

การสรุปใจความสำคัญของเอกสาร

Document Summarization

อดิชาติ ขานทอง , วัลลภา ตันติประสงค์ชัย, ชุติรัตน์ จรัสกุลชัย

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์

มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน

อีเมล : b4304421@ku.ac.th , b4304414@ku.ac.th , fscichj@ku.ac.th

บทคัดย่อ

รายงานฉบับนี้ อธิบายถึงวิธีการสรุปใจความสำคัญของเอกสารที่ได้มีการนำอัลกอริทึมที่มีใช้ในต่างประเทศ โดยใช้วิธีเลือกประโยคที่แสดงถึงเนื้อหาหลักของเอกสาร แล้วนำประโยคที่ได้มาสร้างเป็นใจความสำคัญของเอกสาร รวมถึงหลักการและทฤษฎีที่มีการนำมาใช้ในการเลือกประโยคที่จะนำมาเป็นใจความสำคัญของเอกสาร, การนำทฤษฎีการให้น้ำหนักของคำโดยใช้สูตรการหาค่า TF/IDF และการประเมินผลของใจความสำคัญที่ได้ เพื่อหาแนวทางในการสรุปใจความสำคัญของเอกสารภาษาไทยต่อไป

คำสำคัญ : การสรุปใจความสำคัญ, Information Retrieval, การประมวลผลภาษา

1. บทนำ

โลกแห่งเครือข่ายเป็นโลกที่เต็มไปด้วยมากมายมหาศาล จึงมีระบบค้นหาข้อมูลเกิดขึ้น ได้แก่ Search engine ต่างๆ เช่น AltaVista, Yahoo, Excite, Google ที่มีความสามารถในการค้นคืนเอกสารที่ต้องการจากเว็บมาให้ผู้ใช้อย่างมีประสิทธิภาพ แต่บางครั้งเอกสารที่ได้จากการค้นคืนมีจำนวนมาก ผู้ใช้จึงต้องเสียเวลาไม่น้อยกับการหาเอกสารที่ตรงกับความต้องการจริงๆ มีการเพิ่มประสิทธิภาพในการค้นคืนโดยงานวิจัยทางด้าน IR (Information Retrieval) ตัวอย่างเช่น การจัดกลุ่มเอกสาร (Document Classification) และการสืบค้นค้นข้ามภาษา (Cross-Language Information Retrieval) เป็นต้น การสรุปใจความสำคัญก็เป็นวิธีการหนึ่งที่ถูกนำมาใช้ แก้ปัญหาในส่วนที่ช่วยให้ผู้ใช้ทราบถึงเนื้อหาหลักของเอกสารโดยไม่ต้องอ่านเนื้อหาของเอกสารทั้งฉบับ จึงช่วยลดเวลาที่ใช้ในการค้นหาเอกสารที่ต้องการ

ปัจจุบัน ต่างประเทศได้ทำการวิจัยและพัฒนาแบบที่ใช้ในการประมวลผลทางด้านภาษาออกมาเป็นจำนวนมาก เช่น A General Language Model for Information Retrieval (S. Fei and W. Bruce Croft, 1999) , An Approach to Natural Language for Document Retrieval (B. Croft, 1987) สำหรับภาษาไทยก็มีผู้ทำการวิจัยโดยการนำอัลกอริทึมของต่างประเทศมาปรับปรุงเพื่อให้เหมาะกับภาษาไทยอยู่ไม่น้อย เช่น Thai Text Classification Based on Naïve Bayes (K. Canasai and J. Chuleerat, 2001) , การจัดกลุ่มเอกสารสำหรับข้อความภาษาไทย () แต่งานทางด้านสรุปใจความสำคัญยังมีผู้ทำการวิจัยและพัฒนาอยู่ เนื่องจากโครงสร้างประโยคของภาษาไทยความซับซ้อน การหาขอบเขตของประโยคจึงทำได้ยากกว่า ดังนั้นจึงต้องเน้นการเลือก

ประโยคให้ถูกต้องที่สุด โดยที่ยอมให้เกิดความผิดพลาดของการแยกประโยคบ้าง เพื่อที่จะให้เห็นแนวทางของการสรุปใจความสำคัญของเอกสารภาษาไทยในอนาคตถ้าสามารถแยกประโยคออกได้อย่างถูกต้องเพิ่มขึ้นก็จะทำให้ผลของการสรุปใจความสำคัญของเอกสารภาษาไทยก็จะดีขึ้นตามด้วย

