PRINCE OF SONGKLA UNIVERSITY
FACULTY OF ENGINEERING

Midterm Examination: Semester 2 **Academic Year**: 2012

Date: 18 December 2012 **Time**: 9.00-12.00 (3 hours)

Subject Number: 242-500 **Room**: A401

Subject Title: Research and Development Methodologies

Exam Duration: 3 hours (180 minutes)

This paper has 18 pages (including a 9-page paper) and 8 questions and there are 130 marks (25%) to be collected.

Authorised Materials:

- Writing instruments (e.g. pens, pencils).
- Textbooks, notebooks, handouts, and dictionaries are permitted.

Instructions to Students:

- Scan all the questions before answering so that you can manage your time better.
- Write your answers in Thai only.
- Write your name and ID on every page.
- Any unreadable part will not be marked (wrong answer).

Cheating in this examination

Lowest punishment: Failed in this subject and courses dropped for next semester.

Highest punishment: Expelled.

NO	Time (Min)	Marks	Collected	NO	Time (Min)	Marks	Collected
1	30	23		5	30	30	
2	30	28		6	10	8	
3	10	10		7	10	6	
4	20	20		8	10	5	
Total	150	Raw Marks (130)			Collected (25%)		

Question 1 **(23 marks; 30 minutes)**

From the attached paper, spot format and style mistakes and tell what should be corrected.

Name_____ ID_____

Question 2 **(28 marks; 30 minutes)**

Tell the differences between the following pairs.

a) *Master's Thesis* and *Doctoral Thesis* (4 marks)

b) *Proposal* and *Thesis* (4 marks)

c) *Abstract* and *Introduction* (4 marks)

d) *Literature Review* and *Research Methods* (4 marks)

Name_____ ID_____

e) *Results and Discussion* and *Conclusion* (4 marks)

f) *Conference Proceedings* and *Journals* (4 marks)

g) *Basic Research* and *Applied research* (4 marks)

Question 3 **(10 marks; 10 minutes)**

Answer the following about *Plagiarism*.

a) What is the meaning of Plagiarism? (4 marks)

Name_____ ID_____

b) Tell possible results of plagiarism. (3 marks)

c) Tell how to avoid plagiarism (3 marks)

Question 4 **(20 marks; 20 minutes)**

Answer the following questions about reports.

a) What does the Front Matter contain? (4 marks)

b) What does the Back Matter contain? (3 marks)

c) What are guideline questions to check whether the work is ready for submission? List at least 4 questions (4 marks)

d) How do we write a literature survey? (4 marks)

e) Explain relationships among selling points, new methods, old and new data, same and different results, and research article and general report or review article. (5 marks)

Question 5 **(30 marks; 30 minutes)**

Answer the following questions about postgraduate study.

a) Why do we need semester progress seminars? Give at least 4 reasons. (4 marks)

b) Give at least 3 reasons why we need to do research. (3 marks)

c) Why do we need a logbook for research? (2 marks)

d) Why do we need to do a literature survey? Give at least 3 reasons. (3 marks)

e) What do you need to do at the meeting with your supervisor? (4 marks)

f) Why do we need to go to conferences? (3 marks)

g) Why do we need reviewers/referees for reviewing publications? (2 marks)

h) Give at least 4 approaches to enhance creativity (4 marks)

i) List 5 questions should be asked when starting a research. (5 marks)

Question 6 **(8 marks; 10 minutes)**

What presentation medium to use for the following types of messages?

a) figures and graphs

b) photos of complex objects

c) dynamic material, e.g. animation

d) words

e) the agenda and important points, to be stayed up all the time or a long time

f) working through something, where the process is important

g) complex tables, with lots of figures, equations,

h) anything that can't be understood in 30 seconds

Question 7 **(6 marks; 10 minutes)**

What are pitfalls and shortcomings for the following methods or medium?

a) Copy and Paste

b) Graphs

c) Diagrams

Question 8 **(5 marks; 10 minutes)**

What usage are the following types of outputs for?

a) Bar Graph

b) Circle Graph (Pie)

c) Line

d) Distributed Graph

e) Table

Pichaya Tandayya

Lecturer

Name_____ ID_____

ชื่อเรื่อง **การใช้เอ็นแกรมช่วยในการตัดสินใจแปลอักษรเบรลล์ที่ใช้คำควบกล้ำ สระผสม และอักษรเบรลล์สองเซลล์**

Enhancing Thai Braille translation with n-gram for decision making in the cases of compound consonants, vowels and characters

ชื่อคณะผู้วิจัย นายทศวัฒน์ ชุณหวิทยะธีระ และ ผศ.ดร. พิชญา ตัณฑัยย์

Mr. Totsawat Chunhawitayatera and Asst. Prof. Dr. Pichaya Tandayya

หน่วยงานต้นสังกัด ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์

บทคัดย่อ

การแปลงข้อความรหัสอักษรเบรลล์ภาษาไทยให้เป็นข้อความรหัสแอสกีภาษาไทยในปัจจุบัน มักใช้กฎไวยากรณ์ทางภาษาเข้ามากำกับวิธีการแปลง แต่ยังคงประสบปัญหาการแปลงคำควบกล้ำ, สระผสม, อักษรเบรลล์สองเซลล์และคำทับศัพท์ เนื่องจากไวยากรณ์ที่แตกต่างกันและความไม่แน่นอนในการใช้ภาษา จึงไม่สามารถเขียนโปรแกรมให้ครอบคลุมทุกๆ กรณีที่เกิดขึ้นได้ งานวิจัยนี้ได้นำเสนอวิธีแก้ไขปัญหาการแปลงข้อความรหัสอักษรเบรลล์ภาษาไทยให้เป็นข้อความรหัสแอสกีภาษาไทยเพื่อใช้กับประโยคที่มีคำควบกล้ำ, สระผสม, อักษรเบรลล์สองเซลล์และคำทับศัพท์ โดยใช้เทคนิคเอ็นแกรมเข้ามาช่วยในการตัดคำ ก่อนที่จะแปลงให้เป็นภาษาไทยเพื่อลดความกำกวมของภาษาและนำเสนอจำนวนแกรมที่เหมาะสมสำหรับใช้ตัดคำอักษรเบรลล์ภาษาไทย ซึ่งผลการทดลองสรุปว่าสามารถแปลงข้อความได้ถูกต้องมากขึ้นและใช้ปริมาณหน่วยความจำน้อยลง แต่ใช้เวลาแปลงมากกว่าเดิมโดยอยู่ในเกณฑ์ที่ยอมรับได้ และจำนวนแกรมที่เหมาะสมคือ 4 แกรม

คำสำคัญ : อักษรเบรลล์ภาษาไทย, เอ็นแกรมโมเดล, โปรแกรมแปลงข้อความรหัสอักษรเบรลล์ภาษาไทยเป็นรหัสแอสกีภาษาไทย

ABSTRACT

In the present, the Braille Thai to Thai translation applies grammar rules to control the translation process but there're are problems about translating compound consonants and vowels, two-cell Braille characters and transliterated words. The causes are too many grammar rules and the ambiguity of the language. Therefore, it's difficult to write a program to cover all conditions. This paper proposes a method to solve Braille Thai to Thai translation problems in the cases that the sentences contain compound consonants, vowels, two-cell Braille characters and transliterated words by using the N-gram model technique to wrap words before being translated into Thai in order to reduce the ambiguity of the language and propose the optimal N-gram number for wrapping Braille Thai words. The new method can improve the translation correctness and reduce the memory consumption better than the old method but the new method requires more translation time than the old method. We suggest the appropriate N-gram number of 4-gram for the Braille Thai text word wrap.

