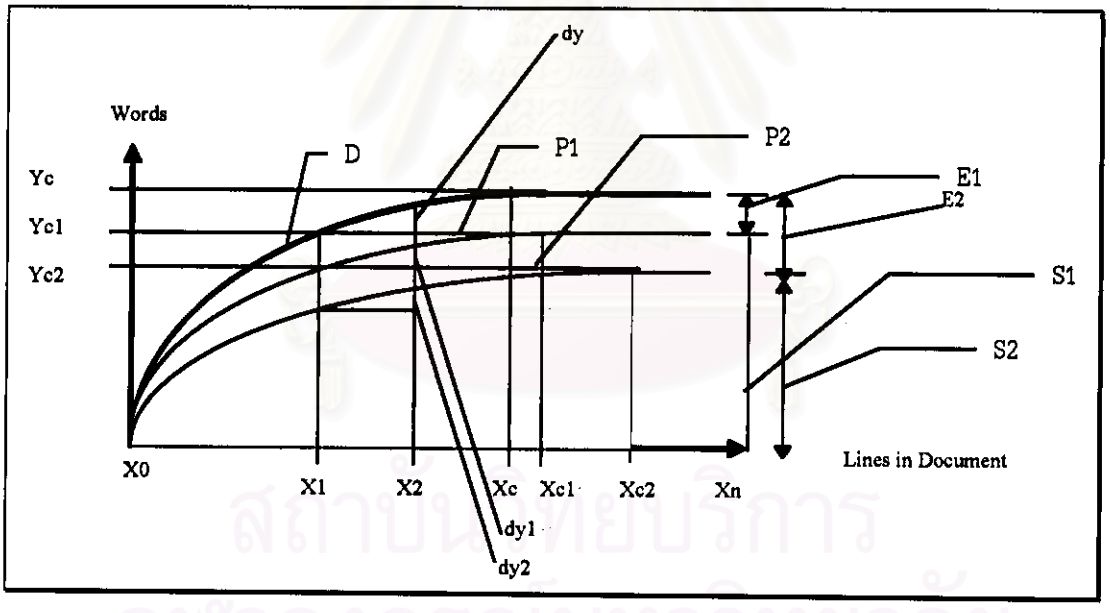
บทที่ 4

ผลกระทบของความยาวของเอกสารต่อประสิทธิภาพ

ปริมาณความยาวของเอกสารภาษาไทยที่จะนำใช้เป็นอินพุทสำหรับการวัดประสิทธิภาพเป็นประเด็นปัญหาที่จำเป็นจะต้องศึกษาเพื่อให้ได้ผลการทดลองเปรียบเทียบที่ถูกต้องครอบคลุมขอบเขตที่สมบูรณ์

4.1 คุณลักษณะเฉพาะที่น่าสนใจของเอกสารภาษาไทย

งานวิจัยนี้ได้ทำการศึกษาคุณลักษณะเฉพาะของเอกสารภาษาไทยทั่วไปพบว่า ถ้านับหน่วยคำที่ไม่ซ้ำกันที่ประกอบขึ้นเป็นเอกสารหนึ่งเล่มหรือหนึ่งเรื่องใด ๆจะมีจำนวนเพิ่มขึ้นเป็นสัดส่วนแบบลอการิธึมเมื่อเทียบกับปริมาณหน่วยคำทั้งหมดในเอกสารนั้นดังรูป 4.1



รูปที่ 4.1 ความสัมพันธ์ระหว่างจำนวนหน่วยคำที่ไม่ซ้ำกันกับขนาดของเอกสารตั้งแต่บรรทัดแรก (Xo) จนถึงบรรทัดสุดท้าย (Xn)