2. แนวคิดทางด้านการสรุปใจความสำคัญ

ถึงแม้ว่าใจความสำคัญสามารถอธิบายได้ในหลาย ๆ รูปแบบ แต่ในที่นี้เราจะอธิบายเฉพาะหลักในการสร้างใจความสำคัญของเอกสารซึ่งประกอบไปด้วย วัตถุประสงค์ (intent), จุดความสนใจ (focus), และ ขอบเขต (coverage) (Firmin, T. and Chrzanowski, M.J., 1999)

วัตถุประสงค์ อธิบายถึงการนำใจความสำคัญไปใช้ซึ่งแบ่งเป็นการชี้แนะ (indicative) , การให้ข้อมูล (informative) และ การประเมินค่า (evaluative)

- ใจความสำคัญแบบชี้แนะ จะมีข้อมูลเพียงพอที่จะทำให้ผู้อ่านเข้าใจประเด็นหลักของเอกสารฉบับเต็ม หรือให้คำชี้แนะสั้นๆ เกี่ยวกับหัวเรื่องหลักของเอกสาร โดยทั่วไปแล้วใจความสำคัญประเภทนี้จะนำไปใช้แสดงให้แก่ผู้อ่านก่อนแสดงเอกสารฉบับเต็ม

- ใจความสำคัญแบบให้ข้อมูล เป็นเหมือนตัวแทนของเอกสารฉบับเต็ม จะเก็บข้อมูลรายละเอียดที่สำคัญ โดยในขณะเดียวกันก็ลดปริมาณข้อมูลที่ต้องแสดงแก่ผู้อ่านด้วย

- ใจความสำคัญแบบประเมินค่า จะเก็บเนื้อหาตรงจุดที่เป็นผู้แต่งในหัวเรื่องที่กำหนดให้

งานวิจัยทางด้านการสรุปใจความสำคัญในสมัยก่อน จะมุ่งไปยังการสร้าง ใจความสำคัญประเภทการชี้แนะ (indicative) ตัวอย่างงานวิจัย เช่น Luhn (Luhn H. P., 1958), Edmundson (Edmundson, H. P., 1969), และ Brandow, Mitze and Rav (Brandow, R, Mitze K and Rau ,L. F., 1995) ก็ได้สร้างใจความสำคัญประเภทนี้ ถึงแม้ว่าเทคนิคที่ใช้จะแตกต่างกัน แต่วิธีการเหล่านี้ได้ถูกยืนยันแล้วว่าสามารถนำไปใช้ได้กับเอกสารทั่วไปถึงแม้ว่าจะทำการทดสอบเฉพาะกับเอกสารประเภทใดประเภทหนึ่งเท่านั้น เช่น ข้าว (Brandow, R, Mitze K and Rau ,L. F., 1995) , สิ่งตีพิมพ์เฉพาะทาง (Edmundson, H. P., 1969) เป็นต้น

จุดความสนใจ หมายถึงขอบเขตของข้อมูลที่สนใจอาจจะเป็นแบบข้อมูลทั่วไป (generic) หรือข้อมูลที่ตรงกับคำขออย่างใดอย่างหนึ่ง ใจความสำคัญแบบข้อมูลทั่วไปมีพื้นฐานอยู่บนสาระสำคัญหลักหรือ

สาระสำคัญของเอกสาร ในขณะที่ใจความสำคัญแบบข้อมูลตรงกับคำขอ ถูกสร้างจากหัวเรื่องโดยเฉพาะซึ่งกำหนดโดยผู้ที่ต้องการใจความสำคัญนั้น

ขอบเขต จะเป็นตัวบอกว่าใจความสำคัญนี้มีในเอกสารฉบับเดียวหรือมีเอกสารหลายฉบับที่เกี่ยวข้องในหัวเรื่องนั้นๆ

แนวทางของการสรุปใจความสำคัญในงานวิจัยต่างๆ อาจแบ่งตามลักษณะได้เป็น (Aone, C., Gorlinsky, J. and Others, 1999)

- ยึดตามค่าความถี่เป็นหลักสำคัญ (frequency-base) วิธีการนี้จะดูที่ความถี่และตำแหน่งของคำในเอกสารเป็นหลัก ตัวอย่างเช่น การนับความถี่ของคำสำคัญ, ระยะห่างระหว่างคำสำคัญในประโยค, งานวิจัยที่ใช้วิธีนี้ เช่น งานวิจัยของ Luhn (Luhn, 1958), Edmundson (Edmundson, 1969) และ Rush, Salvador, และ Zamora (Rush, Salvador and Zamora, 1971)