Key Words: Thai Braille, N-gram model, Braille to Thai translation program

---

ที่อยู่ที่สามารถติดต่อได้ นายทศวัฒน์ ชุณหวิทยะธีระ 31 หมู่ 3 ถ.ชุมพร-ระนอง ต.วังใหม่ อ.เมือง จ.ชุมพร 86190
E-mail Address : totsawat@hotmail.com

คำนำ

ในประเทศไทยได้มีการพัฒนาโปรแกรมช่วยแปลงรหัสอักษรเบรลล์ภาษาไทยเป็นรหัสแอสกีภาษาไทยแบบ File-to-file มาอย่างต่อเนื่อง แต่โปรแกรมเหล่านั้นยังคงประสบปัญหาการแปลงคำที่เกี่ยวข้องกับคำควบกล้ำ, คำที่มีการใช้สระผสม, คำที่มีการใช้อักษรเบรลล์สองเซลล์และคำทับศัพท์ ซึ่งในวิทยานิพนธ์เรื่อง "การแปลงเบรลล์และแอสกีภาษาไทยแบบทันทีทันใด" โดยมหาวิทยาลัยสงขลานครินทร์ ได้นำเสนอโปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file ซึ่งโปรแกรมนี้เป็นโปรแกรมแปลงข้อความรหัสอักษรเบรลล์ภาษาไทยเป็นข้อความรหัสแอสกีภาษาไทยแบบ File-to-file ซึ่งยังคงมีความผิดพลาดในการแปลงคำประเภทดังกล่าว

ในบทความนี้ พวกเราได้เสนอวิธีแก้ไขปัญหาดังกล่าวโดยนำเทคนิคเอ็นแกรมเข้ามาช่วยในการหาขอบเขตของคำอักษรเบรลล์ภาษาไทย ซึ่งจะทำให้ทราบขอบเขตที่แน่นอนของคำแต่ละคำ จึงส่งผลให้การแปลงเบรลล์ภาษาไทยเป็นแอสกีภาษาไทยทำได้ง่ายขึ้นและถูกต้องมากยิ่งขึ้น อีกทั้งการนำเทคนิคเอ็นแกรมมาประยุกต์ใช้นี้ ทำให้กระบวนการแปลงเบรลล์ภาษาไทยเป็นแอสกีภาษาไทยมีความซับซ้อนน้อยลง เมื่อเปรียบเทียบกับการแปลงเบรลล์ภาษาไทยเป็นแอสกีภาษาไทยที่ใช้เทคนิคกฎไวยากรณ์ทางภาษา โดยให้เอ็นแกรมคำนวณหาค่าความน่าจะเป็นของอักขระที่เขียนเรียงติดกัน (Character sequence) ที่เกิดขึ้นร่วมกันเป็นคำอักษรเบรลล์ภาษาไทย เพื่อช่วยในการระบุขอบเขตของคำอักษรเบรลล์ภาษาไทยว่าควรแบ่งคำออกมาเป็นรูปแบบไหนจึงจะเหมาะสม และเพิ่มความถูกต้องในการแปล

ในงานวิจัยนี้ พวกเรามุ่งเน้นวิจัยและพัฒนาเพื่อศึกษาการนำเทคนิคเอ็นแกรม เข้ามาประยุกต์ใช้กับการแปลงอักษรเบรลล์ภาษาไทย เพื่อช่วยปรับปรุงกระบวนการแปลงรหัสอักษรเบรลล์ภาษาไทยไปเป็นรหัสแอสกีภาษาไทยให้ถูกต้องมากยิ่งขึ้น โดยมีประเด็นที่ศึกษาดังต่อไปนี้

1) นำเทคนิคเอ็นแกรมเข้ามาประยุกต์ใช้ร่วมกับการแปลงเบรลล์ภาษาไทยเป็นแอสกีภาษาไทย จะช่วยเพิ่มความถูกต้องในการแปลงได้หรือไม่ โดยเปรียบเทียบกับโปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file เดิม ซึ่งได้นำเสนอไว้ในวิทยานิพนธ์เรื่อง "การแปลงเบรลล์และแอสกีภาษาไทยแบบทันทีทันใด"

2) หาค่าของจำนวนแกรมที่เหมาะสม เพื่อหาค่าของจำนวนแกรมที่ให้ผลลัพธ์ในการแบ่งคำที่ดีที่สุดโดยเปรียบเทียบระหว่าง 3-แกรม, 4-แกรม และ 5-แกรม โดยใช้วิธีพิจารณาจากความถูกต้องของการแบ่งคำอักษรเบรลล์ภาษาไทยระดับ 1 ซึ่งจะต้องถูกต้องตามหลักไวยากรณ์ของอักษรเบรลล์ภาษาไทยและเมื่อนำไปแปลงให้เป็นภาษาไทยแล้วจะต้องถูกต้องตามหลักไวยากรณ์ของภาษาไทยด้วย

โดยมีวัตถุประสงค์ในการเสนอวิธีการแปลงอักษรเบรลล์ภาษาไทยไปเป็นภาษาไทยแบบ File-to-file เพื่อแก้ปัญหากรณีคำควบกล้ำ, คำที่มีการใช้สระผสม, คำที่มีการใช้อักษรเบรลล์สองเซลล์และคำทับศัพท์ โดยนำเทคนิคเอ็นแกรมมาตัดคำอักษรเบรลล์ภาษาไทย และแปลงให้เป็นภาษาไทยโดยใช้วิธีจับคู่คำ และเสนอจำนวนแกรมที่เหมาะสมในการตัดคำอักษรเบรลล์ภาษาไทย

อุปกรณ์และวิธีการวิจัย

วิธีการวิจัยแบ่งออกเป็นสองส่วน คือ 1) ศึกษางานวิจัยและเทคโนโลยีที่เกี่ยวข้อง 2) ส่วนออกแบบและพัฒนาโปรแกรม เพื่อรองรับแนวคิดที่นำเสนอ

1) ศึกษางานวิจัยและเทคโนโลยีที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้อง

งานวิจัยและเทคนิคต่างๆ ที่เกี่ยวข้อง ซึ่งได้นำมาใช้ในการวิจัยนี้โดยประกอบไปด้วย 1) โปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file เดิม 2) เทคนิคการตัดคำ 3) เอ็นแกรมโมเดล 4) Smoothing Technique 5) คลังข้อมูลภาษา และ 6) CMU-SLM Toolkit

1. โปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file เดิม

โปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file ได้นำเสนอไว้ในวิทยานิพนธ์เรื่อง "การแปลงเบรลล์และแอสกีภาษาไทยแบบทันทีทันใด" ของมหาวิทยาลัยสงขลานครินทร์ ซึ่งโปรแกรมดังกล่าวทำ

หน้าที่แปลงเบรลล์ภาษาไทยเป็นแอสกีภาษาไทยในรูปแบบ File-to-file โดยกระบวนการทำงานของโปรแกรมนี้ ได้ใช้กฎไวยากรณ์ทางภาษาของอักษรเบรลล์ภาษาไทยมาควบคุมการแปลง ซึ่งสามารถให้ผลลัพธ์ในการแปลงถูกต้องระดับหนึ่ง

จากที่พวกเราศึกษางานวิจัยนี้พบว่ายังคงมีปัญหาในการแปลงเบรลล์ภาษาไทยเป็นแอสกีภาษาไทยอยู่ด้วยกัน 2 ประการคือ 1) ปัญหาการแปลงคำควบกล้ำและสระผสม ซึ่งปัญหานี้เกิดจากการตัดคำที่มีความกำกวนในการต่อคำ ทำให้โปรแกรมไม่สามารถแยกแยะได้ว่าอักขระที่เขียนเป็นอักขระของคำหน้าหรืออักขระของคำหลัง 2) ปัญหาการแปลงคำที่มีการใช้อักษรเบรลล์สองเซลล์ ซึ่งปัญหานี้เกิดจากมีอักษรเบรลล์สองเซลล์บางตัวมีรูปพ้องกันกับอักษรเบรลล์เซลล์เดียวสองตัวที่เขียนติดกัน ทำให้โปรแกรมไม่สามารถทราบได้ว่าตรงไหนเป็นอักษรเบรลล์สองเซลล์หนึ่งตัว หรือเป็นอักษรเบรลล์เซลล์เดียวสองตัวที่เขียนติดกัน

2. เทคนิคการตัดคำ

รู้ไหมว่าเทคนิคการตัดคำ ใช้ในการหาขอบเขตของคำในงานด้านประมวลผลภาษาธรรมชาติ เพื่อทำให้คอมพิวเตอร์สามารถเข้าใจความหมายของคำในภาษานั้นๆ ได้ และเนื่องจากภาษาบางภาษา เช่น ภาษาไทย ภาษาจีน ภาษาญี่ปุ่น ภาษาลาว เป็นต้น มีลักษณะการเขียนประโยคที่ประกอบไปด้วยคำย่อยๆ หลายคำเรียงติดกันโดยไม่มีการเว้นช่องว่างระหว่างคำเหมือนกับภาษาอังกฤษทำให้เกิดปัญหา “ขอบเขตของคำ” เมื่อนำข้อมูลเหล่านั้นไปประมวลผลด้วยคอมพิวเตอร์ ทำให้คอมพิวเตอร์ไม่สามารถทราบได้ว่าในประโยคนั้นๆ ประกอบด้วยคำกี่คำ และแต่ละคำประกอบด้วยตัวอักษรอะไรบ้าง จึงจำเป็นต้องตัดคำเสียก่อน ก่อนที่จะนำไปประมวลผลต่อ เพื่อแบ่งขอบเขตของคำแต่ละคำอย่างชัดเจน

อักษรเบรลล์ภาษาไทยมีลักษณะการเขียนเช่นเดียวกันกับภาษาไทยคือ เขียนคำแต่ละคำติดๆ กันเป็นประโยค และเว้นวรรคระหว่างประโยค แต่จะเรียงตัวอักษรนำ ตัวอักษรตาม สระ ตัวสะกด และวรรณยุกต์ที่แตกต่างไปจากภาษาไทย

ดังนั้นหากใช้เทคนิคการตัดคำเข้ามาช่วยแบ่งคำอักษรเบรลล์ภาษาไทย เพื่อให้ทราบขอบเขตของคำที่ชัดเจน และลดความกำกวม ทำให้การแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยทำได้ง่ายขึ้นและถูกต้องมากยิ่งขึ้นอีกด้วย นอกจากนี้ยังไม่มีงานวิจัยใดๆ เลยที่จะนำวิธีการตัดคำมาใช้กับอักษรเบรลล์ภาษาไทย ซึ่งในปัจจุบันมีเพียงงานวิจัยที่เกี่ยวข้องกับการตัดคำภาษาไทยเท่านั้น

ในปัจจุบันพบว่าเทคนิคการตัดคำโดยใช้คลังข้อมูลมีความถูกต้องแม่นยำมากที่สุด และเทคนิคการตัดคำโดยใช้เอ็นแกรมโมเดลเป็นเทคนิคการตัดคำโดยใช้คลังข้อมูลที่นิยมใช้ในปัจจุบันมากที่สุด

3. เอ็นแกรมโมเดล (N-gram model)

เอ็นแกรมโมเดล (N-gram model) คือ แบบจำลองที่ใช้คำนวณค่าความน่าจะเป็นของชุดอักขระ (character sequence) ที่เกิดขึ้นร่วมกันเป็นคำ หรือค่าความน่าจะเป็นของคำที่เขียนเรียงกัน (word sequence) ที่เกิดขึ้นร่วมกันเป็นประโยค โดยค่าความน่าจะเป็นของชุดอักขระหรือคำสามารถคำนวณได้จากคลังข้อมูลฝึกที่สร้างไว้

แกรม (Gram) คือ หน่วยที่ใช้ในการสร้างแบบจำลองอาจจะเป็นเสียง คำ หรืออักขระก็ได้และแกรมมีได้หลายขนาดแล้วแต่จะกำหนด ตั้งแต่ 1 จนถึง N โดย N เป็นจำนวนนับตั้งแต่ 1, 2, 3 ..., n

หลักการทำงานของเอ็นแกรมอาศัยหลักความเป็นจริงที่ว่าเมื่อมีการเขียนอักขระใดๆ หรือคำใดๆ เรียงติดกันเพื่อสร้างเป็นคำหรือเป็นประโยค แต่ละอักขระหรือคำที่เขียนนั้น จะมีความสัมพันธ์กับอักขระหรือคำที่เขียนไว้ก่อนหน้าด้วยค่าความน่าจะเป็นค่าหนึ่ง ตัวอย่างเช่น มีคำ 3 คำคือ ”ข้าว” “กิน”และ“ฉัน” เมื่อนำมาเขียนเรียงกันเพื่อสร้างเป็นประโยค จะสามารถเรียงได้หลายรูปแบบ คือ “ข้าว+กิน+ฉัน” “กิน+ฉัน+ข้าว” “ฉัน+กิน+ข้าว” เป็นต้น แต่รูปแบบ “ฉัน+กิน+ข้าว” จะสามารถพบได้มากที่สุดหรืออีกนัยหนึ่งก็คือรูปแบบนี้มีค่าความน่าจะเป็นสูงที่สุดนั่นเอง แต่ทั้งนี้เทคนิคแบบเอ็นแก

