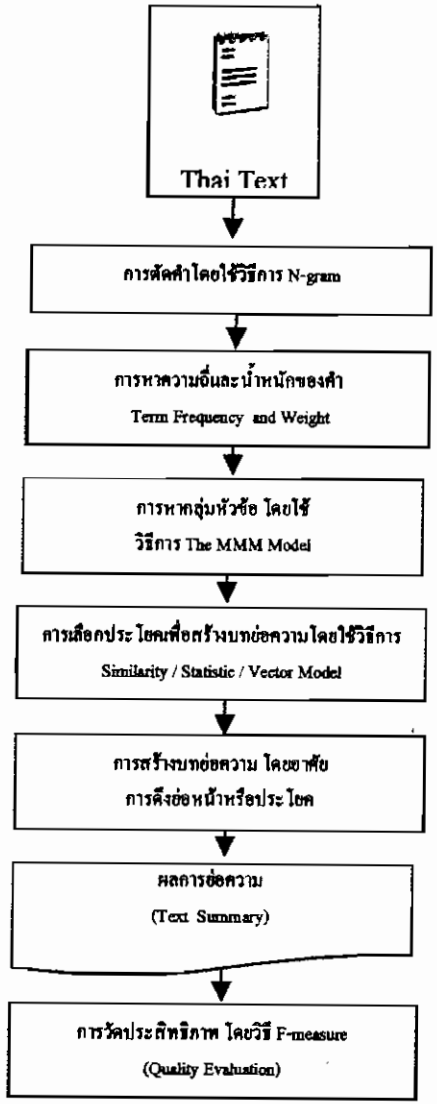
.: ขั้นตอนการดำเนินงาน :.



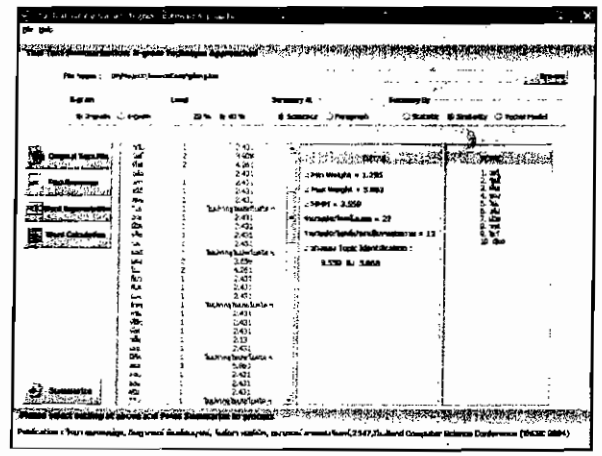
.:Publication:.

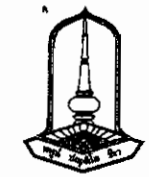
วีรยา อมรพงษ์กูล, อัษฎาภรณ์ พิมพ์สมบูรณ์, จันทิมา พลพินิจ, อุมาภรณ์ สายแสงจันทร์, 2547, การประยุกต์อินแกรมและเวกเตอร์โมเดลสำหรับระบบย่อข้อความภาษาไทย, Thailand Computer Science Conference (ThCSC 2004), มหาวิทยาลัยกษตรศาสตร์, กรุงเทพฯ.

.: สรุป :.