- ยึดตามฐานความรู้ (knowledge-base) โครงสร้างหลักของเอกสารที่ได้ จะขึ้นอยู่กับการศึกษาจากฐานความรู้ที่มีอยู่ โดยใช้เทคนิคของการแยกแยะ (classification) เช่น KPC ของ Kupiec, Pederson, และ Chen (Kupiec, Pederson, and Chen, 1995), TOPIC ของ Reimer และ Hahn (Reimer and Hahn, 1988) จะเน้นทางด้านใดด้านหนึ่งโดยเฉพาะ อาจมีการสร้างเป็นโครงรูป (template) ตามประเภทของเอกสาร

- ยึดตามความสอดคล้องและความเข้ากันได้ (discourse-base) วิธีการนี้จะดูที่ความเข้ากันได้และความสอดคล้องกันของประโยค โดยยึดตามโครงสร้างประโยคของภาษานั้นๆ ใช้แก้ปัญหาประโยคที่ได้จาก frequency-base ไม่สอดคล้องกัน โดยใช้วิธีของการทำดัชนี (indexing) เช่น Brandow, Mitze, และ Rau (Brandow, Mitze, and Rau, 1995)

วิธีการสรุปใจความสำคัญ ยังสามารถแบ่งออกเป็น 2 ประเภทตามวิธีการสร้าง คือ วิธีดึงจากต้นฉบับ (extract) และ วิธีจัดทำเป็นบทคัดย่อ (abstract) ในระบบการสรุปใจความสำคัญโดยทั่วไป จะทำการสร้างใจความสำคัญโดยใช้ประโยคหรือย่อหน้า (paragraph) เป็นหน่วยพื้นฐานโดยจะให้ระดับความสำคัญของแต่ละหน่วย, ทำการเรียงหน่วยต่างๆ ตามระดับความสำคัญ, และรวบรวมประโยคที่สำคัญมาเป็นใจความสำคัญของเอกสาร กล่าวโดยสรุปใจความสำคัญก็คือ กลุ่มของประโยคที่ถูกแยกออกมาจากต้นฉบับ เราจะเรียกใจความสำคัญประเภทนี้ว่า ใจความสำคัญที่ดึงจากต้นฉบับ (extraction summary) (Mani, I. and Maybury, M. T., 1999)

การสรุป ใจความสำคัญแบบดึงจากต้นฉบับ จะมีวิธีการแตกต่างจากการสรุป ใจความสำคัญแบบจัดทำเป็นบทคัดย่อ ในรายงานนี้ จะอธิบายถึงขั้นตอนในการสรุป ใจความสำคัญแบบดึงจากต้นฉบับ ซึ่งมีขั้นตอนหลักๆ 3 ขั้นตอน คือ การแยกคำสำคัญ, การแยกประโยค, และการเรียงประโยคเป็นใจความสำคัญ (Buyukkokten, O., Garcia-Molina, H. and Peapcke, A., 2000) ส่วนการสรุป ใจความสำคัญแบบจัดทำเป็นบทคัดย่อ สามารถแบ่งออกเป็น 4 ขั้นตอน คือ

1. เอกสารต้นฉบับจะถูกนำไปทำการแปลเพื่อให้ระบบทราบความหมายที่จะเป็นตัวแทนของเอกสาร

2. ตัวแทนของเอกสาร จะถูกนำไปใช้ในการแยกข้อมูลที่เป็นประเด็นสำคัญ ซึ่งรวมถึงหัวเรื่องของเอกสารด้วย

3. ข้อมูลเหล่านั้นจะถูกกลั่นกรองและกำจัดข้อมูลที่ซ้ำๆ กันออกไป

4. ข้อมูลจะถูกนำเสนอต่อผู้อ่านในรูปแบบใหม่

โดยส่วนใหญ่วิธีการสร้างใจความสำคัญจะมุ่งความสนใจไปยังขั้นตอนที่ 1. และ 2. (Saggion, H. and Lapalme, G., 2000)

ปัจจุบันได้มีการพัฒนาระบบที่ใช้ในการสรุปใจความสำคัญของเอกสารขึ้นหลายระบบด้วยกัน ในการสรุปใจความสำคัญของเอกสารระบบจะทำการคัดเลือกประโยคที่แสดงถึงเนื้อหาหลักของเอกสารขึ้นมาจำนวนหนึ่ง เพื่อนำมาสร้างเป็นใจความสำคัญของเอกสาร ซึ่งจะพิจารณาในแต่ละประโยคในเอกสารว่าประโยคใดควรเป็นใจความสำคัญของเอกสาร โดยในรายงานนี้จะนำเสนอทฤษฎีของ H.P. Luhn ซึ่งเป็นพื้นฐานและง่ายต่อการนำไปพัฒนาเพื่อพิจารณาความสำคัญของประโยค กระบวนการทำงานแบ่งออกเป็นขั้นตอนได้ดังนี้

