

การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย

นายฟิลิธี พรหมจันทร์



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2540

ISBN 974-638-133-4

ลิขสิทธิ์ของ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

**ANALYSIS OF GUIDELINES FOR PERFORMANCE COMPARISON OF
THAI WORD SEPARATION PROGRAMS**

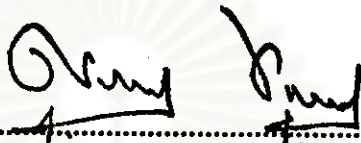
Mr. Pisit Promchan

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย


**A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
Graduate School
Chulalongkorn University
Academic Year 1997
ISBN 974-638-133-4**

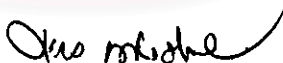
หัวข้อวิทยานิพนธ์ การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรม
แยกคำภาษาไทย
โดย นายพิสิทธิ์ พรหมจันทร์
ภาควิชา วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา อาจารย์ ดร. ชรรยง เต็งอำนวยการ

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้วิทยานิพนธ์ฉบับนี้เป็น
ส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาโทบัณฑิต

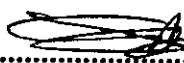

.....คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ นายแพทย์ศุภวัฒน์ ชุตินวงศ์)

คณะกรรมการสอบวิทยานิพนธ์


.....ประธานกรรมการ
(รองศาสตราจารย์สมชาย ทยานง)


.....อาจารย์ที่ปรึกษา
(อาจารย์ ดร.ชรรยง เต็งอำนวยการ)

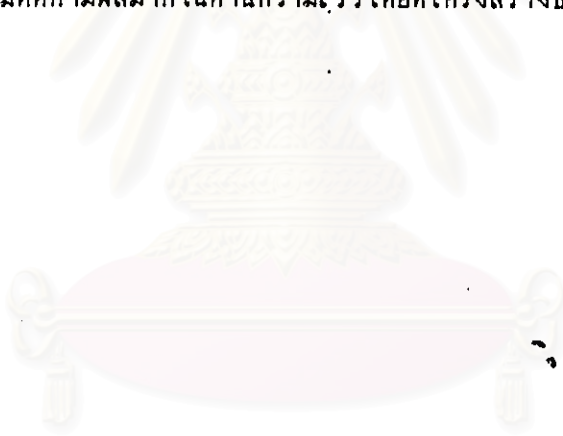

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สมชาย ประสิทธิ์จตุระกุล)


.....กรรมการ
(อาจารย์จารย์มาตร ปันทอง)

พิสิทธิ พรหมจันทร์: การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย (Analysis of Guidelines for Performance Comparison of Thai Word Separation Program) อ.ที่ปรึกษา: อ.ดร.ยรรยง เต็งอ้วนวย, 71 หน้า, ISBN 974-638-133-4

งานวิจัยนี้ได้ทำการวิเคราะห์ หาแนวทางในการเปรียบเทียบสมรรถนะของโปรแกรมและอัลกอริธึมตัดคำภาษาไทย โดยเริ่มจากการสังเคราะห์ตัวอย่างมาตรฐานที่จะใช้ในการวัดและเปรียบเทียบประสิทธิภาพ ศึกษาคุณลักษณะเฉพาะของเอกสารภาษาไทย ที่มีผลต่อประสิทธิภาพของโปรแกรมตัดคำภาษาไทย รวบรวมโปรแกรมและอัลกอริธึมตัดคำภาษาไทย ที่ได้มีการพัฒนาและเผยแพร่ใช้งานในปัจจุบัน รวบรวมข้อมูลภาษาไทยที่ใช้อ้างอิง รวมไปถึงพจนานุกรมที่ใช้ในการตรวจสอบความถูกต้องของการตัดคำ จากนั้นจึงทำการพัฒนาวิธีการวัดประสิทธิภาพ และทำการวัดประสิทธิภาพ