งานวิจัยฉบับนี้นำเสนอระบบย่อข้อความเอกสารเดียว แบบอัตโนมัติโดยการสกัดย่อหน้า ซึ่งเป็นการประยุกต์วิธีการแบบ *N-gram* เข้ามาใช้แทนการตัดคำด้วยพจนานุกรม จากนั้นจึงนำไปหา *Topic Identification* เพื่อใช้เป็นส่วนสำคัญในการสกัดย่อหน้าที่สำคัญจากเอกสาร นอกจากนั้นยังได้นำอาวิธีการเข้าคู่เอกสารแบบเวกเตอร์โมเดล มาผสมผสานกับวิธีการวัดค่าความคล้ายคลึง เพื่อสร้างบทย่อความที่มีความเกี่ยวข้องระหว่างกลุ่มคำที่เป็น *Topic Identification* และเอกสารต้นฉบับ ซึ่งผลจากการวัดประสิทธิภาพของระบบ ด้วย *F -measure* ปรากฏว่าวัดค่าของประสิทธิภาพได้ระหว่าง 41%-42% เมื่อเป็นการย่อความที่ระดับ 20% จากขนาดของเอกสารทั้งหมดและได้ค่าระหว่าง 54%-56% เมื่อเป็นการย่อความที่ระดับ 40%

.: ตัวอย่างโปรแกรม :.

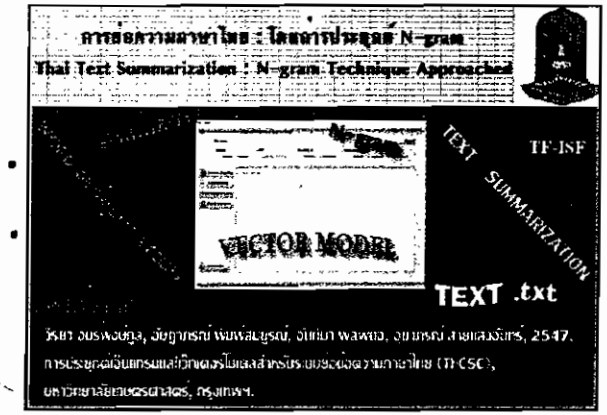




PPA115

การย่อความภาษาไทย : โดยการประยุกต์ N-gram

Thai Text Summarization : N-gram Technique Approached



วีรยา อมรพงษ์กูล

อัษฎาภรณ์ พิมพ์สมบูรณ์

อาจารย์จันทิมา พลพินิจ

อาจารย์อุมาภรณ์ สายแสงจันทร์

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

โครงการการแข่งขันพัฒนาโปรแกรมคอมพิวเตอร์แห่งประเทศไทยครั้งที่ 7

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

.: หัวข้อ :.

การย่อความภาษาไทย : โดยการประยุกต์ N-gram

Thai Text Summarization : N-gram Technique Approached

.: คณะผู้จัดทำ :.

วีรยา อมรพงษ์กูล

อัษฎาภรณ์ พิมพ์สมบูรณ์

อาจารย์จันทิมา พลพินิจ

อาจารย์อุมาภรณ์ สายแสงจันทร์

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

.: บทคัดย่อ :.

เอกสารฉบับนี้นำเสนอระบบย่อข้อความภาษาไทยแบบการสกัดย่อหน้าที่สำคัญที่สุดจากเอกสารต้นฉบับที่เป็นข้อความภาษาไทย โดยระบบนี้จะแบ่งการทำงานออกเป็น 2 ส่วนหลักคือ ส่วนแรกเป็นขั้นตอนของการสร้างกลุ่มหัวข้อที่สำคัญ โดยการใช้การตัดคำแบบเอ็นแกรม จากนั้นนำคำที่ได้ทั้งหมดไปให้น้ำหนักและกำหนดช่วงของคำเพื่อสร้างเป็นกลุ่มหัวข้อ ส่วนที่สองคือการสกัดย่อหน้าที่มีใจความสำคัญออกจากเอกสารเพื่อสร้างบทย่อความ โดยอาศัยการเข้าคู่ด้วยตัวแบบเว็กเตอร์ผสมผสานกับการวัดค่าความคล้ายคลึงระหว่างกลุ่มคำที่เป็นหัวข้อและเอกสารในแต่ละย่อหน้า ซึ่งการย่อข้อความจะทำการย่อความในระดับ 20% และ 40% ของเอกสาร และค่าความถูกต้องอยู่ระหว่าง 28-43% สำหรับการย่อความที่ 20% และค่าความถูกต้องอยู่ระหว่าง 42-63% สำหรับการย่อความที่ 40% ซึ่งเป็นการวัดด้วยค่าเอฟ

คำสำคัญ: การย่อข้อความ, เอ็นแกรม, ตัวแบบเว็กเตอร์

.: หลักการและเหตุผล :.

