รหัสโครงการ 34S001

Thai Word Segmentation a Hybrid Approach

การแบ่งคำภาษาไทยด้วยเทคนิคไฮบริด

ประเภท การแข่งขันสุดยอดซอฟต์แวร์แบ่งคำภาษาไทย

รายงานฉบับสมบูรณ์

เสนอต่อ

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ
สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ
กระทรวงวิทยาศาสตร์และเทคโนโลยี

ประจำปีประมาณ 2552

โดย
นายวรศักดิ์ ตั้งกุลทวีทรัพย์
นาย พีระศักดิ์ รัตนมณี
นายธนพล จินดาพิทักษ์

อาจารย์ที่ปรึกษาโครงการ
นาย สุธน แซ่ว่อง

สถาบันการศึกษา
ภาควิชาคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

บทคัดย่อ

เป็นที่ทราบกันทั่วไปว่าการประมวณผลการแบ่งคำในภาษาต่างๆเป็นเทคโนโลยีพื้นฐานที่จะนำไปสู่ระบบประยุกต์ที่มีคุณค่าเป็นอย่างมาก เช่น การแปลภาษาอัตโนมัติ การรู้จำเสียงและสังเคราะห์เสียงพูด การย่อความอัตโนมัติ การพัฒนาหุ่นยนต์ เป็นต้น การประมวณผลการแบ่งคำในภาษาต่างๆก็จะมีความยากง่ายแตกต่างกันออกไปโดยเฉพาะภาษาไทย ซึ่งเป็นภาษาที่ประมวลผลได้ยากมาก และยังมีความล่าช้าในการประมวลผลอีกด้วย

เนื่องจากภาษาไทยเป็นภาษาที่ไม่มีการเขียนแบ่งพยางค์ คำ กลุ่มคำ หรือประโยค ไม่มีขอบเขตที่ชัดเจนของคำ ไม่มีหลักเกณฑ์ตายตัวในการใช้ช่องว่ างในภาษาเขียน ไม่มีเครื่องหมายที่ใช้ในการเว้นวรรค ไม่ใช้อักขระพิเศษเพื่อแสดงการขึ้นประโยคใหม่หรือแสดงชื่อเฉพาะ มีรูปแบบการสะกดที่ซับซ้อน และมีคำยืมจำนวนมาก ทำให้การแบ่งคำที่เป็นคำกำกวมทำได้ยาก และยังมีอีกหลายรูปแบบที่เป็นอุปสรรคต่อการพัฒนาระบบประมวลผลภาษาไทย

ในปัจจุบันมีงานวิจัยเพื่อแก้ปัญหาในการประมวลผลการแบ่งคำภาษาไทยโดยใช้หลักการต่างๆเข้ามาช่วยในการประมวลผล เช่นหลักการสร้างพยางค์ การใช้พจนานุกรม การใช้เทคนิคการเรียนรู้ด้วยเครื่อง การแบ่งตามคำศัพท์ เป็นต้นแต่ก็ยังไม่มีวิธีการใดที่เหมาะสมที่สุดสำหรับการแบ่งคำภาษาไทย

คำนำ

ความรู้ความสามารถทางภาษาเป็นสิ่งสำคัญสำหรับการเริ่มต้นการศึกษาหาความรู้ทางด้านต่างๆ ซึ่งคนไทยส่วนใหญ่จะมีความรู้ทางด้านภาษาไทยเป็นพื้นฐานอยู่แล้ว แต่ก็ยังมีบ้างส่วนที่ไม่เข้าใจในพื้นฐานทางด้านภาษา หรือเข้าใจผิดในเรื่องการใช้ภาษา เช่น กลุ่มเด็กที่ใช้สื่อการเรียนรู้อย่างอินเตอร์เน็ต หรือการใช้งานซอฟแวร์ที่ไม่สามารถแบ่งคำภาษาไทยได้อย่างถูกต้อง ทำให้เป็นผลในการอ่านภาษาไทยแบบผิดๆ

ด้วยเหตุผลนี้ ทางผู้จัดทำจึงได้เข้าร่วมการแข่งขันสุดยอดซอฟตีแวร์แบ่งคำภาษาไทย เพื่อเป็นแนวทางหนึ่งในการพัฒนาซอฟแวร์สำหรับการแบ่งคำภาษาไทย