จากผลการวัดประสิทธิภาพพบว่าแบบเปรียบเทียบคำที่ยาวที่สุดจะตัดได้จำนวนคำที่ถูกต้องออกมามากที่สุด แบบการแก้ไขย้อนกลับจะได้คำผิดน้อยที่สุด แบบอาศัยความถี่ของการใช้คำจะได้อัตราความถูกต้องต่อจำนวนคำในพจนานุกรมสูงสุด แบบใช้พจนานุกรมลดความกำกวมสามารถจัดการกับคำกำกวมได้ดีที่สุด และแบบเปรียบเทียบคำที่สั้นที่สุดจะตัดออกมาได้จำนวนคำสูงสุด นอกจากนี้พบว่าโครงสร้างข้อมูลสำหรับพจนานุกรมที่ใช้ในโปรแกรมตัดคำมีผลมากในด้านความเร็ว โดยที่โครงสร้างข้อมูลแบบหรัสให้ความเร็วสูงสุดในปัจจุบัน



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิทยาการสารสนเทศคอมพิวเตอร์
ปีการศึกษา 2540

ลายมือชื่อนิสิต
ลายมือชื่ออาจารย์ที่ปรึกษา
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

C718612 : MAJOR Computer Science
KEY WORD: Algorithm/Analysis/Performance/Comparison/Thai/Word/Separation/
Segmentation

Pisit Eromchan : Analysis of Guidelines for Performance Comparison of Thai Word Separation Programs. Thesis Advisor : Yunyong Teng-Amnuay, PH.D., 71 pp. ISBN 974-638-133-4.

In this thesis, the guidelines for performance comparison of Thai Words Separation Programs have been analyzed. The thesis begin from synthesis of example of performance indicators, study the characteristics of Thai documents that effect performance of the Thai Words Separation Programs. Then, collect Thai Words Separation Programs and algorithms that had been developed and announced to be used currently, collect the Thai reference data which include the reference dictionary to validate the accuracy of Thai Words separation, and develop the measurement methodology. Finally, I do the performance measurement using the developed methodology.

Experimental results show that the Longest Pattern Matching gives the most accurate words output while the Back Tracking Algorithm gives the least error words. Words Usage Frequency gives the highest valid words ratio per number of words in its dictionary. The usage of ambiguity dictionary gives the best ambiguous case resolution, whereas the Shortest Pattern Matching gives the highest number of words output. Additionally, it is found that the data structure for dictionary that used in Thai Words Separation Programs extremely effects in term of speed, meanwhile the Trie structure is the most popular method that has been used in the present due to its outperform speed.

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา.....วิศวกรรมคอมพิวเตอร์.....

สาขาวิชา.....วิทยาศาสตร์คอมพิวเตอร์.....

ปีการศึกษา..... 2540.....

ลายมือชื่อนิสิต..... .....

ลายมือชื่ออาจารย์ที่ปรึกษา..... .....

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ ดร. ยรรยง เต็งอำนาจ ที่ได้ให้คำปรึกษาแนะนำแนวทาง ปรับปรุงแก้ไขและขัดเกลาจนทำให้วิทยานิพนธ์ฉบับนี้ประสบความสำเร็จลุล่วงจนได้ฉบับที่สมบูรณ์นี้ ผู้วิจัยขอขอบคุณ อ.วันทนี พันธชาติ และ ดร.สุรพันธ์ เมฆนาวิน ที่ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้ (Linguistics and Knowledge Science Laboratory: LINKS) ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (National Electronics and Computer Technology Center: NECTEC) ที่ได้ให้คำแนะนำที่เป็นประโยชน์มาก รวมทั้งอนุเคราะห์ฐานข้อมูลอ้างอิงภาษาไทยที่ผู้วิจัยได้เลือกใช้บางส่วนสำหรับงานวิจัยนี้ และได้ให้โปรแกรมตัดคำภาษาไทยที่พัฒนาที่ห้องวิจัยนี้มาร่วมวัดประสิทธิภาพ

ผู้วิจัยขอขอบคุณครอบครัวของผู้วิจัยเอง ที่มีส่วนสนับสนุนด้วยดียิ่งตลอดมา งานวิจัยนี้สำเร็จลุล่วงอย่างสมบูรณ์



