บทที่ 5

โปรแกรมตัดคำภาษาไทยที่ใช้ในปัจจุบัน

โปรแกรมตัดคำภาษาไทยได้มีการพัฒนาและใช้งานกว้างขวางพอสมควรทั้งภาคเอกชนและ ห้องวิจัยของทางราชการทั่วไป ในบทนี้จะทำการรวบรวมคุณลักษณะ เทคนิคของโปรแกรมและอัลกอริธึมตัดคำที่ใช้ในปัจจุบันเหล่านี้

1 โปรแกรมตัดคำภาษาไทยแบบย้อนกลับ (Backtracking algorithm)
2 โปรแกรมตัดคำภาษาไทยของบริษัทไมโครซอฟท์ (Microsoft Windows 95 Thai Edition)
3 โปรแกรมตัดคำภาษาไทยแบบการเทียบคำที่ยาวที่สุด (Longest pattern-matching)
4 โปรแกรมตัดคำภาษาไทยแบบการเทียบคำที่สั้นที่สุด (Shortest pattern-matching)
5 โปรแกรมตัดคำภาษาไทยแบบที่ใช้ความถี่ของการใช้คำ (Word usage frequency)
6 โปรแกรมตัดคำภาษาไทยแบบที่ใช้พจนานุกรมลดความกำกวม (Ambiguity dictionary)
7 โปรแกรมตัดคำภาษาไทยแบบที่เลือกประโยคที่มีจำนวนคำน้อยที่สุด (Maximal pattern-matching)

5.1 โปรแกรมตัดคำภาษาไทยแบบย้อนกลับ (Backtracking algorithm)

โปรแกรมตัดคำภาษาไทยนี้ได้พัฒนาโดยสัมพันธ์ ระรื่นรมย์ ซึ่งเป็นโปรแกรมตัดคำแบบที่ใช้พจนานุกรมช่วยโดยใช้โครงสร้างข้อมูลแบบมัลติเวย์ลิงค์ทรัย (Multiway Linked Trie) ที่จะทำให้ผลตอบสนองต่อการตัดคำกับเอกสารภาษาไทยขนาดใด ๆ เป็นแบบเชิงเส้น และใช้เทคนิคการแก้ไขความผิดพลาดแบบย้อนกลับ (Back Tracking) ซึ่งโปรแกรมตัด คำภาษาไทยนี้ปัจจุบันได้ใช้อยู่ในโปรแกรม ประมวลผลคำจุฬาเวิร์ด (CU Writer Version 1.6)

5.2 โปรแกรมตัดคำภาษาไทยของบริษัทไมโครซอฟท์ (Microsoft Windows 95 Thai Edition)

บริษัทไมโครซอฟท์ได้พัฒนาโปรแกรมตัดคำภาษาไทยสำหรับผู้ใช้งานได้เรียกใช้งานผ่านเอพีไอภาษาไทย (Thai API call) โดยจะอยู่ในรูปของไดนามิกลิงค์ไลบรารี่ (Dynamic Link Library DLL) ซึ่งจะอยู่ในไฟล์ FTLX041E.DLL ที่มากับซอฟท์แวร์ไมโครซอฟท์วินโดว์ 95 ฉบับภาษาไทย (Microsoft Windows 95 Thai Edition) โดยที่โปรแกรมฟังก์ชั่นจะใช้ชื่อ FindThaiWordBreak() สำหรับการเรียกใช้จะเป็นดังภาคผนวก ก งานวิจัยนี้ได้พบปัญหาในการใช้งานโปรแกรมตัดคำนี้ กล่าวคือเมื่อโปรแกรมทำการตัดคำไปจนถึงถึงประมาณ 30 บรรทัดก็จะ

ไม่สามารถทำงานต่อไปได้ เนื่องจากข้อจำกัดทางเวลาผู้วิจัยจึงเลือกที่จะไม่รวมผลของการวัดประสิทธิภาพจากโปรแกรมตัดคำนี้เข้ามาเปรียบเทียบด้วย

5.3 โปรแกรมตัดคำภาษาไทยแบบการเทียบคำที่ยาวที่สุด (Longest pattern-matching)

งานวิจัยของ Nontarat Thongpumppurksar [10] ได้เสนอแนวความคิดไว้ว่า ตัดคำให้ได้คำที่ยาวที่สุดที่จะรวมขึ้นเป็นประโยคได้ โดยใช้โครงสร้างข้อมูลแบบ ดับเบิลอะเรย์ (double array) เพื่อสร้างทรัย (Trie) สำหรับเป็นพจนานุกรม แต่เพื่อให้การจำลองแนวความคิดนี้ง่ายขึ้น งานวิจัยนี้ได้สร้างโปรแกรมตัดคำขึ้นมาโดยอาศัยแนวความคิดของ Nontarat โดยได้ใช้โครงสร้างข้อมูลแบบลิงค์ลิสต์ที่เรียงลำดับโดยใช้ความยาวของคำ และใช้การค้นหาแบบเรียงลำดับ (Sequentail search) ลักษณะของโหนดในพจนานุกรมจะเป็นดังภาคผนวก ข. การค้นหาคำจากพจนานุกรมจะเป็นภาคผนวก ค. การโหลดพจนานุกรมเข้าสู่โครงสร้างข้อมูล จะเป็นดังภาคผนวก ง.