รมเป็นการคำนวณในเชิงสถิติเท่านั้นและไม่ได้นำเอาเรื่องของกฎไวยากรณ์ภาษาเข้ามาเกี่ยวข้องด้วย โดยค่าความน่าจะเป็นมีสมการคำนวณดังนี้

ค่าความน่าจะเป็นที่จะเกิดคำหรืออักขระ x_i โดยมีชุดของคำหรือชุดอักขระ $x_{i-n},\ldots,x_{i-3},x_{i-2},x_{i-1}$ นำหน้า $= P(x_i| x_{i-1},x_{i-2},x_{i-3},\ldots,x_{i-n})$

ซึ่ง $P(x_i| x_{i-1},x_{i-2},x_{i-3},\ldots,x_{i-n}) = c(x_{i-n},\ldots,x_{i-3},x_{i-2},x_{i-1},x_i)/c(x_{i-1},x_{i-2},x_{i-3},\ldots,x_{i-n})$

โดยที่ $c(x_{i-n},\ldots,x_{i-3},x_{i-2},x_{i-1},x_i)$ คือจำนวนครั้งในคลังข้อมูลฝึกที่เกิด x_i ร่วมกับ $x_{i-n},\ldots,x_{i-3},x_{i-2},x_{i-1}$

$c(x_{i-1},x_{i-2},x_{i-3},\ldots,x_{i-n})$ คือจำนวนครั้งที่เกิด $x_{i-n},\ldots,x_{i-3},x_{i-2},x_{i-1}$ ในคลังข้อมูลฝึก

4. Smoothing Techniques

Smoothing techniques คือ เทคนิคการประมาณค่าความน่าจะเป็นให้กับคำหรือชุดของคำที่ไม่มีอยู่หรือไม่พบในคลังข้อความ ซึ่งเทคนิคนี้อาศัยวิธีการประมาณค่าทางสถิติในการคำนวณหาค่าความน่าจะเป็นของคำที่ไม่พบในคลังข้อความ และเทคนิคนี้เป็นกระบวนการที่จำเป็นสำหรับงานด้านสถิติ อีกทั้งเป็นการป้องกันปัญหาการหารด้วยศูนย์อีกด้วย และจากข้อจำกัดของคลังข้อความที่ไม่สามารถจัดเก็บคำและคู่ของคำทั้งหมดที่เกิดขึ้นจริงไว้ในคลังข้อมูลได้ทั้งหมดทำให้คำบางคำหรือคู่ของคำบางคำอาจไม่มีอยู่ในคลังข้อมูล จากการศึกษาพบว่าเทคนิคของ Jelinek-Mercer Smoothing (Interpolation) จะให้ผลลัพธ์ดีที่สุดเมื่อเปรียบเทียบกับวิธีอื่นๆ โดยมีสมการคำนวณค่าความน่าจะเป็นดังนี้

$$P_{interp}(w_i | w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+1}^{i-1}} P_{ML}(w_i | w_{i-n+1}^{i-1}) + (1-\lambda_{w_{i-n+1}^{i-1}}) P_{interp}(w_i | w_{i-n+2}^{i-1})$$

โดยที่ $P_{ML}(w_i | w_{i-n+1}^{i-1}) = C(w_{i-n+1}^{i-1} w_i) \quad C(w_{i-n+1}^{i-1})$

และ $$\lambda_{w_{i-n+1}^{i-1}} = \frac{C(w_{i-n+1}^{i-1})}{|W_i : C(w_{i-n+1}^{i-1})> 0|}$$

5. คลังข้อมูลภาษา

คลังข้อมูลภาษา คือ ข้อมูลภาษาเขียนหรือภาษาพูดที่เป็นภาษาที่ใช้จริง ซึ่งถูกรวบรวมขึ้นมาในปริมาณที่มากเพียงพอตามเงื่อนไขที่กำหนดเพื่อให้ได้คลังข้อมูลที่เป็นตัวแทนของภาษาที่ต้องการ คลังข้อมูลภาษานำไปใช้งานในด้านภาษาศาสตร์ การสร้างแบบจำลองภาษารวมถึงการประมวลผลภาษาธรรมชาติอีกด้วย เช่น นำไปใช้สร้างพจนานุกรม นำข้อมูลด้านสถิติของคำไปสร้างเป็นแบบจำลองเพื่อนำมาใช้ในงานประมวลผลภาษาธรรมชาติด้วยคอมพิวเตอร์

ในงานวิจัยนี้ พวกเราได้นำคลังข้อความภาษาไทยมาประยุกต์ใช้งานในการฝึกฝนเอ็นแกรมโมเดล โดยได้เลือกคลังข้อความ BEST ซึ่งเป็นคลังข้อความภาษาไทยและได้คัดเลือกเฉพาะประโยคที่มีการใช้คำควบกล้ำ, สระผสม, อักษรเบรลล์สองเซลล์และคำทับศัพท์เท่านั้น ทั้งนี้เพราะ โปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file เดิมสามารถแปลงประโยคทั่วๆ ไปได้ถูกต้องอยู่แล้วและต้องนำมาแปลงให้อยู่ในรูปแบบของอักษรเบรลล์ภาษาไทยระดับ 1 เสียก่อน จึงจะสามารถนำไปฝึกฝนเอ็นแกรมโมเดลได้ โดยในที่นี้จะใช้โปรแกรมแปลงภาษาไทยเป็นอักษรเบรลล์ (thai2brl) ซึ่งนี้ได้นำเสนอในวิทยานิพนธ์เรื่อง "การแปลงเบรลล์และแอสกีภาษาไทยแบบทันทีทันใด" ของมหาวิทยาลัยสงขลานครินทร์

คลังข้อความ BEST เป็นคลังข้อความภาษาไทยที่มีการเก็บรวบรวมคำอย่างเป็นระบบและมีขนาดใหญ่ที่สุดในปัจจุบัน ซึ่งพัฒนาโดยหน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา (HLT) ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) โดยคัดเลือกข้อความจากงานเขียนลักษณะต่าง ๆ 3 ประเภท คือ 1) ข้อความจากหนังสือประเภทนวนิยายในฐานะเป็นตัวแทนของภาษาพูดทั่วไป 2) ข้อความจากเว็บไซต์ www.midnightuniv.org และข้อความจากสารานุกรมสำหรับเยาวชนไทยในฐานะเป็นตัวแทนของภาษาเขียนอย่างเป็นทางการ 3) ข้อความจากหนังสือพิมพ์บนอินเตอร์เน็ตในฐานะเป็นตัวแทนของภาษาข่าว

6. CMU-SLM Toolkit

ชุดเครื่องมือ CMU-SLM Toolkit เป็นชุดเครื่องมือที่ใช้ในการสร้างโมเดลภาษา (Language model) ซึ่งชุดเครื่องมือนี้ได้รับการพัฒนาโดยมหาวิทยาลัย Carnegie Mellon โดยในงานวิจัยนี้ได้เลือกใช้ชุดเครื่องมือ CMU-SLM Toolkit

เพื่อช่วยอำนวยความสะดวกในการสร้างโมเดลภาษาของอักษรเบรลล์ภาษาไทยระดับ 1 เนื่องจากชุดเครื่องมือนี้ มีการใช้งานกันอย่างแพร่หลายและมีเทคนิคการ Smoothing แบบ Interpolation ชุดเครื่องมือนี้มีการเผยแพร่เป็นสาธารณะ มีคู่มืออธิบายการใช้งานชัดเจน ทำให้ใช้งานได้ง่าย และเป็นชุดเครื่องมือที่ครบถ้วนสมบูรณ์ โดยในชุดเครื่องมือนี้จะประกอบด้วยโปรแกรมย่อยจำนวนหลายโปรแกรมด้วยกันเช่น text2 wfreq, wfreq2vocab, text2wngram, text2idngram, idngram2lm, interpolate เป็นต้น

2) ส่วนออกแบบและพัฒนาโปรแกรม

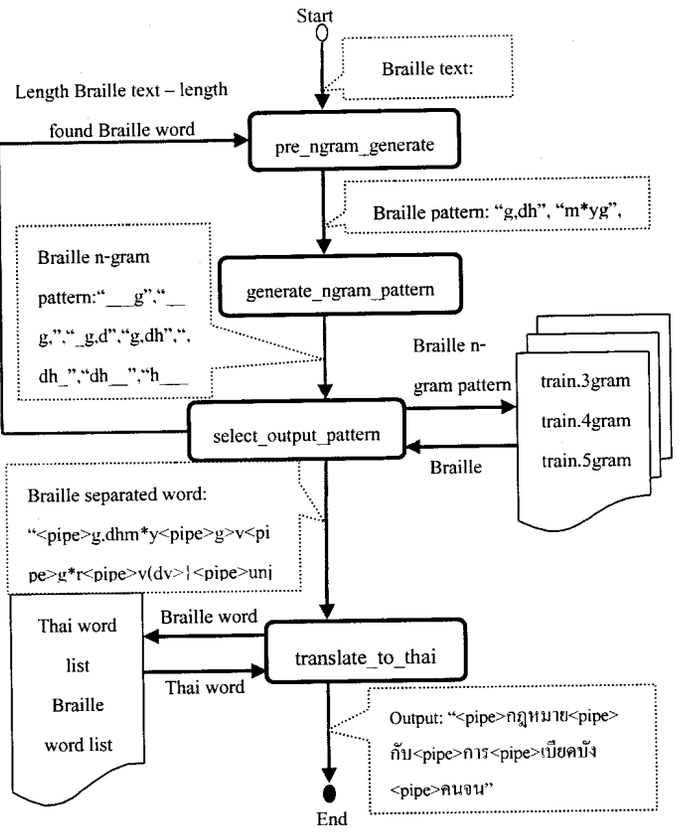
เครื่องมือที่ใช้ในการวิจัยนี้ประกอบด้วย

1. CMU-SLM toolkit เป็นชุดเครื่องมือสำหรับใช้สร้างโมเดลภาษา
2. คลังข้อความ BEST เป็นคลังข้อความภาษาไทย ซึ่งต้องนำไปแปลงให้อยู่ในรูปอักษรเบรลล์ภาษาไทยเสียก่อนจึงจะนำไปใช้ฝึกฝนเอ็นแกรมโมเดล
3. โปรแกรมแปลงภาษาไทยเป็นอักษรเบรลล์ภาษาไทย (thai2brl) เป็นโปรแกรมที่นำเสนอไว้ในงานวิจัยเรื่อง "การแปลงเบรลล์และแอสกีภาษาไทยแบบทันทีทันใดของมหาวิทยาลัยสงขลานครินทร์" ใช้แปลงข้อความภาษาไทยเป็นอักษรเบรลล์ภาษาไทย
4. ชุดข้อมูลสำหรับทดสอบโปรแกรมแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยที่ใช้เทคนิคเอ็นแกรมเข้ามาช่วยตัดคำ ที่ได้นำเสนอไว้ในงานวิจัยนี้

งานวิจัยนี้ พัฒนาโปรแกรมโดยใช้ภาษา C/C++

ภาพรวมของระบบ

ในงานวิจัยนี้ได้นำเสนอวิธีการปรับปรุงประสิทธิภาพในการแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยในรูปแบบ File-to-file โดยนำเทคนิคเอ็นแกรมเข้ามาช่วยในการตัดคำอักษรเบรลล์ภาษาไทยเพื่อลดความกำกวมที่เกิดขึ้นจากลักษณะการเขียนของอักษรเบรลล์ภาษาไทยระดับ 1 หลังจากนั้นจะแปลงให้เป็นภาษาไทยโดยใช้วิธีเทียบแบบคำต่อคำ ซึ่งจำเป็นจะต้องสร้างรายการคำอักษรเบรลล์ภาษาไทยและรายการคำภาษาไทยไว้ล่วงหน้า ในภาพที่ 1 แสดงภาพรวมของระบบที่ได้ออกแบบไว้



รูปที่ 1 แสดงองค์ประกอบต่างๆของระบบที่ได้ออกแบบไว้และลักษณะการเชื่อมต่อกันของฟังก์ชันหลักรวมทั้งการเชื่อมต่อองค์ประกอบอื่นๆ ที่จำเป็นต้องใช้ร่วมในการแปลงอักษรเบรลล์ภาษาไทยระดับ 1 ให้เป็นภาษาไทย ซึ่งประกอบไปด้วยส่วนต่างๆ ทั้งหมด 6 ส่วนด้วยกันคือ

1) ขั้นตอน pre_ngram_generate
2) ขั้นตอน generate_ngram_pattern
3) ขั้นตอน select_output_pattern
4) ไฟล์ฝึกฝนเอ็นแกรมโมเดล
5) ขั้นตอน translate_to_thai
6) รายการคำภาษาไทยและรายการคำอักษรเบรลล์ภาษาไทย

โดยอินพุต/เอาท์พุทและรายละเอียดแต่ละส่วนมีดังนี้

อินพุตและเอาท์พุทของระบบ

โปรแกรมแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยแบบ File-to-file ที่ใช้เทคนิคเอ็นแกรมเข้ามาช่วยในการตัดคำที่

ได้นำเสนอไว้ในงานวิจัยนี้จะรับอินพุตเป็นไฟล์ข้อความธรรมดา ซึ่งภายในได้บรรจุข้อความที่เขียนให้อยู่ในรูปแบบอักษรเบรลล์ภาษาไทยระดับ 1 โดยการทำงานของโปรแกรมนี้จะประมวลผลข้อความอักษรเบรลล์ภาษาไทยระดับ 1 ทีละ 1 ประโยค โดยจะแบ่งแต่ละประโยคออกจากกันด้วยการเว้นวรรค เช่นเดียวกันกับการเขียนภาษาไทย เมื่อโปรแกรมแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยที่ได้นำเสนอนี้ประมวลผลเสร็จแล้ว จะได้เอาท์พุทเป็นประโยคภาษาไทย และนำประโยคภาษาไทยที่ได้นั้นเขียนใส่ในไฟล์เอาท์พุทตามที่ผู้ใช้งานโปรแกรมนี้ได้ตั้งชื่อไฟล์เอาไว้ ซึ่งจะเป็นไฟล์ข้อความธรรมดา และการทำงานของโปรแกรมก็จะเป็นเช่นนี้ไปเรื่อยๆ จนกว่าข้อความอักษรเบรลล์ภาษาไทยระดับ 1 ในไฟล์อินพุตจะหมดลง