ระบบย่อความ คือ กระบวนการของการสรุปเนื้อหาสาระที่สำคัญจากเอกสารต้นฉบับ เพื่อให้ได้เนื้อหาที่กระชับและสั้นลง โดยระบบการย่อความนั้นจะขึ้นอยู่กับความต้องการของผู้ใช้งานและงานที่นำไปประยุกต์ใช้ โดยทั่วไปกระบวนการของการย่อความ จะเริ่มต้นจากอ่านเอกสารที่ต้องการย่อความเข้ามาในระบบ จากนั้นจึงผ่านกระบวนการประมวลผล (Processing) ซึ่งอาจจะเป็นวิธีการ *ประมวลผลทางภาษาธรรมชาติ* (*Natural Language Processing* : *NLP*) การประมวลผลเชิงสถิติ (Statistic) เป็นต้น และเมื่อผ่านกระบวนการดังกล่าวแล้ว จะรวบรวมประโยคต่างๆ ที่ได้มาสร้างบทย่อความและใช้วิธีการที่เรียกว่า *"การวัดความคล้ายคลึง"* (*Similarity*) เพื่อทำให้บทย่อความนั้นมีความสละสลวยมากยิ่งขึ้น

.: วัตถุประสงค์และขอบเขต :.

งานวิจัยฉบับนี้ได้จัดทำขึ้นเพื่อนำเสนอระบบการย่อความภาษาไทยโดยใช้วิธีการ N-gram ในการตัดคำเพื่อใช้ในการสร้างคำขอ (Query) ผสมผสานกับการใช้วิธีการหาความถี่ของคำขอ เพื่อหากลุ่มหัวข้อ และการหาค่าความคล้ายคลึง (*Similarity*) หรือ วิธีการทางด้านสถิติ หรือ ตัวแบบเวคเตอร์ (Vector Model) ของประโยคสำหรับการสร้างบทย่อความและเป็นระบบที่สามารถย่อความจากเอกสารที่มีลักษณะเป็นข้อความเท่านั้น ไม่รวมบทกลอน รูปภาพ สูตร หรือสมการใด ๆ

.: ทฤษฎีที่เกี่ยวข้อง :.

1. ทฤษฎี N-gram

วิธี N-gram คือ การตัดบางส่วนของข้อความนั้นออกมาเป็นข้อความที่มีขนาดเท่ากับ N ซึ่งใช้แทนการตัดคำ

2. การให้น้ำหนักคำ (Term Word Weighting)

$$ISF = 1 + log(|S| / SF)$$

$$TF - ISF = TF \times ISF$$

3. The Mixed Min and Max Model (The MMM Model)

ทำการหาค่า Minimum ของช่วง Topic Identification

$$Sim(Q_{and}D) = C_{and1} * \text{Max}(d_{A1}, d_{A2}, \ldots, d_{An}) + C_{and2} * \text{Min}(d_{A1}, d_{A2}, \ldots, d_{An})$$

4. Vector Model

$$\sigma(D_i, Q) = \frac{(D_i)(Q)}{|D_i||Q|}$$

5. การวัดความคล้ายคลึง (Similarity Measure)

$$Sim(t, S_k) = \sum_{w_i \in S_k, t} tf(w_i, t).tf(w_i, S_k).\left[1 - \frac{\log(sf(w_i)+1)}{\log(n+1)}\right]^2$$

6. การประเมินประสิทธิภาพ (*Quality Evaluation*)

$$Precision\ (\text{ความแม่นยำ}) = J / K$$

$$Recall\ (\text{ความระลึก}) = J / M$$

$$F = (2*R*P) / (R+P)$$