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

หน้า

| | |
|---|----|
| บทคัดย่อภาษาไทย | จ |
| บทคัดย่อภาษาอังกฤษ..... | จ |
| กิตติกรรมประกาศ..... | ฉ |
| สารบัญ..... | ช |
| | |
| บทที่ 1 บทนำ | 1 |
| 1.1 ความเป็นมาของปัญหา..... | 1 |
| 1.2 วัตถุประสงค์ของวิทยานิพนธ์ | 2 |
| 1.3 ขอบเขตของวิทยานิพนธ์ | 3 |
| 1.4 ขั้นตอนการวิจัย | 3 |
| 1.5 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย | 4 |
| | |
| บทที่ 2 แนวคิดและทฤษฎีที่เกี่ยวข้อง | 5 |
| 2.1 การแยกคำหรือการตัดคำ..... | 5 |
| 2.2 เครื่องหมายและการแบ่งวรรคตอนภาษาไทยที่ถูกต้อง | 5 |
| 2.3 การพัฒนาและขั้นตอนวิธีการตัดคำภาษาไทยแบบต่าง ๆ | 8 |
| 2.4 การวิเคราะห์เลือกประโยคที่ต้องการตัดคำ..... | 14 |
| | |
| บทที่ 3 ตัวอย่างมาตรวัดประสิทธิภาพของขั้นตอนวิธีการตัดคำแบบต่าง ๆ | 16 |
| 3.1 ความสามารถที่จะตัดคำได้ | 16 |
| 3.2 ความถูกต้องของคำหลังจากผ่านการตัดแล้ว | 16 |
| 3.3 สัดส่วนความถูกต้องของคำที่ตัดออกมาได้ต่อจำนวนคำที่ใช้เป็นพจนานุกรม..... | 16 |
| 3.4 ความถูกต้องเชิงไวยากรณ์ของประโยคหลังจากการตัดแล้ว | 17 |
| 3.5 ความถูกต้องเชิงความหมายของประโยคหลังจากตัดแล้ว..... | 17 |
| 3.6 ความสามารถ ที่จะรู้จักเครื่องหมายแบ่งวรรคตอนภาษาไทย..... | 17 |
| 3.7 ความสามารถที่จะปรับสระและวรรณยุกต์ที่ติดกันอย่างไม่ถูกต้อง..... | 17 |
| 3.8 มาตรวัดประสิทธิภาพเชิงความเร็ว | 18 |
| 3.9 มาตรวัดประสิทธิภาพการใช้ทรัพยากร | 18 |
| 3.10 มาตรวัดประสิทธิภาพการแก้ไขความผิดพลาด..... | 18 |