5.4 โปรแกรมตัดคำภาษาไทยแบบการเทียบคำที่สั้นที่สุด (Shortest pattern-matching)

งานวิจัยเดียวกันกับข้างบนได้เสนอแนวความคิดไว้ว่า ตัดคำให้ได้คำ ที่สั้นที่สุดที่จะรวมขึ้นเป็นประโยคได้ ในการจำลองแนวความคิดนี้ งานวิจัยนี้ได้สร้างโปรแกรมตัดคำขึ้นมาโดยอาศัยแนวความคิดของ Nontarat โดยสามารถใช้โครงสร้างข้อมูลแบบเดียวกับ 5.3 และเทคนิคการค้นหาแบบเดียวกันนั้นได้ แต่เรียงลำดับคำในพจนานุกรมเสียใหม่ โดยเอาคำที่สั้นที่สุดขึ้นก่อน

5.5 โปรแกรมตัดคำภาษาไทยแบบที่ใช้ความถี่ของการใช้คำ (Word usage frequency)

งานวิจัยเดียวกันกับข้างบนและดร.รัตติกร วรากูลศิริพันธุ์ กับ สง่า คงสุพานิช [8] ได้กล่าวถึงแนวทางหนึ่งในการแก้ปัญหาความกำกวมของประโยคภาษาไทยโดยการวิเคราะห์ความถี่ของการใช้คำในชีวิตประจำวัน สำหรับงานวิจัยนี้ได้จำลองวิธีการดังกล่าวโดยใช้การตัดคำแบบเทียบคำที่ยาวที่สุดเพื่อค้นหาความถี่ของการใช้คำในข้อมูลภาษาไทยขนาด 1,387,068 ตัวอักษร แล้วได้จัดเรียงลำดับพจนานุกรมเสียใหม่ตามความถี่ของการใช้คำเพื่อที่จะทำการตัดคำโดยสามารถใช้วิธีการค้นหาแบบเดียวกันกับ 5.3 จากการศึกษาดังกล่าวพบว่า 20 คำแรกที่มีความถี่ในการใช้งานสูงสุดจะเป็นดังตาราง 5.1

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

I47824361

ลำดับ	คำไทย	ความถี่	ลำดับ	คำไทย	ความถี่
1	การ	19,931	11	ทาง	2,835
2	และ	8,239	12	ประเทศ	2,807
3	ใน	7,207	13	มี	2,675
4	ของ	6,576	14	ราย	2,618
5	ที่	5,962	15	ให้	2,338
6	งาน	4,543	16	กับ	2,254
7	ไทย	4,103	17	เป็น	1,969
8	ความ	3,406	18	กิจ	1,960
9	พัฒนา	3,312	19	ศร	1,872
10	ศึกษา	3,146	20	สาร	1,785

ตาราง 5.1 คำภาษาไทย 20 คำแรกที่มีความถี่ในการใช้งานสูงสุด

5.6 โปรแกรมตัดคำภาษาไทยแบบที่ใช้พจนานุกรมลดความกำกวม (Ambiguity dictionary)

จากการทดลองพบว่าโปรแกรมตัดคำที่กล่าวมาแล้วข้างต้นทั้งหมดไม่สามารถแก้ปัญหาคำกำกวมได้ดีพอ จะขอยกตัวอย่างประโยคที่โปรแกรมข้างต้นไม่สามารถตัดคำได้อย่างถูกต้องดังนี้

คณะกรรมการกรมพละศึกษารอยกร่าง
ฉันมารอกราบ
เรือโคลงเพราะโคลงเรือ
ปลานอนตากลม

ได้ทำการเพิ่มคำศัพท์ที่จะใช้แก้ปัญหาคำกำกวมเหล่านี้ลงไปในพจนานุกรมที่มีอยู่เพื่อแก้ปัญหาดังกล่าวโดยจัดให้คำศัพท์เหล่านี้อยู่เป็นลำดับต้น ๆ ในพจนานุกรม จากการทดลองพบว่าสามารถแก้ปัญหาความกำกวมนี้ได้ในระดับที่ดีมากทีเดียว

นอกจากนี้แล้ว งานวิจัยของ Nontarat [10] ได้พูดถึงการตัดคำแบบตัดทุกกรณี (All possible case patern-matching) แต่เนื่องจากการตัดคำดังกล่าวนี้จะได้ประโยคทุกประโยคที่สามารถเป็นไปได้จากการตัดคำในประโยคใด ๆ โดยที่ไม่มีวิธีการเลือกประโยคที่ถูกต้องที่สุดแต่อย่างใด งานวิจัยนี้จึงไม่ได้ทำการวัดประสิทธิภาพของวิธีการนี้

5.7 โปรแกรมตัดคำภาษาไทยแบบที่เลือกประโยคที่มีจำนวนคำน้อยที่สุด (Maximal pattern-matching)

โปรแกรมตัดคำภาษาไทยแบบนี้จะทำการตัดคำออกมาทุกกรณีที่เป็นไปได้แต่จะเลือกประโยคที่ประกอบด้วยจำนวนคำที่น้อยที่สุดรายละเอียดสำหรับวิธีการแบบนี้สามารถดูได้เพิ่มเติมจากงานวิจัยของ Witoon Kanlayanawat, Somchai Prasitjutrakul [14] โปรแกรมตัดคำแบบนี้ผู้วิจัยได้รับความอนุเคราะห์จาก ดร. สุรพันธ์ ที่ห้องวิจัยลิงคส์ (LINKS) ของเนคเทค (NECTEC) ให้ใช้โปรแกรมตัดคำที่ใช้เทคนิคการตัดคำแบบนี้มาร่วมวัดประสิทธิภาพ

ในบทนี้ได้รวบรวมเทคนิคแนวความคิดของโปรแกรมตัดคำภาษาไทยที่มีการคิดค้นและพัฒนาไว้ซึ่งบางโปรแกรมได้มีการใช้งานจริงในปัจจุบันและบางโปรแกรมเป็นแนวทางความคิดโดยถ้าเป็นแบบหลังผู้เขียนได้สร้างต้นแบบจำลองโดยอาศัยแนวความคิดพื้นฐานเหล่านั้น

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย