การปรับปรุงกฎสำหรับตัดคำในเอกสารไทย
Improved Rule-Based for Thai Documents

ปโยธร อุราธรรมกุล
นักศึกษาระดับบัณฑิตศึกษา
ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ หาวิทยาลัยขอนแก่น
Email:payothorn@gmail.com

กานดา รุณนะพงศา
ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น
Email: krunapon@kku.ac.th

บทคัดย่อ

การตัดคำไทย (Thai Word Segmentation) คือการแยกแต่ละคำในเอกสารไทยออกจากกันเพื่อนำไปใช้ประโยชน์ในด้านอื่นๆ เช่น การสังเคราะห์เสียงพูด การแปลภาษา เป็นต้น เอกสารที่มีอยู่ ณ ปัจจุบันไม่เพียงแต่จะมีคำไทยแท้เท่านั้น คำบางคำที่มาจากภาษาต่างประเทศที่ถูกสะกดอยู่ในรูปของคำอ่านภาษาไทยบางคำจะมีการผสมอักษรที่แตกต่างนอกเหนือออกไปจากกฎการตัดคำ (Rule-based) แบบเดิมที่มีอยู่ เนื่องจากคำเหล่านี้มีอยู่มากมายและเกิดใหม่อยู่เสมอ บทความนี้นำเสนอการปรับปรุงกฎการตัดคำให้มีความยืดหยุ่นมากขึ้นจะเป็นประโยชน์สำหรับการตัดคำที่ไม่รู้จักหรือไม่มีความหมายอยู่ตามพจนานุกรม จากผลการทดลองพบว่าการตัดคำระดับพยางค์เพิ่มขึ้นถึง 2.14 เปอร์เซ็นต์ และระดับคำเพิ่มขึ้นถึง1.76 เปอร์เซ็นต์จากเทคนิคเดิม

คำสำคัญ การตัดคำ, กฎการตัดคำ

1. บทนำ

ลักษณะของประโยคในภาษาไทยมีการเขียนติดกันไป ทำให้ยากต่อการนำไปใช้งานในบางด้าน เช่น การสังเคราะห์เสียงพูด การแปลภาษา เป็นต้น ได้มีผู้คิดค้นวิธีที่จะแยกคำแต่ละคำออกจากประโยคซึ่งมีการเขียนติดกันไปอย่างต่อเนื่องทั้งประโยค ในงานวิจัยนี้จะกล่าวถึงการตัดคำโดยอาศัยอักขระวิธีเป็นหลักการพื้นฐาน การประสมคำซึ่งมีความแตกต่างไปจากภาษาอังกฤษ หรือภาษาจีน เนื่องจากคำไทยหนึ่งคำเกิดการการประสมกันของอักษรไทยหลายตัวเข้าด้วยกัน การเขียนติดกันไปอาจทำให้การแยกแยะคำมีปัญหา ดังนั้นในการแยกแยะระดับย่อยของคำสามารถนำหลักเกณฑ์ที่เรียกว่าอักขระวิธีมาใช้ให้เป็นประโยชน์

ปัจจุบันคำที่มาจากภาษาต่างประเทศที่ถูกนำมาใช้ร่วมกับภาษาไทยมีเป็นจำนวนมากขึ้น และคำเหล่านั้นนอกจากจะไม่ปรากฏในพจนานุกรมแล้ว หลายคำที่พบมีลักษณะการรวมกันของอักษรไทยที่แตกต่างออกไป เช่น เพาน์ด ฟิล์ม การ์ด เลาน์จ เพื่อให้กฎครอบคลุมและตัดคำได้อย่างมีประสิทธิภาพจึงได้เพิ่มกฎที่จัดการในส่วนนี้ไว้ด้วย

ในส่วนเนื้อหาส่วนที่ 2 จะกล่าวถึงลักษณะโครงสร้างส่วนประกอบของคำในภาษาไทย ส่วนที่ 3 คือวิธีการตัดคำภาษาไทยที่ใช้เป็นหลักพร้อมทั้งเปรียบเทียบถึงข้อดีข้อเสียของทั้งสามวิธี เนื้อหารายละเอียดของการตัดคำด้วยกฎที่นำเสนอในงานวิจัยนี้ จะอยู่ภายในส่วนที่ 4 และส่วนที่ 5 โดยวิธีการพัฒนาและการวัดผลการทดลองจะอธิบาย

ไว้ในเนื้อหาส่วนที่ 6 พร้อมทั้งแสดงตัวอย่างขั้นตอนจากการนำกฎที่นำเสนอไปใช้ไว้งานจริงในส่วนที่ 7 ส่วนสุดท้ายคือบทสรุปสำหรับการใช้กฎที่นำเสนอตัดคำในเอกสารไทย

2. ลักษณะของภาษาไทย

ประโยคภาษาไทยประกอบไปด้วยคำ และคำในภาษาไทยก็ประกอบไปด้วยส่วนต่างๆ ซึ่งสามารถแบ่งได้เป็นแบบสามส่วน สี่ส่วน และห้าส่วน สามส่วนได้แก่ พยัญชนะ สระ และวรรณยุกต์ เช่นคำว่า กา ก่า ก้า ก๊า ก๋า เป็นต้น โดยคำว่า กา มีไม่มีรูปวรรณยุกต์ แต่มีเสียงวรรณยุกต์สามัญ ส่วนแบบสี่ส่วนจะเพิ่มเติมตัวสะกดเข้ามา เช่น กาย บิน รวม และสุดท้ายแบบห้าส่วนจะเพิ่มในส่วนของการันต์ได้แก่คำว่า การณ์ จลน์ เป็นต้น หลักการเหล่านี้จะเรียกว่าอักขระวิธี [4] การประสมกันระหว่างตัวอักษรก็มีหลักการอีกมากมาย อันได้แก่ การแบ่งอักษรไทยทั้ง 44 ตัวออกเป็นมาตรา คือ อักษรสูง อักษรกลาง และอักษรต่ำ การวางตำแหน่งของสระในคำ อักษรบางตัวที่ไม่นำไปเป็นตัวสะกด จากหลักการที่มีอยู่นี้ค่อนข้างแน่นอนพอสมควรในการนำไปแยกแยะคำไทยแต่ละคำ แต่ก็ยังไม่เพียงพอ เพื่อความถูกต้องยิ่งขึ้นจึงมีการพัฒนาและปรับปรุงกฎเพื่อให้ได้ความถูกต้องเพิ่มขึ้น

3. การตัดคำ

การใช้อักขระวิธีในการตัดคำสามารถทำได้ในระดับหนึ่งเนื่องจากคำบางคำเป็นคำที่เลียนเสียงจากภาษาต่างประเทศ ดังนั้นอาจมีการประสมคำนอกเหนือไปจากอักขระวิธี จึงมีการพัฒนาวิธีการต่างๆ ในการตัดคำในเอกสารไทย เพื่อให้ได้ความถูกต้องสูงสุด วิธีการหลักสำหรับตัดคำในเอกสารไทยมีดังนี้

3.1 การใช้กฎ

ลักษณะของการใช้กฎเพื่อตัดคำในภาษาไทย จะใช้ไวยากรณ์ทางภาษา โดยภาษาไทยจะแบ่งตัวอักษรเป็นหมวดหมู่ตามลักษณะการใช้งาน ได้แก่ กลุ่มพยัญชนะ กลุ่มสระ กลุ่มวรรณยุกต์ กลุ่มตัวเลขและกลุ่มตัวอักษรพิเศษ ขั้นตอนการตัดพยางค์จะทำจากซ้ายไปขวาเป็นส่วนใหญ่ ส่วนคำที่ไม่เป็นไปตามกฎที่สร้างไว้จะถูกเก็บไว้ในแฟ้มข้อมูล การวิเคราะห์โดยการหาขอบเขตหน้า (Front boundary recognition rule) และกฎการหาขอบเขตหลัง (Tail boundary recognition rule) [7] ได้เสนอกฎที่ได้จากคุณสมบัติการนำไปประสมกับอักษรไว้ในกฎกลุ่ม A และคุณสมบัติการนำสระไปไว้ในกฎกลุ่ม B การใช้กฎในการตัดคำนี้ยังคงประสบปัญหาการหาขอบเขตของคำ เนื่องจากคำหนึ่งคำอาจประกอบไปด้วยพยางค์เดียวหรือหลายพยางค์ จึงต้องมีการนำวิธีการอื่นเข้ามาในการตัดคำนอกเหนือไปการการตัดคำด้วยกฎเพียงอย่างเดียว

3.2 การใช้พจนานุกรม

การนำพจนานุกรมมาใช้ จะทำให้ผลลัพธ์ที่ได้อยู่ในระดับคำ โดยมีหลักการว่าให้ทำการตรวจสอบสายอักขระ (String) ซึ่งเป็นชุดของตัวอักษรที่ได้จากเอกสาร จากนั้นจะนำอักษรแต่ละตัวไปค้นหาคำจากพจนานุกรม หากพบคำในพจนานุกรมที่สามารถเป็นคำในสายอักขระนั้นได้มากกว่าหนึ่งคำ จะทำการเลือกคำที่ยาวที่สุด (Longest matching) หากอักษรตัวต่อมาไม่สามารถพบคำที่ตรงกับที่ในพจนานุกรมมีอยู่จะทำการย้อนกลับไปเลือกคำที่สั้นกว่าแทนเรียกวิธีการนี้ว่าวิธีการย้อนรอย (Back tracking) [6] นอกจากนี้ยังมีการนำเสนอวิธีแยกพยางค์ด้วยกฎก่อนแล้วจึงใช้พจนานุกรมเพื่อตัดและรวบรวมให้เป็นคำ [1]

โดยที่เทคนิคนี้จะใช้โปรแกรมเล็กซ์ (Lex) [1] สร้างกลุ่มตัวอักษร (Token) ที่ยาวที่สุดก่อนโดยไม่รวมตัวสะกด เนื่องจากตัวสะกดที่อยู่ท้ายคำอาจมีโอกาสเป็นพยัญชนะต้นของคำถัดมาได้

จากนั้นจึงวิเคราะห์อีกครั้งโดยพิจารณาตัวสะกดร่วมด้วยพร้อมกับการใช้พจนานุกรม ภาพรวมของระบบการตัดคำที่นำเสนอปรากฏในรูปที่ 1

3.3 การใช้คลังข้อความ

การใช้คลังข้อความ (Corpus) ในการตัดคำในเอกสารไทยเป็นการนำค่าทางสถิติมาร่วมพิจารณา เช่นค่าสถิติการ

ใช้คำ ค่าสถิติหน้าที่ของคำ งานวิจัยที่ใช้คลังข้อความมาใช้ในการตัดคำมีจุดประสงค์เพื่อเพิ่มความถูกต้องในการตัดคำและลดคำกำกวม ยกตัวอย่างเช่น ที่อยู่ ที่ตั้ง (ซึ่งอาจเป็นได้ทั้ง ที่ อยู่, ที่ ตั้ง) จะนำคำที่ผ่านการตัดคำมาระดับหนึ่งแล้วอาจเป็นจากการใช้กฎหรือการใช้พจนานุกรมที่กล่าวมาข้างต้น มาผ่านการวิเคราะห์ด้วยคลังข้อความ ความรู้ภายในคลังข้อความอาจเป็นค่าสถิติหรือลักษณะไวยากรณ์ เป็นต้น