ตำแหน่ง	คำอธิบาย
Xc	จุดคอนเวอร์จ

Yc	จำนวนคำที่จุดคอนเวอร์จ
dy	ความชันของคำ
E	อัตราการผิดพลาด
S	ความสามารถตัดคำได้

ตาราง 4.1 มาตรวัดประสิทธิภาพของโปรแกรมตัดคำภาษาไทย

จากรูปที่ 4.1 เส้นกราฟ D แสดงคุณลักษณะเฉพาะของเอกสารภาษาไทยใด ๆ จะพบว่าที่ตำแหน่ง Xc ในเอกสารนี้ เป็นต้นไปจนจบเอกสาร ปริมาณหน่วยคำภาษาไทยใหม่ ๆ จะเพิ่มขึ้นมาน้อยมาก ในงานวิจัยนี้จะเรียกจุดนี้ว่า จุดคอนเวอร์จ (Convergence Point) ซึ่งจะมีจำนวนหน่วยคำเท่ากับ Yc ค่า dy จะเป็นความชันของปริมาณหน่วยคำที่ตำแหน่งใด ๆ ก่อนจุดคอนเวอร์จของเอกสาร D ซึ่งความชันของเอกสารต่าง ๆ จะเกิดจากลักษณะการใช้คำและภาษาของผู้แต่งเอกสารแต่ละราย ยกตัวอย่างเช่นผู้แต่งเอกสารรายหนึ่งใช้คำศัพท์ใหม่ ๆ เพิ่มขึ้นเรื่อย ๆ จนกระทั่งถึง 80 เปอร์เซ็นต์ของเอกสารก็จะไม่ปรากฏคำศัพท์ใหม่เกิดขึ้นอีก ในขณะที่อีกรายอาจจะใช้คำศัพท์ใหม่ ๆ จนถึงแค่ 40 เปอร์เซ็นต์ก็จะไม่ปรากฏคำศัพท์ใหม่ ๆ เกิดขึ้นแล้วเป็นต้น

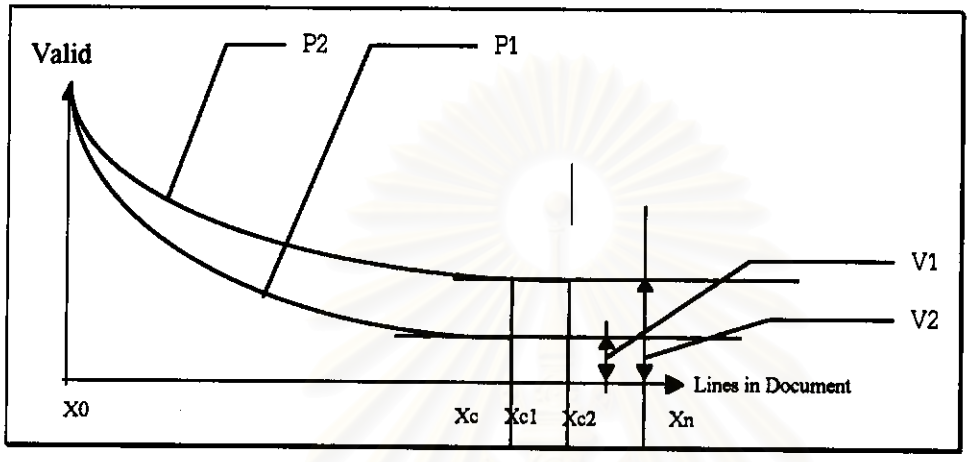
เส้นกราฟ P1 และ P2 แสดงพฤฒิกรรมของโปรแกรมตัดคำใด ๆ สองโปรแกรมเมื่อนำมาทำการตัดคำที่อยู่ในเอกสาร D ที่ตำแหน่ง Xc1 และ Xc2 จะแสดงให้เห็น จุดคอนเวอร์จของโปรแกรม P1 และ P2 ตามลำดับ Yc1 และ Yc2 เป็นจำนวนหน่วยคำที่จุดคอนเวอร์จของโปรแกรม P1 และ P2 ตามลำดับ dy1 และ dy2 จะแสดงให้เห็นความชันของหน่วยคำก่อนจุดคอนเวอร์จ ของโปรแกรม P1 และ P2 ตามลำดับ ความผิดพลาดที่เกิดจากการตัดคำ(E1, E2) ของโปรแกรมตัดคำ P1 และ P2 หรือในทางกลับกัน ความสามารถในการตัดคำ (S1, S2) ของโปรแกรมตัดคำ P1 และ P2 จะสังเกตได้จากค่า Xc, Yc และค่า dy ที่ได้จากโปรแกรมตัดคำนั้นๆ ตาราง 4.1 จะรวบรวมคุณลักษณะเฉพาะและคำอธิบายความหมายของเอกสารภาษาไทยทั่วๆไปซึ่งจะสามารถใช้เป็นมาตรวัดประสิทธิภาพของโปรแกรมตัดคำภาษาไทยได้เช่นกัน อาจจะสามารถเขียนเป็นสมการคณิตศาสตร์ได้ดังนี้

$$E=1-S=1-F(Xc, Yc, dy, X)$$

เมื่อ E : ความผิดพลาด
S : ความสามารถในการตัดคำ
Xc: จุดคอนเวอร์จ
Yc: จำนวนหน่วยคำที่จุดคอนเวอร์จ
dy: ความชันของคำ
X: จำนวนบรรทัดในเอกสาร

4.2 ความถูกต้องของโปรแกรมตัดคำ

ถ้าพิจารณาความถูกต้องที่เกิดขึ้นจากการใช้โปรแกรมตัดคำหลาย ๆ โปรแกรมทำการตัดคำลงบนเอกสารใด ๆจะพบว่าเกิดขึ้นไม่เท่ากัน ในปัจจุบันนี้ยังไม่มีโปรแกรมตัดคำใด ๆ ที่ตัดคำแล้วไม่ปรากฏความผิดพลาดเกิดขึ้น รูปที่ 4.2 แสดงความถูกต้องที่เกิดขึ้นสำหรับโปรแกรมตัดคำ P1 และ P2