1. การแยกคำสำคัญ

การแยกคำสำคัญในเอกสารจะขึ้นอยู่กับการศึกษาความสำคัญของคำในเอกสาร ความสำคัญของคำ W จะขึ้นอยู่กับจำนวนครั้งของคำ W ที่ปรากฏในประโยค โดยปกติแล้ว คำที่อยู่ในประโยคจะถูกพิจารณาว่าเป็นคำสำคัญถ้าคำนั้นปรากฏหลายครั้งในประโยค แต่ปรากฏน้อยครั้งในจำนวนเอกสารทั้งหมด หลักการนี้ถูกนำมาใช้ในสูตรการหาค่า TF/IDF (Salton, G., 1989)

$$w_{ij} = tf_{ij} \times \log_2 \frac{N}{n_j}$$

โดยที่ w_{ij} = น้ำหนักของคำ T_j ในเอกสาร D_i
 tf_{ij} = ความถี่ของคำ T_j ในเอกสาร D_i
 N = จำนวนเอกสารทั้งหมดในระบบ
 n_j = จำนวนเอกสารที่มีคำ T_j ปรากฏอยู่อย่างน้อยหนึ่งครั้ง

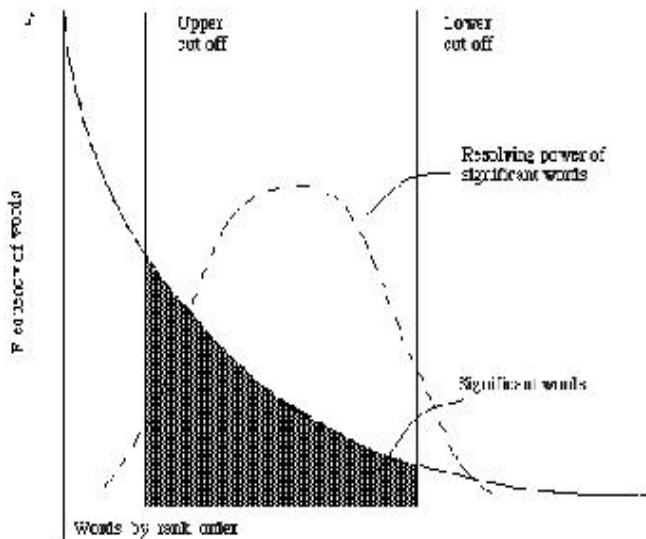
รูปที่ 1 แสดงสมการการคำนวณค่า TF/IDF

พารามิเตอร์ n_j ในสมการนี้จะหาได้เมื่อระบบได้รับรู้ทุกเอกสารที่มีในระบบ ในกรณีนี้เราจะต้องมีฐานข้อมูลของคำไว้ให้ระบบ โดยคำในฐานข้อมูลนี้ได้มีการคัดเลือกคำหยุด (stop word) ออกไปแล้ว เพื่อกรองคำที่ไม่สื่อถึงความสำคัญของเอกสาร ดังนั้น คำที่มีค่าความสำคัญในขอบเขตที่กำหนดจะถูกเลือกเป็นคำสำคัญ (significant word) และจะถูกจัดเก็บลงในฐานข้อมูล

2. การแยกประโยค

ในการสรุปใจความสำคัญ เอกสารที่ต้องการสรุปใจความสำคัญจะถูกนำไปผ่านขั้นตอนของการแยกประโยค และนำมาวิเคราะห์ถึงความสำคัญของประโยคเป็นผลต่อเนื่องมาจากความถี่ของคำสำคัญและตำแหน่งของคำสำคัญที่พบในประโยคนั้น

เริ่มแรก เราทำการพิจารณาคำทั้งหมดเพื่อหาคำสำคัญที่จะใช้เป็นกลุ่มคำซึ่งเป็นตัวแทนของเอกสารโดยดูจากความถี่ของคำทั้งหมดในเอกสารดังรูปที่ 2 โดยในส่วนของคำที่มีความถี่สูงเกินไป (ทางซ้ายของเส้น upper cut-off) จะถูกตัดทิ้ง และส่วนของคำที่มีความถี่น้อยเกินไป (ทางขวาของเส้น lower cut-off) ก็จะถูกตัดทิ้ง ก็จะได้ขอบเขตของคำที่มีประโยชน์ในการพิจารณาความสำคัญของประโยค