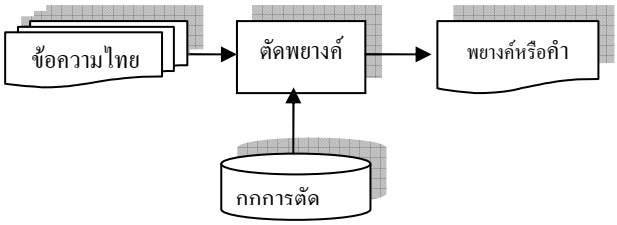
จากวิธีการสามวิธีหลักที่กล่าวมาสามารถสรุปเป็นข้อดีและข้อด้อยดังตารางที่ 1

ตารางที่ 1 ข้อดีข้อเสียของการตัดคำด้วยวิธีต่างๆ

	ความเร็ว	ความถูกต้อง		ขนาดของคลังข้อความ	การตัดคำที่ไม่มีอยู่ในพจนานุกรม
		ระดับพยางค์	ระดับคำ		
กฎ	///	//	/	/	//
พจนานุกรม	//	//	//	//	/
คลังข้อความ	/	//	///	///	/

///มาก // ปานกลาง / น้อย

4. วิธีการตัดคำโดยใช้กฎที่นำเสนอ



รูปที่ 1 ภาพรวมของระบบการตัดคำด้วยกฎ

ลักษณะประโยคหรือวลีของภาษาไทยเกิดจากการรวมกันของคำ ซึ่งแต่ละคำอาจเป็นคำเดี่ยวหรือเกิดจากการรวมกันของคำคำอื่น เมื่อผ่านขั้นตอนแรกเพื่อคัดแยกข้อมูลจากเอกสารให้เป็นอนุประโยค ดังนั้นหากแทนค่า D เป็นเอกสารที่ต้องการนำมาแยกคำและ s[i] เป็นอนุประโยคหรือประโยคที่ได้จากการแบ่งครั้งแรก ณ ตำแหน่งเว้นวรรคหรืออักขระพิเศษที่กำหนดไว้จากเอกสาร D ดังสมการ (1)

$$Document \quad D = \sum_{i=1}^{N} s[i] \qquad (1)$$

คำที่อยู่ในประโยค s[i] ให้แทนด้วย w[i,j] และพยางค์ที่ประกอบขึ้นเป็น w[i,k] แทนด้วย c[i,j] ภายใน s[i] จะประกอบไปด้วย Character c[i,j] ซึ่งเป็นอักขระต่างๆ ใน s[i] ซึ่ง i มีค่าตั้งแต่ 1 ถึง N และ j มีค่าตั้งแต่ 1 ถึง L โดยที่ L คือความยาว String s[i] c[i,j] อาจเป็นอักขระไทยหรือไม่ก็ได้ ดังนั้นการพิจารณาคำไทยจึงต้องตรวจสอบว่า c[i,j] เป็นอักษรไทยที่เป็นส่วนหนึ่งของคำไทยได้หรือไม่ นั่นคือ character in Thai ct[i,j] และWord w[i,k] คือคำไทยที่ทำการตัดออกมาได้จากประโยคที่ i โดยที่ w[i,k] เกิดจากการรวมกันของ ct[i,j] โดยที่ค่า k และ j มีค่าระหว่าง 1 กับ L และ i มีค่าระหว่าง 1 กับ N ซึ่งจะได้ตามดังสมการ(2)-(3)

$$s[i] = \sum_{j=1}^{L} c[i,j] \qquad (2)$$

$$w[i,j] = \sum_{j=first}^{end} ct[i,j] \qquad (3)$$

โดยที่ $1 \le first \le end \le L$

และ $ct[i,j] \in c[i,j]$

4.1 กฎที่นำเสนอ

เมื่อนำ ct[i,j] มาพิจารณาโดยใช้กฎจะได้ผลลัพธ์ในระดับพยางค์ จากกฎที่มีอยู่เดิม [1][7] ได้เพิ่มกฎสำหรับการตัดคำที่มีความเกี่ยวข้องกับภาษาต่างประเทศและคำที่อยู่นอกเหนือจากการประสมตามอักขระวิธี