สารบัญ

บทคัดย่อ i

คำนำ ii

สารบัญ iii

บทที่ 1 บทนำ 1

1.1 ที่มาของโครงงาน 1

1.2 วัตถุประสงค์และเป้าหมาย 1

บทที่ 2 รายละเอียดของการพัฒนา 2

2.1 ทฤษฎีหลักการและเทคนิคที่ใช้ 2

2.2 เครื่องมือที่ใช้ในการพัฒนา 3

NetBeans 3

2.3 รายละเอียดโปรแกรมที่ได้พัฒนาในเชิงเทคนิค (Software Specification) 3

2.3.1 Input/output Specification 3

2.3.2 Functional Specification 3

บทที่ 3 ผลการทดสอบโปรแกรม 4

3.1 ผลการทดสอบโปรแกรม 4

3.2 ตัวอย่างการทดสอบโปรแกรม 4

บทที่ 4 ปัญหาและอุปสรรค 5

4.1 ปัญหาและอุปสรรค 5

บทที่ 5 ภาคผนวก 6

5.1 การใช้งานโปรแกรม 6

เอกสารอ้างอิง 10

บทที่ 1 *บทนำ*

1.1 *ที่มาของโครงงาน*

แม้ว่าในปัจจุบัน ได้มีซอฟแวร์ที่ใช้ในการแบ่งคำภาษาไทยอยู่หลากหลายโปรแกรม หรืองานวิจัยในการแบ่งคำภาษาไทย แต่ก็ยังคงมีปัญหาอยู่ในปัจจุบัน เนื่องจากภาษาไทยเป็นภาษาที่ไม่ได้มีการเว้นวรรคเมื่อสิ้นสุดคำ ทำให้การแบ่งคำด้วยซอฟแวร์พัฒนาให้มีความสามารถในการแบ่งคำได้ไม่ครบหนึ่งร้อยเปอร์เซ็นต์ ทำให้ยังไม่มีอัลกอริธึมใดที่เหมาะสมที่สุด

1.2 *วัตถุประสงค์และเป้าหมาย*

เพื่อปรับปรุงการแบ่งคำภาษาไทยให้สามารถแบ่งคำ ตามเทคนิคที่ได้กำหนดไว้ได้

บทที่ 2 *รายละเอียดของการพัฒนา*

2.1 *ทฤษฎีหลักการและเทคนิคที่ใช้*

- การประมวลผลการแบ่งคำไทยจะใช้เทคนิค การแบ่งคำภาษาไทยด้วยดิกชันนารี โดยการเลือกคำใดๆนั้นจะเลือกจากคำที่มีขนาดยาวที่สุดก่อน

ประโยคตัวอย่าง	คำที่ถูกเลือก(ตรวจสอบจากพจนานุกรม)
เรือโคลงเพราะโคลงเรือ	ย้อนกลับ
เรือโคลงเพราะโคลงเรี	ย้อนกลับ
เรือโคลงเพราะโคลง	ย้อนกลับ
เรือโคลงเพราะโค	ย้อนกลับ...
เรือ	เรือ
โคลงเพราะโคลงเรือ	ย้อนกลับ
โคลงเพราะโคลงเ	ย้อนกลับ
โคลงเพราะโค	ย้อนกลับ...
โคลง	โคลง
เพราะโคลงเรือ	ย้อนกลับ
เพราะโคลงเ	ย้อนกลับ...
เพราะ	เพราะ
โคลงเรือ	ย้อนกลับ...
โคลง	โคลง
เรือ	เรือ

- การเลือกคำด้วยหลักภาษาไทย เช่นกฎของคำที่ขึ้นต้นด้วยการ, ความ เป็นต้น

2.2 *เครื่องมือที่ใช้ในการพัฒนา*

NetBeans

NetBeans เป็นเครื่องมือที่ใช้ในการพัฒนาโปรแกรมในครั้งนี้ โดยการเขียน Code ส่วนใหญ่จะเขียนด้วยเครื่องมือชนิดนี้ ในตัว NetBeans นี้ยังมีเครื่องมือที่อำนวยความสะดวกในการทำงานอีกด้วย เช่น การแสดงเลขบรรทัดในการเขียนCode การค้นหาข้อความและการแทนที่ข้อความ