| | |
|--|----|
| บทที่ 4. ความยาวของเอกสารภาษาไทยที่มีผลต่อการเปรียบเทียบประสิทธิภาพ..... | 20 |
| 4.1 คุณลักษณะเฉพาะที่น่าสนใจของเอกสารภาษาไทย..... | 20 |
| 4.2 ความถูกต้องของโปรแกรมตัดคำ..... | 21 |
| บทที่ 5 โปรแกรมและอัลกอริธึมตัดคำภาษาไทย..... | 24 |
| 5.1 โปรแกรมตัดคำภาษาไทยแบบย้อนรอยกลับ..... | 24 |
| 5.2 โปรแกรมตัดคำภาษาไทยของไมโครซอฟท์วินโดวส์ 95..... | 24 |
| 5.3 โปรแกรมตัดคำภาษาไทยแบบการเทียบคำที่ยาวที่สุด..... | 25 |
| 5.4 โปรแกรมตัดคำภาษาไทยแบบการเทียบคำที่สั้นที่สุด..... | 25 |
| 5.5 โปรแกรมตัดคำภาษาไทยแบบที่ใช้ความถี่ของการใช้คำ..... | 25 |
| 5.6 โปรแกรมตัดคำภาษาไทยแบบที่ใช้พจนานุกรมลดความกำกวม..... | 26 |
| 5.7 โปรแกรมตัดคำภาษาไทยแบบที่เลือกประโยคที่มีจำนวนคำน้อยที่สุด..... | 27 |
| บทที่ 6 ข้อมูลภาษาไทยที่ใช้อ้างอิงในการวัดประสิทธิภาพ..... | 28 |
| 6.1 ฐานข้อมูลภาษาไทย..... | 28 |
| 6.2 พจนานุกรมอ้างอิง..... | 31 |
| บทที่ 7 ขั้นตอนในการวัดประสิทธิภาพ..... | 33 |
| 7.1 ทำการตัดคำภาษาไทย..... | 33 |
| 7.2 ปรับเปลี่ยนรูปแบบผลลัพธ์..... | 33 |
| 7.3 เรียงลำดับผลลัพธ์..... | 33 |
| 7.4 ลบคำที่ซ้ำกันออกไป..... | 33 |
| 7.5 ตรวจสอบความถูกต้องและรวบรวมข้อมูลสถิติ..... | 33 |
| 7.6 วัดความผิดพลาดที่เกิดจากความกำกวมของภาษาไทย..... | 34 |
| บทที่ 8 ผลการทดลองเปรียบเทียบประสิทธิภาพของโปรแกรมตัดคำภาษาไทย..... | 36 |
| 8.1 จำนวนคำที่ตัดออกมาถูกต้องและจำนวนคำที่ตัดออกมา | 36 |
| 8.2 สัดส่วนความถูกต้องต่อจำนวนคำที่ตัดออกมา..... | 37 |
| 8.3 เปรียบเทียบจำนวนคำภาษาไทยที่ตัดออกมาได้ถูกต้อง | 38 |

| | | |
|---------|---|----|
| 8.4 | เปรียบเทียบจำนวนคำภาษาไทยที่ตัดออกมาได้ทั้งหมด..... | 39 |
| 8.5 | ความถูกต้องโดยใช้ผลของโปรแกรมแบบเทียบคำที่ยาวที่สุดเป็นฐานอ้างอิง..... | 40 |
| 8.6 | ความถูกต้องโดยใช้ผลของโปรแกรมแบบเทียบคำที่สั้นที่สุดเป็นฐานอ้างอิง..... | 41 |
| 8.7 | สัดส่วนความถูกต้องต่อจำนวนคำในพจนานุกรมที่ใช้งาน..... | 42 |
| 8.8 | เปรียบเทียบความผิดพลาดที่เกิดขึ้นของแต่ละโปรแกรม..... | 44 |
| 8.9 | เปรียบเทียบการใช้งานทรัพยากร..... | 45 |
| 8.10 | เปรียบเทียบความสามารถในการแก้ปัญหาค่ากำกวม..... | 46 |
| 8.11 | ผลการทดลองที่ได้จากฐานข้อมูลภาษาไทยดาด้าแบงค์..... | 47 |
| 8.12 | การวิเคราะห์ผลการทดลอง..... | 51 |
| 8.13 | ความเหมาะสมของการประยุกต์ใช้โปรแกรมตัดคำ..... | 52 |
| บทที่ 9 | บทสรุปและข้อเสนอแนะ..... | 53 |
| 9.1 | ผลของความยาวของเอกสารต่อประสิทธิภาพของโปรแกรมตัดคำ..... | 53 |
| 9.2 | เอกสารประเภทต่าง ๆ กับผลการเปรียบเทียบประสิทธิภาพการตัดคำ..... | 53 |
| 9.3 | สรุปประสิทธิภาพของโปรแกรมตัดคำ..... | 53 |
| 9.4 | ปัญหาต่าง ๆ ที่พบในงานวิจัยนี้..... | 54 |
| 9.5 | ข้อเสนอแนะและงานวิจัยที่สามารถทำเพิ่มเติมได้..... | 55 |
| | รายการอ้างอิง..... | ซ |
| | ประวัติผู้ทำวิจัย..... | ณ |