รูปที่ 2 แสดงกราฟของคำสำคัญ

ถัดมา เราจะทำการพิจารณาคำสำคัญในประโยค คำที่พิจารณาจะเป็นคำสำคัญก็ต่อเมื่อค่า TF/IDF ของคำนั้นมีค่ามากกว่า คำน้หนักที่ตัดออก (weight cutoff) ที่เรากำหนดไว้ คำน้หนักที่ตัดออก เป็นพารามิเตอร์ที่ต้องมีการปรับเปลี่ยนค่าตามความเหมาะสม ในขั้นตอนต่อมา เราจะหา กลุ่มคำ (cluster คือ คำที่เรียงต่อเนื่องกันในประโยค) ทั้งหมดในประโยค S ซึ่งเมื่อเงื่อนไขต่อไปนี้เป็นจริง

- คำเริ่มต้นและคำสุดท้ายของกลุ่มคำที่ต่อเนื่องกันเป็นคำสำคัญ
- ระยะห่างระหว่างกลุ่มคำที่ต่อเนื่องกันต้องมีค่าไม่เกินค่าที่กำหนดไว้ โดยเรียกค่านีว่า ระยะของการตัดออก (distance cutoff) และยังเป็นพารามิเตอร์ที่ต้องมีการปรับเปลี่ยนเช่นกัน

ในประโยคหนึ่งๆ อาจมีได้หลายกลุ่มคำ หลังจากเราหากลุ่มคำทั้งหมดในประโยคได้แล้ว เราจะทำการคำนวณค่าน้หนักของแต่ละกลุ่มคำ คำน้หนักที่มีค่าสูงสุดที่ได้จากการคำนวณจะถูกนำมาเป็นค่าน้หนักของประโยค Luhn (Luhn, H. P., 1999) คำน้หนักของกลุ่มคำหาได้จากจำนวนคำสำคัญที่มีในกลุ่มคำยกกำลังสอง หารด้วยจำนวนคำทั้งหมดที่มีในกลุ่มคำ จากรูปที่ 3 คำน้หนักของกลุ่มคำมีค่าเท่ากับ $4^4/7 = 2.2857142857$

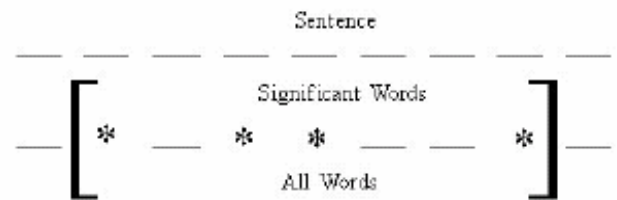
$$\text{term factor} = \frac{n_s^2}{n_c}$$

โดยที่ term factor = คำน้หนักของกลุ่มคำ

n_s = จำนวนคำสำคัญที่พบในกลุ่มคำ

n_c = จำนวนคำทั้งหมดในกลุ่มคำ

รูปที่ 3 แสดงสมการการคำนวณค่าน้หนักของกลุ่มคำ



รูปที่ 4 แสดงการหากลุ่มคำในประโยค

3. การเรียงประโยคเพื่อสร้างใจความสำคัญ

หลังจากขั้นตอนการหาคำความสำคัญของแต่ละประโยคในเอกสาร ขั้นตอนต่อไปคือ นำประโยคที่ได้มาเรียงลำดับตามค่าน้หนักจากมากไปน้อยแล้วเลือกประโยคที่มีค่าน้หนักสูงตามที่กำหนดมาเป็นใจความสำคัญของเอกสาร และนำประโยคที่ผ่านการพิจารณาแล้วนั้นมาเรียงตามใจความสำคัญของเอกสารต้นฉบับ

4. การประเมินคุณภาพของใจความสำคัญ

การประเมินคุณภาพของใจความสำคัญก็เป็นอีกหัวข้อหนึ่งที่มีการศึกษากันอย่างลึกซึ้งและยังเป็นปัญหาวិถยในการหาวิธีการประเมินผลที่เหมาะสม มีหลักการหลายอย่างที่นำไปได้สำหรับการเปรียบเทียบประสิทธิภาพของการสรุปใจความสำคัญ เช่น เปรียบเทียบระหว่างใจความสำคัญกับเอกสารต้นฉบับ, ใจความสำคัญที่สร้างโดยระบบกับใจความสำคัญที่สร้างโดยมนุษย์, ใจความสำคัญที่สร้างโดยระบบหนึ่งกับอีกระบบหนึ่ง โดยทั่วไปวิธีสำหรับการประเมินใจความสำคัญสามารถแบ่งออกกว้างๆ ได้ 2 ประเภท (Mani, I. And Maybury, M. T., 1999) คือ