นอกจากนี้ NetBeans ยังสามารถทำการเพิ่ม Library ที่ต้องการใช้ในการพัฒนาโปรแกรมได้อีกด้วย ซึ่งทำให้ง่ายต่อการพัฒนาโปรแกรม

2.3 *รายละเอียดโปรแกรมที่ได้พัฒนาในเชิงเทคนิค (Software Specification)*

2.3.1 Input/output Specification

Input ของโปรแกรม:

- ไฟล์ข้อความ (.txt) เพื่อใช้ในการประมวลผลการแบ่งคำ

Output ของโปรแกรม:

- ไฟล์ข้อความที่มีการแบ่งคำเรียบร้อยแล้วโดยใช้เครื่องหมาย| ในการแบ่งคำ ที่เป็นไปตามกฎเกณฑ์ที่วางไว้

2.3.2 Functional Specification

- โปรแกรมมีส่วนของการ ผลการทดสอบโปรแกรม

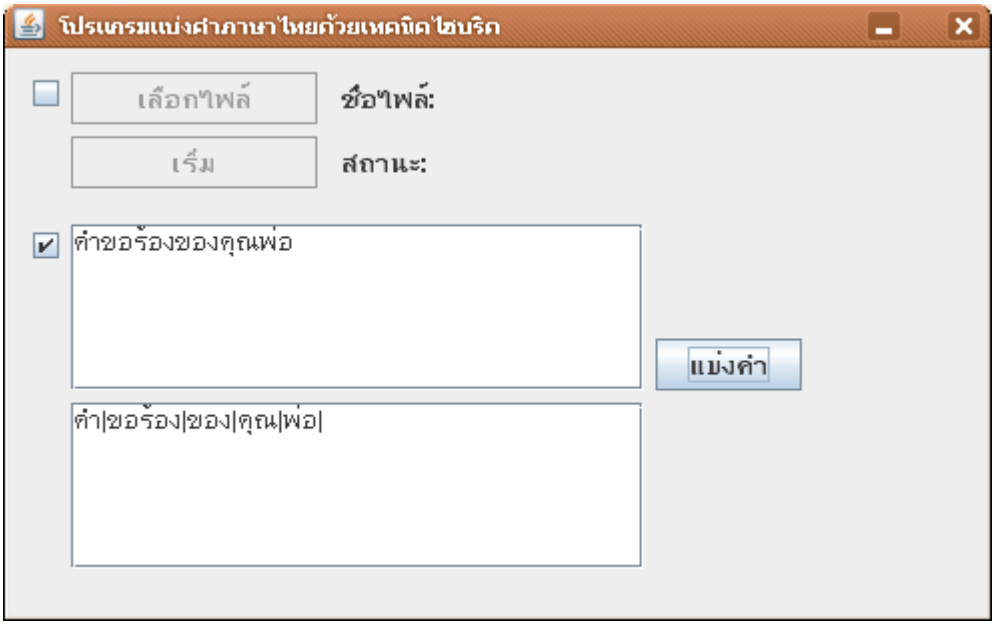
ในบทนี้จะเป็นการกล่าวถึงการใช้งานของซอฟแวร์ที่พัฒนา และผลลัพธ์ที่ได้จากของซอฟแวร์

บทที่ 3 *ผลการทดสอบโปรแกรม*

3.1 *ผลการทดสอบโปรแกรม*

โปรแกรมสามารถตัดคำตามเทคนิคที่ใช้ได้

3.2 *ตัวอย่างการทดสอบโปรแกรม*



รูป ทดสอบโปรแกรมด้วยคำว่า “คำขอร้องของคุณพ่อ”

บทที่ 4 *ปัญหาและอุปสรรค*

4.1 *ปัญหาและอุปสรรค*

- ไม่สามารถแยกแยะคำที่นำมาจากภาษาต่างประเทศได้ เนื่องจากไม่มีคลังข้อมูลของคำภาษาต่างประเทศ

- ไม่สามารถแยกแยะชื่อที่มีนามสกุลมีความหมายและต่อเนื่องกับประโยคที่ตามมาได้นื่อง เช่น นายชาติชาย หวังดีต่อนายสุพรชัย อาจตัดได้ว่า หวังดี| หรือ หวัง|ดี|

- ไม่สามารถแยกแยะคำที่เขียนเหมือนกันแต่มีความหมายแตกต่างกันเมื่ออยู่ในประโยคต่างกันที่ซับซ้อนได้

- ประโยคบางประโยค ไม่สามารถทราบได้ว่าตัดถูกหรือไม่ เนื่องจากไม่มีที่ให้ตรวจสอบ

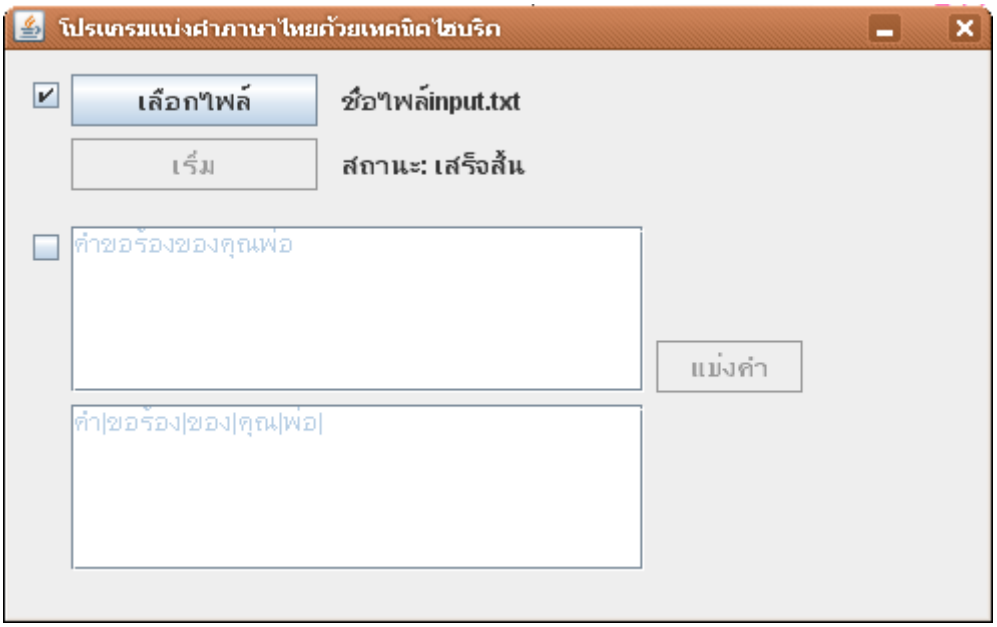
บทที่ 5 *ภาคผนวก*

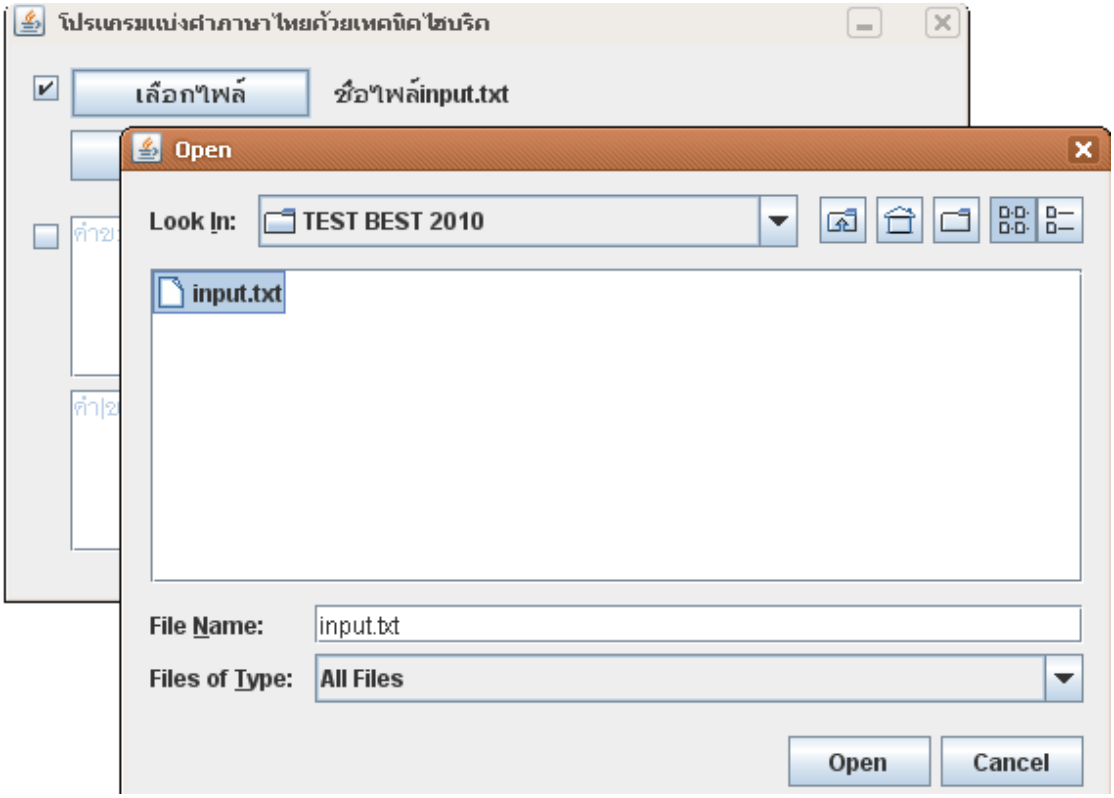
5.1 *การใช้งานโปรแกรม*

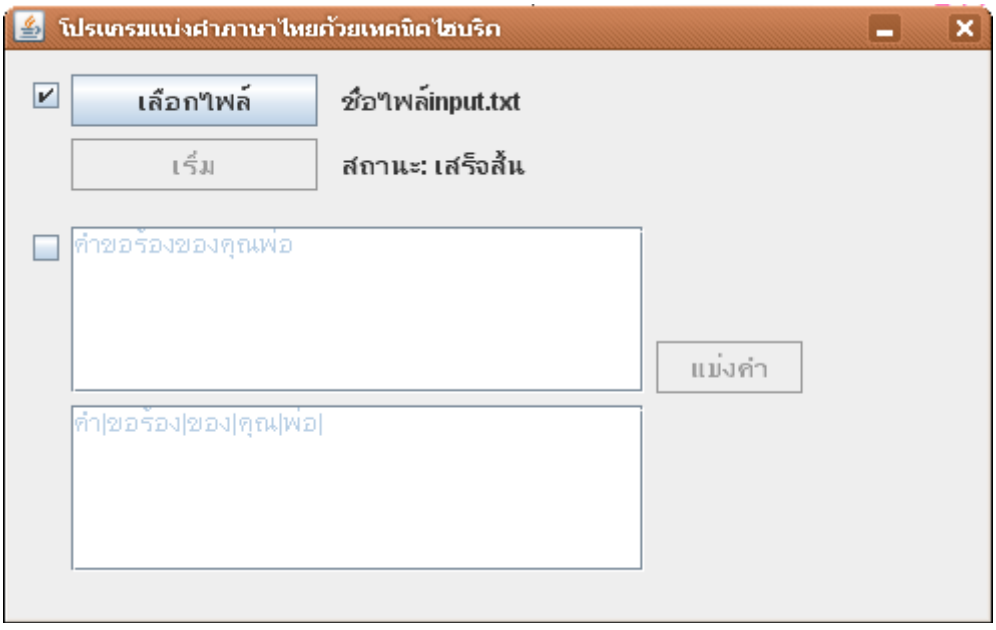