1. ขั้นตอน pre_ngram_generate

ขั้นตอนนี้รับอินพุตเป็นข้อความอักษรเบรลล์ภาษาไทย และแบ่งออกเป็นแต่ละประโยคโดยใช้การเว้นวรรค จากนั้นตัดประโยคที่รับมาเป็นส่วนๆ ตามขนาดของจำนวนแกรมที่ได้กำหนดไว้เพื่อเพิ่มอัตราการจับคู่คำของแต่ละอักขระให้มากขึ้น และให้เอาท์พุทเป็นชุดของอักขระอักษรเบรลล์ภาษาไทยโดยแต่ละชุดมีความยาวเท่ากับจำนวนแกรมที่ได้กำหนดไว้

2. ขั้นตอน generate_ngram_pattern

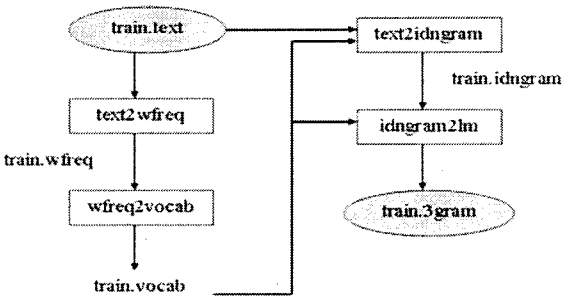
ขั้นตอนนี้รับอินพุตเป็นชุดของอักขระอักษรเบรลล์ภาษาไทยจากขั้นตอน pre_ngram_generate และนำมาสร้างรูปแบบของ n-gram pattern โดยอ้างอิงกระบวนการนี้จากเว็บไซต์ http://www.hlt.nectec.or.th เพื่อนำไปใช้ค้นหาเปรียบเทียบรูปแบบของคำที่เขียนเรียงติดกันตามหลักการของเอ็นแกรมโมเดลในขั้นตอนต่อไป โดยให้เอาท์พุทเป็นชุดของรูปแบบ n-gram pattern

3. ขั้นตอน select_output_pattern

ขั้นตอนนี้รับอินพุตเป็นชุดรูปแบบของ n-gram pattern จากนั้นเปรียบเทียบแต่ละ pattern กับไฟล์ฝึกฝนเอ็นแกรมโมเดลโดยพิจารณาว่า pattern นั้นๆ ใกล้เคียงกับคำใดมากที่สุดและ pattern รูปแบบไหนมีค่าความน่าจะเป็นสูงที่สุด และหากมี pattern นั้นปรากฏอยู่ในประโยคก็เลือกเอา pattern นั้นเป็นคำๆหนึ่ง หากไม่มีก็จะเลือกเอาคำที่มีความคล้ายคลึงกับ pattern นั่นมากที่สุดมาแทน แต่จะต้องมี pattern นั้นปรากฏอยู่ในประโยคด้วย จากนั้นใส่เครื่องหมาย “<pipe>” เพื่อเป็นสัญลักษณ์ในการแบ่งคำแต่ละคำออกจากกัน และส่งค่าความยาวของประโยคที่เหลือกลับไปให้ขั้นตอน pre_ngram_generate เพื่อสร้างชุดของอักขระใหม่และทำแบบนี้วนไปเรื่อยๆ จนกว่าจะจบประโยค เมื่อจบขั้นตอนนี้แล้วจะให้เอาท์พุทเป็นประโยคอักษรเบรลล์ที่มีการแบ่งคำแต่ละคำออกจากกันโดยใช้เครื่องหมาย “<pipe>” คั่นคำแต่ละคำ

4. ไฟล์ฝึกฝนเอ็นแกรมโมเดล

ไฟล์ที่ใช้ฝึกฝนเอ็นแกรมโมเดลนี้ สร้างขึ้นจากการใช้ชุดเครื่องมือ CMU-SLM toolkit โดยในการทดลองได้ใช้ขนาดของจำนวนแกรมเท่ากับ 3-gram 4-gram และ 5-gram เพื่อเปรียบเทียบว่าจำนวนแกรมเท่าใดจึงจะให้ผลลัพธ์ที่ดีที่สุด ซึ่งวิธีการฝึกฝนเอ็นแกรมโมเดล (Training N-gram model) คือ กระบวนการทางสถิติเพื่อให้เอ็นแกรมโมเดลคำนวณค่าความน่าจะเป็นของคำหนึ่งๆ ว่าสามารถที่จะเกิดขึ้นร่วมกับคำใดได้บ้าง โดยขั้นตอนการฝึกฝนเอ็นแกรมโมเดลอ้างอิงมาจากเว็บไซต์ www.hlt.nectec.or.th



ภาพที่ 2 การฝึกฝนเอ็นแกรมโมเดล

โดยการฝึกฝนเอ็นแกรมโมเดลจะได้เป็นไฟล์นามสกุล .gram และในไฟล์ .gram นี้จะบรรจุค่าความน่าจะเป็นของคำนั้นๆ ว่ามีโอกาสที่จะเกิดขึ้นร่วมกับคำใดได้บ้างด้วยค่าความน่าจะเป็นเท่าไร และบรรจุค่า back off weight ที่จะนำไปใช้ในกระบวนการ Smoothing อีกด้วย ซึ่งมีการจัดเก็บให้อยู่ในรูปแบบของ ARPA format ซึ่งเป็นรูปแบบไฟล์มาตรฐานของโมเดลภาษา (Language Model)

5. ขั้นตอน translate_to_thai

ขั้นตอนนี้เป็นขั้นตอนแปลงข้อความอักษรเบรลล์ภาษาไทยที่ได้ตัดคำแล้วโดยใช้สัญลักษณ์ “<pipe>” คั่นระหว่างคำ แปลงให้เป็นภาษาไทย โดยใช้เทคนิควิธีการจับคู่คำ ซึ่งอาศัยรายการคู่คำอักษรเบรลล์ภาษาไทยและภาษาไทยที่สร้างไว้ก่อนหน้านี้แล้ว โดยแทนที่คำอักษรเบรลล์ภาษาไทยแต่ละคำด้วยคำภาษาไทยที่ได้จับคู่เอาไว้ และลบเครื่องหมาย “<pipe>” ออก ซึ่งผลลัพธ์ท้ายสุดก็จะได้ข้อความภาษาไทยเพื่อนำไปเปรียบเทียบประสิทธิภาพกับโปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file[1] เดิมได้

