกับปัญหาการสู้รบของเขมรแดงที่เป็นภัยต่อความมั่นคงของชาติได้สำเร็จก็ต่อ **- ร้องเรียนว่าเวียดนามยังคงกีดกั้นการใช้แม่น้ำโขงขนส่งสินค้าเข้าสู่กรุงพนมเปญนครหลวงของกัมพูชานายเต็งกล่าวว่าการขนส่งสินค้าผ่านแม่น้ำระหว่างประเทศเช่นแม่น้ำโขงไม่ควรจะถูกควบคุม** - พนมเปญ(เอเอฟพี)๒๒ต.ค.-สมเด็จเจ้านโรดมสีหนุกษัตริย์แห่งกัมพูชาเสด็จกลับถึงกรุงพนมเปญนครหลวงของประเทศแล้วในวันนี้หลังจากทรงใช้เวลาเกือบ - กรุงพนมเปญเมื่อวันเสาร์ที่ผ่านมาซึ่งทำให้มีผู้บาดเจ็บถึงอย่างน้อย๓๔คนแถลงการณ์ดังกล่าวออกโดยศูนย์สิทธิมนุษยชนของสหประชาชาติในกรุงพนมเปญมีใจความว่านายอ - พนมเปญ(รอยเตอร์)๑ก.ย.-เจ้าหน้าที่สถานทูตฝรั่งเศสประจำกรุงพนมเปญนครหลวงของกัมพูชาเปิดเผยในวันนี้ว่าประธานาธิบดีฌาคส์ชีรัคแห่งฝรั่งเศสได้กราบบังคมทูลเชิญ

จากรายการของประโยคคำตอบดังแสดงในตารางที่ 2 จะนำมาทำการวัดประสิทธิภาพของระบบโดยใช้การตรวจสอบผลลัพธ์ภายใน 5 อันดับที่ได้จากการประมวลผล โดยเลือกจากรายการของคำตอบที่มีค่าอันดับที่ดีที่สุด จะไม่คำนึงถึงการได้รับคำตอบที่ถูกต้องในหลายๆอันดับ

จากงานวิจัยของระบบคำถาม-คำตอบในงาน TREC (Text Retrieval Conference) จะใช้วิธีการวัดประสิทธิภาพของระบบโดยใช้ค่า Mean Reciprocal Answer Rank (MRAR) โดยใช้ผลลัพธ์ที่ได้จากรายการคำตอบใน 5 อันดับเฉลี่ยด้วยจำนวนคำถามทั้งหมดที่ใช้ในการทดสอบ ถ้าคำตอบไม่ได้อยู่ภายใน 5 อันดับ หรือไม่มีคำตอบที่ถูกต้องจะได้คะแนนเป็นศูนย์

$$\text{Reciprocal Answer Rank} = 1 / \text{Answer rank} \qquad (2)$$

จากการทำวิจัยจะทดสอบกับคำถามจำนวน 60 คำถาม แบ่งเป็นคำถามเกี่ยวกับบุคคล 20 คำถาม, คำถามเกี่ยวกับสถานที่ 20 คำถามและคำถามเกี่ยวกับเวลา 20 คำถาม การวัดประสิทธิภาพของระบบแสดงได้ดังตารางที่ 3

Table 3 The evaluation of question answering system

Correct rank #1	26
Correct rank #2	7
Correct rank #3	4
Correct rank #4	1
Correct rank #5	0
No correct answer	22
MRAR	0.518

ผลลัพธ์จากประโยคคำตอบที่ไม่ถูกต้องเนื่องมาจากข้อผิดพลาดที่เกิดจากการตัดคำในประโยคและการจัดเอกสารที่ไม่ตรงกับหมวดหมู่ที่ถูกต้อง

สรุปผลการทดลอง

งานวิจัยนี้นำเสนอระบบคำถาม-คำตอบที่ใช้กับข้อความภาษาไทย โดยการประยุกต์งานทางด้านการจัดหมวดหมู่เอกสารและอัตราการเปรียบเทียบการเกิดร่วมกัน เข้ามาเพื่อช่วยในการหาคำตอบให้กับคำถาม จากการทดลองประมวลผลป้อนคำถามให้กับระบบพบว่าให้ผลลัพธ์ที่ส่วนใหญ่จะถูกพบได้ จากเปอร์เซ็นต์ของการได้รับคำตอบ 63.33%

จากงานวิจัยระบบคำถาม-คำตอบที่มีการทดลองกับข้อความภาษาอังกฤษซึ่งไม่มีปัญหาจากการแยกคำได้ใช้การรวมวิธีของการสร้างตัวแทนฐานความรู้ (knowledge representation), การค้นคืนสารสนเทศ และ การประมวลผลภาษาธรรมชาติ ทดลองกับคำถามจำนวน 200 คำถาม ได้ประสิทธิภาพจากระบบคำถาม-คำตอบเพียง 0.381 และจากงานวิจัยที่ใช้ระบบการค้นคืนสารสนเทศ และการสกัดสารสนเทศ โดยทำการแบ่งเอกสารออกเป็นข้อความสั้นมีความยาว 250 ไบต์ จะได้ค่าการวัดประสิทธิภาพ คือ 0.545 ดังนั้นค่าการวัดประสิทธิภาพ

ของระบบคำถาม-คำตอบในงานวิจัยนี้ คือ 0.518 ผลที่ได้รับถือว่าเป็นค่าที่น่าพอใจในการให้คำตอบของระบบคำถาม-คำตอบที่ใช้กับข้อความภาษาไทย

ในงานวิจัยฉบับนี้ผลลัพธ์ที่ได้จะเป็นรายการของประโยคคำตอบ ซึ่งเป็นสารสนเทศที่ผู้ใช้จะต้องอ่านจากรายการเหล่านั้นเพื่อหาข้อมูลที่ต้องการจากคำถาม ในงานต่อไปสามารถที่จะนำรายการคำตอบที่ได้มาประยุกต์โดยการนำระบบการสกัดสารสนเทศเข้ามาเพื่อสกัดคำตอบที่ผู้ใช้ต้องการให้มีความชัดเจนมากขึ้น เช่น ถ้าเป็นคำถามเกี่ยวกับบุคคล คำตอบจะได้เฉพาะคำนามที่เป็นชื่อบุคคล เป็นต้น

เอกสารอ้างอิง

Bailey, S. M. 2001. Providing Question Coverage through Answer Type Classification in Natural Language Question Answering (NLQA) Systems. Master Thesis, Georgetown University, USA, April 2001.

Cooper, R. J., and S. M. Ruger. 2000. A Simple Answering System. In Proceedings of the Ninth Text REtrieval Conference (TREC 9).

Goharian, N., Ghazawi, T. E. , D. Grossman, and A. Chowdhury. 2000. On the Enhancements of a Sparse Matrix Information Retrieval Approach. International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2000).

Hovy, E., L. Gerber, U. Hermjakob, M. Junk, and C.Y. Lin. 2000. Question Answering in Webclopedia. In Proceedings of the Ninth Text REtrieval Conference (TREC 9).

Joho, H. 1999. Automatic detection of descriptive phrases for Question Answering System: A Simple pattern matching approach. The degree of Master of Science in Information Management, The University of Sheffield, United Kingdom.

Kim, S-M., D-H. Baek, S-B. Kim, and H-C. Rim. 2000. Question Answering Considering Semantic Categories and Co-occurrence Density. In Proceedings of the Ninth Text REtrieval Conference (TREC 9).

Kruengkrai, C., and C. Jaruskulchai. 2001. Thai Text Document Clustering using Parallel Spherical K-Means Algorithm on PIRUN Linux Cluster, The Fifth National Computer Science and Engineering Conference.

Li, X., and B. Croft. 2001. Evaluating Question-Answering Techniques in Chinese. In Proceedings of HLT 2001, First International Conference on Human Language Technology Research.

Michell, T. 1997. Machine Learning. McGraw-Hill.

Salton, G., and McGill J. M. 1983. Introduction to modern information retrieval. McGraw-Hill.

Sornlertlarmvanich, V. 1993. Word Segmentation for Thai in Machine Translation System. Machine Translation, Nation Electronics and Computer Technology Center, Bangkok.

Zheng, Z. 2002. AnswerBus Question Answering System. In Proceeding of HLT Human Language Technology Conference (HLT 2002).