โปรแกรมมีสองส่วนให้เลือกใช้งาน

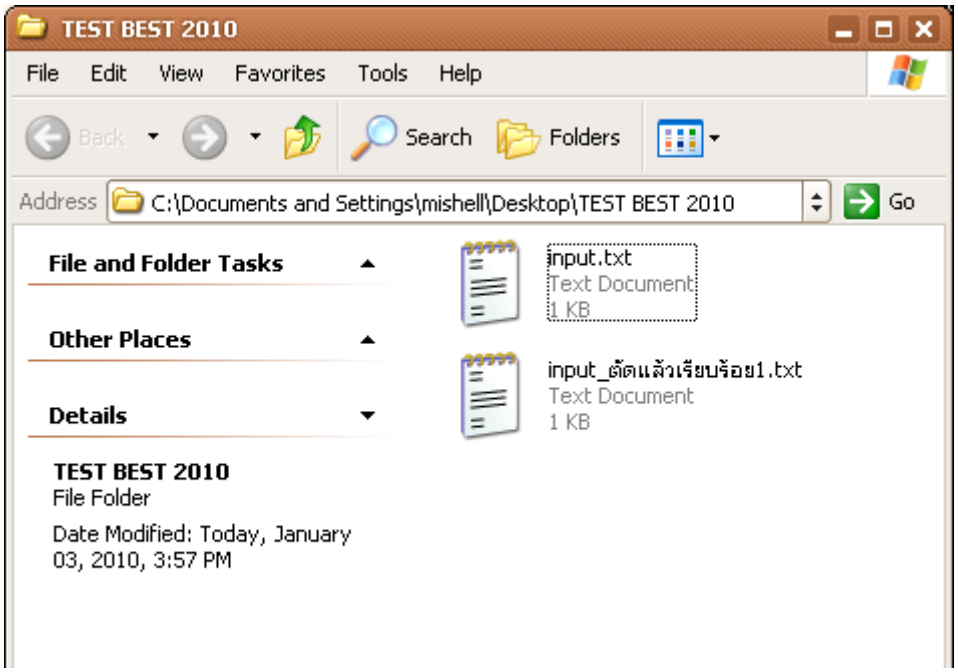
ส่วนแรก เป็นส่วนที่ใช้ไฟล์ เพื่อแปลง และผลลัพธ์ก็จะสร้างเป็นไฟล์ออกมาเป็นชื่อไฟล์เดิม ตามด้วยคำว่า ตัดเสร็จเรียบร้อย และอยู่ที่เดิมกับไฟล์อินพุต