6. รายการคู่คำอักษรเบรลล์ภาษาไทยและภาษาไทย

รายการคำอักษรเบรลล์ภาษาไทยและรายการคำภาษาไทยที่จัดเตรียมขึ้นมานี้ จะนำไปใช้ในขั้นตอน Translate_to_Thai เพราะเอ็นแกรมโมเดลนั้นมีความสามารถในเรื่องของการหาขอบเขตของคำเท่านั้น ไม่สามารถใช้ในการแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยได้ นอกจากนี้ต้องการเปรียบเทียบประสิทธิภาพระหว่างโปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file[1] เดิม กับโปรแกรมแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยแบบ File-to-file ที่ได้นำเสนอนี้ ซึ่งโปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file เดิมนั้นให้เอาท์พุทเป็นข้อความภาษาไทย ทำให้จำเป็นต้องมีขั้นตอน Translate_to_Thai เพิ่มเติมเข้ามาเพื่อทำให้เอาท์พุทเป็นข้อความภาษาไทยเหมือนกันเพื่อที่จะเปรียบเทียบกันได้โดยตรง

ดังนั้นจึงได้ออกแบบให้การแปลงจากอักษรเบรลล์ภาษาไทยเป็นภาษาไทยใช้วิธีการแบบเทียบเป็นคำต่อคำหลังจากใช้เทคนิคเอ็นแกรมในการตัดคำแล้ว ตัวอย่างเช่น เมื่อตัดคำอักษรเบรลล์ภาษาไทยได้เป็นคำว่า “g,dhm*y” ซึ่งตรงกันกับคำว่า “กฎหมาย” ในภาษาไทย เมื่อใช้วิธีเทียบเป็นคำต่อคำก็จะแทนที่คำ “g,dhm*y” ด้วยคำว่า “กฎหมาย” ลงไปแทนในตำแหน่งเดียวกันของข้อความ ทำให้ไฟล์เอาท์พุทเป็นข้อความภาษาไทยเช่นเดียวกัน

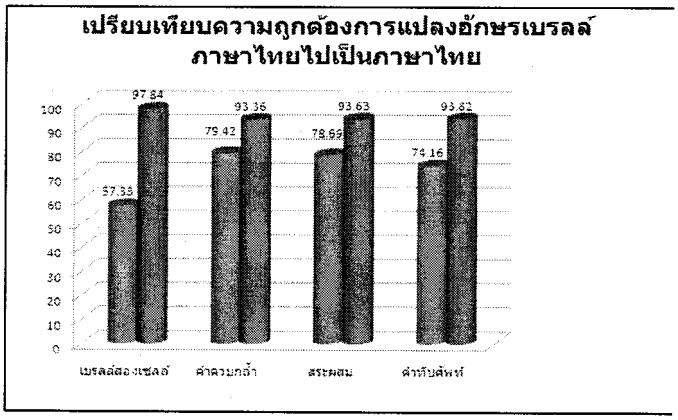
ผลการทดสอบเปรียบเทียบประสิทธิภาพ

ในการทดสอบเปรียบเทียบประสิทธิภาพระหว่างโปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file เดิมกับโปรแกรมแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยที่ใช้เทคนิคเอ็นแกรมเข้ามาช่วยในการตัดคำที่ได้นำเสนอนี้ จะทดสอบเปรียบเทียบทั้งหมด 3 ด้าน ด้วยกันคือ 1) ความถูกต้องของการแปลง 2) ปริมาณการใช้หน่วยความจำ และ 3) ความรวดเร็วในการแปลง ซึ่งมีรายละเอียดดังนี้

1) ด้านความถูกต้องของการแปลง

ในการทดสอบด้านความถูกต้องของการแปลงข้อความอักษรเบรลล์ภาษาไทยเป็นภาษาไทยได้ใช้ชุดข้อมูลทดสอบซึ่งแบ่งเป็น 4 ชุดด้วยกันคือ ชุดที่ 1) ข้อความที่ใช้อักษรเบรลล์สองเซลล์ จำนวน 8,713 บรรทัด ชุดที่ 2) ข้อความที่ใช้คำควบกล้ำ จำนวน 231,115 บรรทัด ชุดที่ 3) ข้อความที่ใช้สระผสม จำนวน 219,022 บรรทัด ชุดที่ 4) ข้อความที่ใช้คำทับศัพท์ จำนวน 18,881 บรรทัด โดยแต่ละบรรทัดจะมี 1 ประโยคและในแต่ละประโยคจะมีคำประเภทนั้นๆ อย่างน้อย 1 คำ

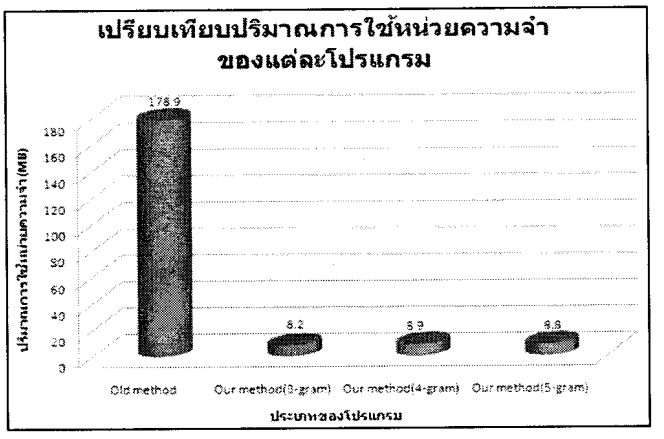
โดยชุดข้อมูลทดสอบแต่ละชุดจะมีลักษณะการเขียนที่แตกต่างกัน 3 ประเภทคือ 1) ภาษาเขียนแบบเป็นทางการ 2) ภาษาเขียนแบบไม่เป็นทางการ 3) ภาษาข่าว ทั้งนี้เพื่อต้องการให้ข้อมูลทดสอบมีลักษณะใกล้เคียงกับการเขียนโดยทั่วไปมากที่สุด โดยอ้างอิงจากโครงสร้างการจัดเก็บข้อความของคลังข้อความ BEST ผลการทดลองแสดงในภาพที่ 3



ภาพที่ 3 ผลการเปรียบเทียบความถูกต้อง

2) ด้านปริมาณการใช้หน่วยความจำ

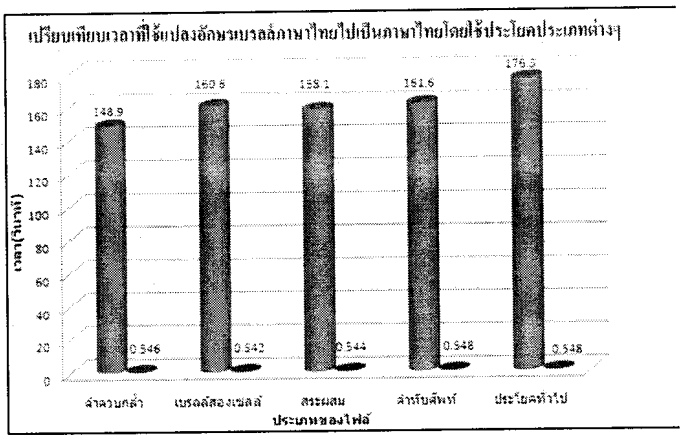
ในการทดสอบด้านปริมาณการใช้หน่วยความจำจะเริ่มวัดปริมาณหน่วยความจำที่โปรแกรมต้องใช้ตั้งแต่เริ่มเรียกใช้งานโปรแกรมจนกระทั่งโปรแกรมสิ้นสุดการทำงาน โดยใช้ชุดข้อมูลทดสอบเช่นเดียวกันกับวิธีการวัดความถูกต้องของการแปลง และวัดทั้งหมด 10 ครั้งและหาค่าเฉลี่ยปริมาณการใช้หน่วยความจำของโปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file เดิมกับโปรแกรมแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยที่ใช้เทคนิคเอ็นแกรมเข้ามาช่วยในการตัดคำ โดยแบ่งเป็น เมื่อใช้จำนวนแกรมเท่ากับ 3-gram, 4-gram และ 5-gram ซึ่งผลที่ได้แสดงในภาพที่ 4