- การประเมินภายใน (intrinsic evaluation) เป็นการเปรียบเทียบระหว่างใจความสำคัญที่สร้างจากระบบ (automatic summary) กับใจความสำคัญที่เป็นใจความที่ถูกต้องสำหรับเอกสารนี้ (ideal summary หรือ target summary) ซึ่ง target summary นี้ อาจสร้างมาจากผู้เชี่ยวชาญทางด้านภาษาในหัวเรื่องนั้น ๆ ในการเปรียบเทียบจะดูถึงความถูกต้อง (precision) และความครบถ้วนของใจความสำคัญที่ได้ ซึ่งใจความสำคัญของระบบที่ดี ควรมีเนื้อหาใกล้เคียงกับเนื้อหาที่มีใน target summary โดยพื้นฐานแล้ว การประเมินใจความสำคัญแบบนี้คำนวณได้จาก

$$\text{precision } A = \frac{n_c}{n_A}$$

โดยที่ A = ใจความสำคัญที่สร้างจากระบบ

n_c = จำนวนประโยคใน A ที่พบในใจความสำคัญที่ถูกต้อง

n_A = จำนวนประโยคทั้งหมดใน A

รูปที่ 5 แสดงสมการการคำนวณค่าความถูกต้องของใจความสำคัญที่สร้างจากระบบ

$$recall\ A = \frac{n_c}{n_r}$$

โดยที่ A = ใจความสำคัญที่สร้างจากระบบ
 n_c = จำนวนประโยคใน A ที่พบในใจความสำคัญที่ถูกต้อง
 n_r = จำนวนประโยคที่ได้จากใจความสำคัญที่ถูกต้อง

รูปที่ 6 แสดงสมการการคำนวณค่าความครบถ้วนของใจความสำคัญที่สร้างจากระบบ

นอกจากนี้ยังสามารถดูได้จากความสอดคล้องกันของประโยคในใจความสำคัญที่สร้างจากระบบ เช่น การสะกด, ไวยากรณ์, ความชัดเจน, ความเห็นส่วนบุคคล, คำย่อ เป็นต้น ซึ่งลักษณะเหล่านี้จำเป็นที่จะต้องมีการประเมินค่าจะเป็นผู้ประเมิน

- การประเมินภายนอก (extrinsic evaluation) หรือ task-oriented เป็นการประเมินประสิทธิภาพของการนำระบบสรุปใจความสำคัญไปใช้ร่วมกับงานประยุกต์ต่าง ๆ เช่น การค้นหาเอกสารที่ต้องการจากกลุ่มเอกสาร, สร้างรายงานหรือข้อนำเสนอลักษณะของการสรุป, ระบบถามตอบปัญหา เป็นต้น โดยทั่วไปจะดูถึงความถูกต้องที่ได้ (accuracy) เทียบกับเวลาที่ใช้ (time) แต่จะเห็นว่าในส่วนของความถูกต้องยังคงต้องมีกลุ่มคนที่จะเป็นผู้ประเมิน อีกตัวอย่างของการประเมินแบบนี้ คือ การตั้งคำถามแล้วให้ผู้ประเมินอ่านใจความสำคัญที่สร้างจากระบบ เพื่อตอบคำถามนั้น

3. ปัญหาของการสรุปใจความสำคัญ

การสรุปใจความสำคัญของเอกสารบางครั้งใจความที่ได้ไม่สามารถสื่อให้ผู้ให้ทราบถึงสาระสำคัญหลักที่แท้จริงได้ พิจารณาตัวอย่างเรื่องต่อไปนี้

John and Bill wanted money. They bought masks and guns and stole an old car from a neighbor. Wearing their masks and waving their guns, the two entered the bank, and within minutes left the bank with several bags of \$100 bills. They drove away happy, throwing away the masks and guns in a sidewalk trash can. They were never caught

ตารางที่ 1 แสดงตัวอย่างของเนื้อหาที่นำมาสรุปใจความสำคัญ

วิธีที่ใช้ในการสรุปใจความสำคัญแบบที่อธิบายในรายงานนี้ จะได้คำสำคัญที่แสดงถึงหลักของเรื่องนี้ คือ masks และ guns ซึ่งคำทั้งสอง