กฎการรวมกันของต่างภาษา

$$s[i] = w[i,j]$$

ถ้าหาก $c[i,j] \in w[i,j]$ และ $c[i,j] \ne ct[i,j]$

ยกตัวอย่างเช่น ส-8 3กช6 หรือ ม.6/5

กฎการเพิ่มตัวสะกด คำที่มาจากภาษาต่างประเทศบางคำมีการสะกดที่แตกต่างไปจากอักขระวิธีที่มีอยู่แต่เดิม การมีตัวสะกดมากกว่าหนึ่ง หรือพบว่าสระที่เป็นตัวสุดท้ายของคำเสมอนั้นไม่ใช่เสมอไป เช่น เลาน์จ เพาว์ เป็นต้น

4.2 การค้นหาจากพจนานุกรม

การค้นหาคำศัพท์จากพจนานุกรมจะนำ w[i,j] ที่ได้จากการตัดคำด้วยกฎที่นำเสนอ เพื่อรวมกันให้เป็นคำศัพท์ที่มีอยู่ในพจนานุกรม การค้นหาคำหากพบมากกว่าหนึ่งคำปรากฏอยู่จะใช้เทคนิค Longest Matching เพื่อเลือกคำศัพท์ที่ยาวกว่า

5. วิธีการพัฒนาและวัดผลการทดลอง

การพัฒนาการตัดคำภาษาไทยโดยใช้กฎที่นำเสนออยู่บนมาตรฐาน ANSI C บนระบบปฏิบัติการ Windows ข้อมูลที่ถูกนำมาทำการตัดคำถูกนำมาจากเอกสารที่พบเห็นทั่วไปได้แก่นิตยสารและหนังสือพิมพ์ รวมทั้งบทความทางวิชาการที่มีทั้งคำไทยและคำที่มาจากภาษาต่างประเทศ คำที่ได้จากเอกสารจะถูกนำไปแบ่งครั้งแรกด้วยช่องว่างระหว่างประโยค และนำไปแยกในระดับพยางค์โดยการใช้กฎ ผลที่ได้จากการแบ่งด้วยกฎจะถูกเก็บเป็นสองส่วน คือ ส่วนที่ต้องนำไปค้นหาต่อในพจนานุกรมและไม่ต้องนำไปค้นหาต่อ

คำที่ไม่ต้องนำไปค้นหาต่อได้แก่คำดังนี้ คำปรากฏสัญลักษณ์ - / , . คำที่มีตัวเลขและคำที่มีตัวอักษรต่างประเทศ อยู่ระหว่างประโยคที่ถูกแบ่งครั้งแรกด้วยช่องว่าง ซึ่งให้ถือเป็นคำสมบูรณ์แล้ว นอกจากคำเหล่านี้คำอื่นๆ ที่ได้จากการแบ่งด้วยกฎจะนำไปค้นหาในพจนานุกรมเพื่อทำการแบ่งในระดับคำต่อไป

การวัดผลการทดลองใช้ค่าความถูกต้องของคำ (Word validity) หน่วยวัดประสิทธิภาพคือสัดส่วนของคำที่ตัดได้ถูกต้องต่อจำนวนคำที่ตัดออกมาได้

6. ตัวอย่างการตัดคำจากวิธีที่นำเสนอ

จากการวิธีการตัดคำโดยใช้กฎที่นำเสนอขึ้นซึ่งได้กล่าวไว้ในส่วนที่สี่และห้า ตัดคำจากเอกสารไทยที่มีส่วนผสมระหว่างคำไทยและคำจากต่างประเทศ จากประโยคตัวอย่างจะแสดงขั้นตอนการทำงานของการตัดคำไว้ดังในตารางที่ 2 นี้