ภาพที่ 4 ผลเปรียบเทียบปริมาณการใช้หน่วยความจำ

3) ด้านความรวดเร็วในการแปลง

ในด้านความรวดเร็วในการแปลงจะใช้วิธีจับเวลาตั้งแต่เรียกใช้โปรแกรมจนกระทั่งโปรแกรมสิ้นสุดการทำงาน โดยไฟล์ที่นำมาทดสอบจะแบ่งออกเป็น 5 ประเภทคือ 1) ประโยคที่ใช้คำควบกล้ำ 2) ประโยคที่ใช้อักษรเบรลล์สองเซลล์ 3) ประโยคที่ใช้สระผสม 4) ประโยคที่ใช้คำทับศัพท์ และ 5) ประเภทประโยคทั่วไป ซึ่งแต่ละประเภทจะใช้ไฟล์ทั้งหมด 5 ไฟล์และประเภทประโยคทั่วไปจะใช้ทั้งหมด 10 ไฟล์โดยมีขนาดประมาณ 10 kB (ประมาณ 220 บรรทัด) โดยแต่ละบรรทัดจะมี 1 ประโยคและในแต่ละประโยคจะมีคำประเภทนั้นๆ อย่างน้อย 1 คำ และทดสอบทั้งหมด 10 ครั้งและหาค่าเฉลี่ย ทั้งนี้เพื่อต้องการให้ใกล้เคียงกับการนำโปรแกรมไปใช้งานจริงมากที่สุด ซึ่งผลที่ได้แสดงในภาพที่ 5



ภาพที่ 5 ผลการเปรียบเทียบเวลาที่ใช้

ผลการวิจัยและการอภิปรายผล

จากการพัฒนาโปรแกรมแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยที่ใช้เทคนิคเอ็นแกรมเข้ามาช่วยในการตัดคำ เพื่อแก้ไขปัญหาของโปรแกรมแปลงเบรลล์เป็นแอสกีภาษาไทยแบบ File-to-file เดิม โดยสามารถแก้ไขปัญหาการแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยได้เป็นอย่างดี และใช้ปริมาณหน่วยความจำน้อยกว่าโปรแกรมเดิมซึ่งใช้วิธีการ recursion มาก ซึ่งเหมาะสมในการนำไปใช้ในอุปกรณ์ Braille Note ที่มีทรัพยากรอยู่อย่างจำกัด

อย่างไรก็ตามการใช้งานโปรแกรมที่ได้พัฒนาขึ้นมาใหม่นี้ก็มีข้อด้อย คือเวลาที่ใช้ในการแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยใช้เวลาแปลงที่มากกว่าโปรแกรมเดิมมาก ทำให้โปรแกรมนี้เหมาะกับการใช้แปลงไฟล์ที่มีขนาดไม่ใหญ่มาก ซึ่งผู้ใช้งานสามารถยอมรับได้ แต่หากเป็นไฟล์ที่มีขนาดใหญ่จะใช้เวลาแปลงนาน จึงนับว่าเป็นข้อด้อยของโปรแกรมนี้ที่ควรนำไปปรับปรุงให้ดีกว่าเดิม

สรุป

บทความนี้เสนอการนำเทคนิคเอ็นแกรมโมเดลมาประยุกต์ใช้ตัดคำอักษรเบรลล์ภาษาไทยเพื่อลดความกำกวมของภาษาก่อนที่จะแปลงไปเป็นภาษาไทย ซึ่งวิธีการนี้สามารถช่วยแก้ปัญหาการแปลงอักษรเบรลล์ภาษาไทยเป็นภาษาไทยได้เป็นอย่างดี โดยเฉพาะปัญหาที่เกี่ยวข้องกับการแปลงคำควบกล้ำ คำที่มีการใช้สระผสม

อักษรเบรลล์สองเซลล์และคำทับศัพท์ และจำนวนแกรมที่เหมาะสมสำหรับใช้ตัดคำอักษรเบรลล์ภาษาไทยคือ 4 แกรม

ข้อเสนอแนะ

การนำเทคนิคเอ็นแกรมเข้ามาช่วยตัดคำอักษรเบรลล์ภาษาไทยก่อนที่จะแปลงให้เป็นภาษาไทยสามารถช่วยลดความกำกวมของคำได้เป็นอย่างดี แต่ยังคงมีประเด็นเรื่องความรวดเร็วในการแปลงที่ทำได้ช้ากว่าโปรแกรมเดิมมาก ซึ่งไม่เหมาะสมในการนำไปใช้แปลงไฟล์ขนาดใหญ่ทั้งหมด แต่ควรใช้งานร่วมกับวิธีการอื่นๆ เพื่อเพิ่มความถูกต้องในการแปลงคำควบกล้ำ คำที่มีการใช้สระผสม อักษรเบรลล์สองเซลล์และคำทับศัพท์

เอกสารอ้างอิง

วรพล ทินกรสูตินุตร. การแปลงเบรลล์และแอสกีภาษาไทยแบบทันทีทันใด. วิทยานิพนธ์ปริญญามหาบัณฑิตสาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์

Stanley F. Chen and Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling, Aiken Computational Laboratory Harvard University.

N-gram. (2011). Retrieved on September 5, 2011, from http://en.wikipedia.org/wiki/N-gram

BEST Corpus. (2011). Retrieved on September 5, 2011, from http://thailang.nectec.or.th/2009/index.php

Philip Clarkson. (2011). CMU-SLM language model toolkit. Retrieved on September 26, 2011, from http://mi.eng.cam.ac.uk/~prc14/toolkit.html

Training N-gram model. (2011). Retrieved on September 5, 2011, from http://www.hlt.nectec.or.th/speech/index.php?option=com_content&view=article&id=70&Itemid=94

Word wrap. (2011). Retrieved on September 5, 2011, from http://en.wikipedia.org/wiki/Word_wrap

Language model. (2011). Retrieved on September 5, 2011, from http://www.hlt.nectec.or.th/speech/index.php?option=com_content&view=article&id=70&Itemid=94