การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย

นายพิสิทธิ์ พรมจันทร์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตรคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลากรณ์มหาวิทยาลัย
ปีการศึกษา 2540
ISBN 974-638-133-4

ANALYSIS OF GUIDELINES FOR PERFORMANCE COMPARISON OF THAI WORD SEPARATION PROGRAMS

Mr. Pisit Promchan

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	การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย
โดย	นายพิสิทธิ์ พรมจันทร์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร. ยรรยง เต็งอำนวย

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้วิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ นายแพทย์ศุภวัฒน์ ชุติวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาตราจารย์สมชาย ทยานยง)

.....อาจารย์ที่ปรึกษา
(อาจารย์ ดร.ยรรยง เต็งอำนวย)

.....กรรมการ
(ผู้ช่วยศาตราจารย์ ดร.สมชาย ประสิทธ์จูตระกูล)

.....กรรมการ
(อาจารย์จารุมาตร ปิ่นทอง)

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสีเขียวนี้เพียงแผ่นเดียว

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

งานวิจัยนี้ได้ทำการวิเคราะห์ หาแนวทางในการเปรียบเทียบสมรรถนะของโปรแกรมและอัลกอริธึมตัดคำภาษาไทย โดยเริ่มจากการสังเคราะห์ตัวอย่างมาตรวัด ที่จะใช้ในการวัดและเปรียบเทียบประสิทธิภาพ ศึกษาคุณลักษณะเฉพาะของเอกสารภาษาไทย ที่มีผลต่อประสิทธิภาพของโปรแกรมตัดคำภาษาไทย รวบรวมโปรแกรมและอัลกอริธึมตัดคำภาษาไทย ที่ได้มีการพัฒนาและเผยแพร่ใช้งานในปัจจุบัน รวบรวมข้อมูลภาษาไทยที่ใช้อ้างอิง รวมไปถึงพจนานุกรมที่ใช้ในการตรวจสอบความถูกต้องของการตัดคำ จากนั้นจึงทำการพัฒนาวิธีการวัดประสิทธิภาพ และทำการวัดประสิทธิภาพ

จากผลการวัดประสิทธิภาพพบว่าแบบเปรียบเทียบคำที่ยาวที่สุดจะตัดได้จำนวนคำที่ถูกต้องออกมามากที่สุด แบบการแก้ไขย้อนกลับจะได้คำผิดน้อยที่สุด แบบอาศัยความถี่ของการใช้คำจะได้อัตราความถูกต้องต่อจำนวนคำในพจนานุกรมสูงสุด แบบใช้พจนานุกรมลดความกำกวมสามารถจัดการกับคำกำกวมได้ดีที่สุด และแบบเปรียบเทียบคำที่สั้นที่สุดจะตัดออกมาได้จำนวนคำสูงสุด นอกจากนี้พบว่าโครงสร้างข้อมูลสำหรับพจนานุกรมที่ใช้ในโปรแกรมตัดคำมีผลมากในด้านความเร็วโดยที่โครงสร้างข้อมูลแบบทรัยให้ความเร็วสูงสุดในในปัจจุบัน

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2540 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

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

Pisit Bromchan : 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.

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา

ปีการศึกษา 2540 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

กิตติกรรมประกาศ

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

ผู้วิจัยขอขอบคุณครอบครัวของผู้วิจัยเอง ที่มีส่วนสนับสนุนด้วยดียิ่งตลอดมา จนงานวิจัยนี้สำเร็จลุล่วงอย่างสมบูรณ์

สารบัญ

หน้า