ตารางที่ 2 ตัวอย่างจากการใช้กฎที่นำเสนอ

ประโยคตัวอย่าง	“เขานั่งที่เลาน์จเพื่อรอเพื่อน ก่อนจะนั่งรถทะเบียน กข3477 ไปที่สนามบิน”
ขั้นที่ 1	เขานั่งที่เลาน์จเพื่อรอเพื่อน ก่อนจะนั่งรถทะเบียน กข3477 ไปที่สนามบิน
ขั้นที่ 2	เขา นั่ง ที่ เลาน์ จ เพื่อ รอ เพื่อน
ขั้นที่ 3	เขา นั่ง ที่ เลาน์จ เพื่อ รอ เพื่อน
ขั้นที่ 4	ก่อน จะ นั่ง รถ ทะ เบียน
ขั้นที่ 5	ก่อน จะ นั่ง รถ ทะเบียน
ขั้นที่ 6	ไป ที่ ส นาม บิน
ขั้นที่ 7	ไป ที่ ส นาม บิน
ผลลัพธ์	เขา นั่ง ที่ เลาน์จ เพื่อ รอ เพื่อน ก่อน จะ นั่ง รถ ทะเบียน กข3477 ไป ที่ สนาม บิน

จากตารางที่ 2 จะข้ามคำว่า กข3477 ซึ่งเป็นคำที่มีตัวเลขอยู่ในประโยคที่แยกออกมาครั้งแรกด้วยช่องว่าง คำนี้ให้ถือเป็นคำสมบูรณ์เนื่องจากเป็นคำที่ไม่พบในพจนานุกรมและการเขียนอักษรต่างภาษาอยู่ติดกันให้ถือว่าทั้งคำเป็นคำคำเดียว

7. เปรียบเทียบผลการตัดคำด้วยกฎเดิมและกฎที่นำเสนอ

การตัดคำนี้ได้นำเอกสาร 3 ประเภทได้แก่ บทความทางวิชาการ นิตยสารและหนังสือพิมพ์ ซึ่งแต่ละประเภทมีจำนวน 10 เอกสาร แต่ละเอกสารมีขนาดประมาณ 600-1000 คำ ค่าความถูกต้องของคำจากการใช้กฎเดิมและกฎที่นำเสนอในการตัดคำระดับพยางค์และระดับคำ ได้ถูกแสดงไว้ ตามตารางที่ 3

ตารางที่ 3 เปรียบเทียบผลการตัดพยางค์และคำที่ถูกต้องที่มาได้จากกฎเดิมและกฎที่นำเสนอ

เอกสารที่ใช้ทดสอบ	ระดับพยางค์ %		ระดับคำ %	
	แบบเดิม	ปรับปรุง	แบบเดิม	ปรับปรุง
บทความ	90.02	92.01	85.60	87.36
นิตยสาร	92.97	95.11	88.45	88.98
หนังสือพิมพ์	90.66	90.76	85.72	85.73

จากตารางที่ 3 จะเห็นได้ว่าการใช้กฎที่นำเสนอจะได้ค่าความถูกต้องของคำที่ได้สูงกว่าการใช้กฎแบบเดิม ซึ่งกฎเดิมและกฎที่นำเสนอใช้ข้อมูลจากพจนานุกรมเดียวกันเพื่อเข้ามาสนับสนุนการตัดในระดับคำ เนื่องจากการตัดคำในระดับคำมีความซับซ้อนกว่าระดับพยางค์ ค่าความถูกต้องของคำโดยเฉลี่ยจึงมีค่าต่ำกว่าการตัดระดับพยางค์ ถึงแม้ว่าเวลาที่ใช้ในการตัดคำโดยการใช้กฎที่นำเสนอใช้เวลามากกว่าเวลาที่ใช้โดยการใช้กฎเดิมแต่ว่าไม่ได้แตกต่างกันมากนัก