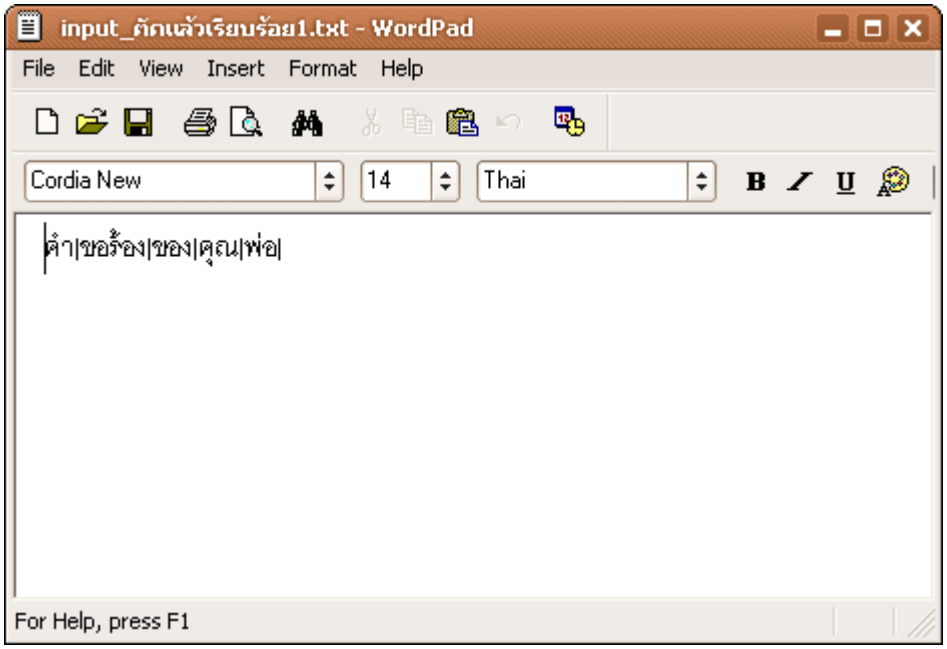
ตัวอย่างการใช้งาน



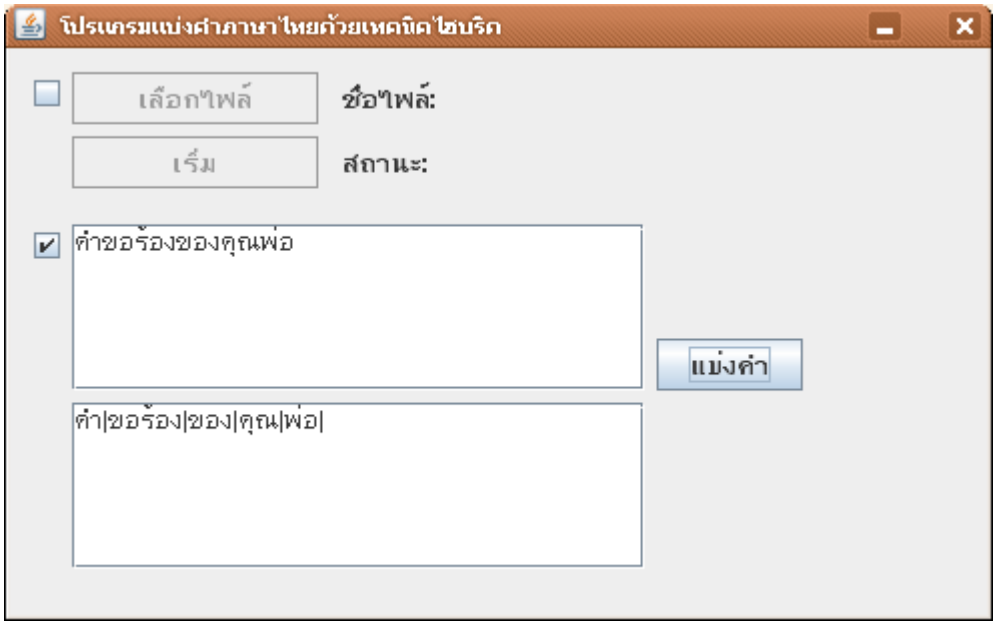

โปรแกรมแบ่งคำภาษาไทยด้วยเทคนิคไฮบริด
เลือกไฟล์
ชื่อไฟล์input.txt
Open
Look In:
TEST BEST 2010
input.txt
File Name:
input.txt
Files of Type:
All Files
Open
Cancel


โปรแกรมแบ่งคำภาษาไทยด้วยเทคนิคไฮบริด
เลือกไฟล์
ชื่อไฟล์input.txt
เริ่ม
สถานะ: เสร็จสิ้น
คำขอร้องของคุณพ่อ
แบ่งคำ
คำ|ขอร้อง|ของ|คุณ|พ่อ|


TEST BEST 2010
File Edit View Favorites Tools Help
Back Search Folders
Address C:\Documents and Settings\mishell\Desktop\TEST BEST 2010
Go
File and Folder Tasks
Other Places
Details
TEST BEST 2010
File Folder
Date Modified: Today, January 03, 2010, 3:57 PM
input.txt
Text Document
1 KB
input_ตัดแล้วเรียบร้อย1.txt
Text Document
1 KB


input_ตัดแล้วเรียบร้อย1.txt - WordPad
File Edit View Insert Format Help
Cordia New
14
Thai
คำ|ขอร้อง|ของ|คุณ|พ่อ|
For Help, press F1

ส่วนที่สอง เป็นพื้นที่ให้ทดลองกรอกข้อความ และ ผลลัพธ์ก็จะออกมาในพื้นที่ด้านล่าง



เอกสารอ้างอิง

[1] สุริทร์ จรรยาพรพงษ์. *A Thai Syllable Seperation Algorithm*. Asian Institute of Technology, 1983.

[2] ดวงแก้ว สวามิภักดิ์, *การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์*. มหาวิทยาลัยธรรมศาสตร์, 1990.

[3] วัชรพงศ์ โกมุทธรรมวิบูลย์และคณะ สำนักพิมพ์พัฒนาศึกษา, *คู่มือเตรียมสอบภาษาไทย ป.6 เข้า ม.1 และ NT,* 2009