คำนี้ได้ปรากฏในเอกสารจำนวน 3 ครั้ง ซึ่งมากกว่าคำอื่น ๆ ในเอกสาร แต่จะเห็นได้ว่า เนื้อเรื่องในเอกสาร เป็นเรื่องที่เกี่ยวข้องกับการปล้น (robbery) ซึ่งคำสำคัญที่ได้มา ไม่สามารถสื่อให้ผู้อ่านทราบได้ว่าเนื้อเรื่องในเอกสารนี้เป็นเรื่องเกี่ยวกับการปล้น (Hovy, E. and Lin, C. Y., 1999)

นอกจากนี้ ยังมีปัญหาที่พบหลักๆ อีก 3 อย่าง (Ganapathiraju, M. K., 2002) คือ

- ใจความสำคัญแบบดึงจากต้นฉบับ บางครั้งจะมีประโยคที่ไม่เป็นใจความสำคัญของเอกสารรวมเข้ามาด้วย ทำให้ใจความสำคัญที่ได้มีความยาวมากเกินไป

- ข้อมูลที่สำคัญ หรือข้อมูลที่สัมพันธ์กันมักกระจายไปในแต่ละประโยค ใจความสำคัญแบบดึงจากต้นฉบับไม่สามารถที่จะจับข้อมูลเหล่านี้ได้ (ยกเว้นว่าใจความสำคัญมีความยาวเพียงพอที่จะรวมข้อมูลเหล่านี้เข้าไปด้วย

- ข้อมูลที่ขัดแย้งกัน อาจถูกแสดงอย่างไม่ชัดเจนทำให้ความหมายของเอกสารผิดเพี้ยนไป

4. ผลการประเมินระบบสรุปใจความสำคัญ

ในการประเมินระบบสรุปใจความสำคัญ ในระบบนี้ได้เลือกวิธีการวัดคุณภาพของใจความสำคัญที่ได้จากระบบโดยการดูจากระดับความพอใจที่ได้จากผู้ใช้ โดยมีวิธีการ คือ ทำการสุ่มเลือกเอกสารมาจำนวน 30 เอกสาร จากนั้น นำเอกสารที่เลือกไว้มาทำการสรุปใจความสำคัญโดยระบบ และทำใจความสำคัญที่ได้ไปให้ผู้ใช้อ่านต้นฉบับและใจความสำคัญของต้นฉบับนั้น และทำการให้ระดับความพอใจแก่ใจความสำคัญ โดยมีให้เลือก 3 ระดับ คือ พอใจมาก, พอใช้ได้ และยอมรับไม่ได้

ระดับ พอใจมาก หมายความว่า ผู้ใช้รู้สึกว่าใจความสำคัญนั้นสามารถสรุปได้ถูกต้อง และสามารถทำให้เข้าใจในเนื้อหาหลักของต้นฉบับได้ ระดับ พอใช้ได้ หมายความว่า ผู้ใช้รู้สึกว่าใจความสำคัญนั้นยังไม่ดีเท่าที่ควร แต่ก็ยังทำให้พอจะรู้เนื้อหาหลักคร่าว ๆ ของต้นฉบับ และอาจจะต้องการการปรับเปลี่ยนอีกเล็กน้อยจึงจะสามารถอยู่ในระดับที่ พอใจมากได้ ระดับสุดท้าย คือ ระดับ ยอมรับไม่ได้ หมายความว่า ใจความสำคัญนั้นเป็นใจความสำคัญที่สรุปได้ไม่ถูกต้อง และเป็นใจความสำคัญที่อ่านแล้วไม่สามารถทราบเนื้อหาหลักของต้นฉบับได้เลย ผลการประเมินที่ได้ คือ ระดับ พอใจมาก 53%, ระดับ พอใช้ได้ 11% และที่เหลือคือระดับ ยอมรับไม่ได้

อย่างไรก็ตาม ในการประเมินคุณภาพของระบบสรุปใจความสำคัญด้วยวิธีนี้ ยังไม่เป็นวิธีที่ได้มาตรฐานนัก เนื่องจาก ประเภทของเอกสารที่สุ่มมายังไม่หลากหลายมากนัก และใช้ผู้ประเมินเพียง 30 คน ซึ่งอาจถือได้ว่า จำนวนผู้ประเมินยังน้อยเกินไปในการประเมินคุณภาพของระบบสรุปใจความสำคัญ

5. บทสรุป

ในการสรุปใจความสำคัญถึงแม้ว่าจะยังไม่มีหลักการในการนำไปใช้อย่างแน่นอน เนื่องจากในงานวิจัยต่างๆ ก็ยังจำกัดประเภทของเอกสารที่ใช้อีกทั้งยังไม่มีผลการประเมินผลที่ได้จากการสรุปใจความสำคัญที่ใช้เป็นมาตรฐานอีกด้วย ทำให้จำเป็นที่จะต้องสร้างมาตรฐานเพื่อใช้วัด

ความถูกต้อง ในปัจจุบันเป็นที่สังเกตว่าได้มีการนำทฤษฎีทางด้าน ปัญญาประดิษฐ์เข้ามาช่วยในส่วนของ การสร้างใจความสำคัญ ทำให้ระบบ มีการเรียนรู้จากตัวอย่างที่ถูกต้องและนำไปใช้ทำนายกับเอกสารจริงเพื่อ ช่วยเพิ่มประสิทธิภาพ ซึ่งแนวทางในการทำการสรุปใจความสำคัญของ ภาษาไทยนั้นอาจจำเป็นต้องนำเอาทฤษฎีเหล่านี้มาช่วยเนื่องจากการใช้ ความถี่ของคำตามทฤษฎีของ Luhn ดังที่กล่าวมานั้นอาจไม่เพียงพอ เพื่อให้ได้ประโยคที่สมบูรณ์ขึ้นแทนที่จะไปปรับปรุงการแยกประโยคให้ ถูกต้องซึ่งต้องพิจารณาตัวแปรที่มีผลอื่นๆ อีกมากซึ่งมีมากกว่า การนำ ระบบการสรุปใจความสำคัญไปประยุกต์ใช้

เอกสารอ้างอิง

1. Aone,C., Gollins, J. and Others. 1999. *A trainable Summarizer with knowledge Acquired from Robust NLP Techniques*. In *Advances in Automatic Text Summarization*, 71-80. The MIT Press, England.
2. Buyukkokten, O., Garcia-Molina, H. and Peapcke, A. 2000. *Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices*. In *Proceedings of the Tenth International World Wide Web Conference*. Hong Kong.
3. Croft B. 1987. *An approach to natural language for document retrieval*. In *Proceeding of the 10th annual international ACM SIGIR conference on research and development in information retrieval*, Louisiana, United States
4. Fei S. and Croft B. 1999. *A General Language Model for Information Retrieval*. In *Proceeding of the eighth international conference on information and knowledge management*, Missouri, United States
5. Firmin, T. and Chrzanowski, M. J. 1999. *An Evaluation of Automatic Text Summarization Systems*. In *Advances in Automatic Text Summarization*, 325-339. The MIT Press, England.
6. Fukusima, T. and Okumura, M. 2001. *Text Summarization Challenge Text summarization evaluation at NTCIR Workshop2*. In *Proceedings of the Second NTCIR Workshop Meeting*, 45-51.
7. Ganapathiraju, M. K. 2002. *Relevance of Cluster size in MMR based Summarizer: A Report*. Self-paced lab in Information Retrieval.
8. Hovy, E. and Lin, C. Y. 1999. *Automated Text Summarization in SUMMARIST*. In *Advances in Automatic Text Summarization*, 81-94. The MIT Press, England.
9. Kruengkrai C. and Jaruskulchai C. 2001. *Thai Text Classification Based on Naive Bayes*, Technical Report , Department of Computer Science, Kasetsart University.
10. Luhn, H. P. 1999. *The Automatic Creation of Literature Abstracts*. In *Advances in Automatic Text Summarization*, 15-21. The MIT Press, England.
11. M.-Y. Kan, J. L. Klavans, and K. R. McKeown. 2001. *Domain-specific informative and indicative summarization for information retrieval*. In *Proceedings of the Document Understanding Conference*, New Orleans, USA
12. Mani, I. And Maybury, M. T. 1999. *Advanced in Automatic Text Summarization*. England, The MIT Press.
13. Saggion, H. and Lapalme, G. 2000. *Selective Analysis for Automatic Abstracting: Evaluating Indicativeness and Acceptability*. RIAO'2000, 12-14. Paris, France.
14. Salton, G. 1989. *Automatic Text Processing*. In Addison-Wesley Chapter 9.
15. ชุติรัตน์ จรัสกุลชัย, เจษฎา กันทะเสนา และ สภาพร คิวสุวรรณสุข, *การจัดกลุ่มเอกสารสำหรับข้อความภาษาไทย*, The Fifth National Computer Science and Engineering Conference, 7 - 9 November 2001, Chiang Mai University, Chiang Mai, Thailand