8. สรุป

จากการเปรียบเทียบผลการตัดคำในเอกสารไทยที่ผ่านมาทำให้สรุปได้ว่าภาษาไทยมีลักษณะซับซ้อน การปรับปรุงกฎเพื่อให้มีความยืดหยุ่นของโครงสร้างคำไทยทำให้การตัดคำด้วยกฎสามารถลดปัญหาคำที่มาจากภาษาต่างประเทศได้ส่วนหนึ่งอีกทั้งการนำพจนานุกรมเข้ามาช่วยเสริมนอกเหนือไปจากการตัดคำด้วยกฎเพียงอย่างเดียวทำให้ได้ผลลัพธ์จากการตัดคำในเอกสารไทยถูกต้องยิ่งขึ้นในระดับพยางค์และคำ ซึ่งในอนาคตการตัดคำโดยการใช้กฎที่นำเสนอนี้จะนำไปใช้กับเอกสารที่มีขนาดใหญ่ขึ้นและประเภทเอกสารที่นำมาทดสอบมีความแตกต่างทางด้านเนื้อหาหลากหลายยิ่งขึ้น

เอกสารอ้างอิง

[1] ดวงแก้ว สวามิภักดิ์, *การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์*: มหาวิทยาลัยธรรมศาสตร์,2533.

[2] บรรจบ พันธุเมธา, *ลักษณะภาษาไทย* กรุงเทพ: สำนักพิมพ์มหาวิทยาลัยรามคำแหง. 1-45. 2540.

[3] พิสิทธิ์ พรมจันทร์, *การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย,* จุฬาลงกรณ์มหาวิทยาลัย. 2540.

[4] พระยาอุปกิตศิลปสาร, *หลักภาษาไทย*, กรุงเทพฯ : โรงพิมพ์ไทยวัฒนาพานิช. 18-28. 2539.

[5] ยืน ภู่วรวรรณ, "การวิเคราะห์ข้อมูลคำไทย ".

[6] ยืน ภู่วรวรรณ และ วิวรรธ อิ่มอารมณ์, "การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี". รายงานการประชุมวิชาการวิศวกรรมไฟฟ้าครั้งที่ 9, 2529.

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

[8] หัชทัย ชาญเลขา, อัศนีย์ ก่อตระกูล. การสกัดนิพจน์ระบุนามในภาษาไทยโดยใช้แมกซิมัมเอนโทรปีโมเดลและอิงความรู้. หน่วยปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติและเทคโนโลยีสารสนเทศอัจฉริยะ. มหาวิทยาลัยเกษตรศาสตร์.

[9] B. Kijsirikul. "Comparing Winnow and RIPPER in Thai Named-Entity Identification", Chulalongkorn.

[10] C. Kooptiwoot. "Segmentation of Ambiguous Thai Words by Inductive Logic Programming". Chulalongkorn. 1999.

[11] D. D. Plamer. "A Trainable Rule-based Algorithm for Word Segmentation".

[12] P. Charoenpornsawat, B. Kijsirkul, "Feature-based Proper Name Identification in Thai", Chulalongkorn. 1998.

[13] P. Charoenpornsawat, B. Kijsirkul, S. Meknavin. "Feature-based Thai Unknown Word Boundary Identification Using Winnow", Chulalongkorn. 1998.

[14] T. Pongthai, V. Sornlertlamvanish. "Grapheme to Phoneme for Thai", NECTEC.

[15] T. Theeramunkong, V. Sornlertlamvanich, T. tanhermhong, W. Chinnan. "Character Cluster Based Thai Information Retrieval", National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA).

[16] V. Sornlertlamvanich, T. Potipiti, T. Charoenporn. "Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm". National Electronics and Computer Technology Center (NECTEC).

[17] V. Tesprasit, P. Charoenpornsawat , V. Sornlertlamvanich. "Learning Pharse Break Detection in Thai Text-to-Speech". EUROSPEECH, 2003.

[18] W. Arronmanakun. "Collocation and Thai Word Segmentation".