รูป 4.2 ความถูกต้องที่เกิดขึ้นจากโปรแกรมตัดคำ

จากรูปจะเห็นว่าโปรแกรมตัดคำ P1 จะมีความถูกต้องเกิดขึ้นมาจากการตัดคำในเอกสาร D ใด ๆ น้อยกว่า โปรแกรม P2 เมื่อทำการตัดคำลงบนเอกสารเดียวกัน สาเหตุของความถูกต้องที่เกิดขึ้นจะเป็นผลมาจากความเป็นไปได้หลายประการ เช่น ความผิดพลาดในการสะกดคำของเอกสารที่นำมาใช้ทำการตัดคำ จำนวนคำศัพท์ที่อยู่ในพจนานุกรมของโปรแกรมตัดคำ วิธีการเลือกคำที่จะมาเป็นพจนานุกรม ขั้นตอนและวิธีการเปรียบเทียบคำศัพท์ในเอกสารกับพจนานุกรม วิธีการแก้ไขความผิดพลาดที่เกิดขึ้น เทคนิคการเลือกประโยคที่ถูกต้องหลังการตัดคำ ความกำกวมของประโยคภาษาไทยในเอกสารที่นำมาใช้ตัดคำ เป็นต้น ขอให้พิจารณาประโยคข้างล่างนี้ประกอบ

"กรรมการกรมพลศึกษารอยกร่าง"

จากประโยคนี้ถ้าใช้โปรแกรมตัดคำใด ๆ ทำการตัดคำจะมีความเป็นไปได้ของประโยคที่ได้จากการตัดคำถึง 34 ประโยคด้วยกันดังที่จะตามมาข้างล่าง

1 กร รม กา รก รม พล ศึกษา รอ ยก ร่าง
2 กร รม กา รก รม พล ศึกษา รอย กร่าง
3 กร รม กา รก รม พลศึกษา รอ ยก ร่าง
4 กร รม กา รก รม พลศึกษา รอย กร่าง

5 กร รม การ กรม พล ศึกษา รอ ยก ร่าง
6 กร รม การ กรม พล ศึกษา รอย กร่าง
7 กร รม การ กรม พลศึกษา รอ ยก ร่าง
8 กร รม การ กรม พลศึกษา รอย กร่าง
9 กร รม การ กรมพลศึกษา รอ ยก ร่าง
10 กร รม การ กรมพลศึกษา รอย กร่าง
11 กร รม การก รม พล ศึกษา รอ ยก ร่าง
12 กร รม การก รม พล ศึกษา รอย กร่าง
13 กร รม การก รม พลศึกษา รอ ยก ร่าง
14 กร รม การก รม พลศึกษา รอย กร่าง
15 กรรม กา รก รม พล ศึกษา รอ ยก ร่าง
16 กรรม กา รก รม พล ศึกษา รอย กร่าง
17 กรรม กา รก รม พลศึกษา รอ ยก ร่าง
18 กรรม กา รก รม พลศึกษา รอย กร่าง
19 กรรม การ กรม พล ศึกษา รอ ยก ร่าง
20 กรรม การ กรม พล ศึกษา รอย กร่าง
21 กรรม การ กรม พลศึกษา รอ ยก ร่าง
22 กรรม การ กรม พลศึกษา รอย กร่าง
23 กรรม การ กรมพลศึกษา รอ ยก ร่าง
24 กรรม การ กรมพลศึกษา รอย กร่าง
25 กรรม การก รม พล ศึกษา รอ ยก ร่าง
26 กรรม การก รม พล ศึกษา รอย กร่าง
27 กรรม การก รม พลศึกษา รอ ยก ร่าง
28 กรรม การก รม พลศึกษา รอย กร่าง
29 กรรมการ กรม พล ศึกษา รอ ยก ร่าง
30 กรรมการ กรม พล ศึกษา รอย กร่าง
31 กรรมการ กรม พลศึกษา รอ ยก ร่าง
32 กรรมการ กรม พลศึกษา รอย กร่าง
33 กรรมการ กรมพลศึกษา รอ ยก ร่าง
34 กรรมการ กรมพลศึกษา รอย กร่าง

จากคุณลักษณะของเอกสารภาษาไทยโดยทั่วไปทำให้เราสามารถได้แนวทางในการเลือกความยาวของเอกสารที่จะมาเป็นอินพุทของการวัดประสิทธิภาพว่าจะต้องครอบคลุมถึงจุดคอนเวอร์จของเอกสารนั้น ๆ จึงจะทำให้ผลการวัดเปรียบเทียบครอบคลุมขอบเขตที่สมบูรณ์