การตัดคำไทยในระบบแปลภาษา

Word Segmentation for Thai in Machine Translation System

วิรัช ศรเลิศล้ำวาณิช

ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ

กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม

1. บทนำ การตัดคำ หรือการแบ่งข้อความที่ต่อเนื่องกันออกเป็นหน่วยคำๆ (Morpheme) หรือลักษณะของการรู้จำ (Recognition) คำหนึ่งๆในข้อความที่ต่อเนื่องกันนั้นเริ่มจะมีความหมายมากยิ่งขึ้นเป็นลำดับเมื่อมีการนำเอาคอมพิวเตอร์เข้ามาช่วยในการประมวลผลข้อมูลทางภาษามากยิ่งขึ้น ความยากง่ายหรือวิธีการที่จะนำมาใช้ในการตัดคำนั้นต้องขึ้นอยู่กับลักษณะเฉพาะของภาษานั้นๆเป็นอย่างมาก ผลงานวิจัยที่นำมาเสนอนี้ไม่ได้เป็นคำตอบสุดท้ายหรือคำตอบที่ดีที่สุด แต่เป็นวิธีที่ใช้ได้ผลดีและก็กำลังใช้อยู่ในการพัฒนาระบบเครื่องแปลภาษา ผลของการใช้งานจริงได้เป็นตัวกระตุ้นการวิจัยและพัฒนาระบบการตัดคำนี้จนได้เป็น Version 2 ในปัจจุบันนี้

2. จุดประสงค์ของการตัดคำ การตัดคำหรือที่เรียกกันว่า Word Segmentation นั้นคือการแบ่ง String ของตัวอักษรเพื่อหาขอบเขตของแต่ละหน่วยคำ (Morpheme) เนื่องจากว่าโดยปกติทั่วไปแล้ว ภาษาไทยมีการเขียนในลักษณะที่ติดต่อกันโดยไม่มีการใช้เครื่องหมายวรรคตอนใดๆ ยกเว้นแต่มีการเว้นวรรคเป็นระยะๆ เพื่อให้ผู้อ่านได้หยุดพัก และทำความเข้าใจความหมายเป็นตอนๆไปเท่านั้น แม้ว่าการเว้นวรรคในการเขียนบทความไม่ได้มีกฎเกณฑ์ที่ชัดเจนก็ตาม แต่ถ้ามีการใช้การเว้นวรรคด้วยความระมัดระวังแล้วก็จะสามารถช่วยลดความคลุมเครือของคำหรือประโยคได้

ไม่ว่าจะด้วยวิธีการใดก็ตามถ้าหากสามารถรู้เขตแบ่งของแต่ละคำได้แล้ว การจัดการกับข้อความนั้นๆก็จะเป็นไปได้อย่างสะดวกและถูกต้อง ในระบบคอมพิวเตอร์จึงจำเป็นอย่างยิ่งที่จะต้องคำนวณหาขอบเขตของแต่ละคำให้ได้เพื่อที่จะสามารถลดภาระ

ของผู้ใช้หรือเพื่อที่จะให้กระบวนการที่อยู่ในระดับที่ลึกกว่าสามารถทำงานต่อไปได้ ดังเช่น ฟังก์ชั่นการขอบขวา (Word wrapping) ใน Word processor การตรวจคำผิด การค้นหาคำใน text หรือเป็นตัวช่วยในการกำหนดคำเพื่อทำการวิเคราะห์ต่อไปใน ระบบของเครื่องแปลภาษา

3. อัลกอริธึมสำหรับการตัดคำโดยใช้พจนานุกรม การตัดคำคือการหาขอบเขตของหน่วยคำในข้อความที่ต่อเนื่อง ฉะนั้นถ้าหากเก็บทุกคำที่มีอยู่ในภาษาลงในพจนานุกรมทั้งหมด จากนั้นก็ค้นหาและเปรียบเทียบหาคำศัพท์นั้นๆว่ามีอยู่ในพจนานุกรมหรือไม่ เพียงเท่านั้นก็จะสามารถหาขอบเขตของคำแต่ละคำได้ แต่ในความเป็นจริงแล้วไม่สามารถจะกระทำได้ เนื่องจากว่าเป็นไปไม่ได้ที่จะบรรจุคำทุกคำลงในพจนานุกรมได้ทั้งหมด โดยเฉพาะในส่วนที่เป็นวิสามานยนาม (คำนามที่เป็นชื่อเฉพาะ---พระยาอุปกิตศิลปสาร, 1990) ตัวเลข หรือคำที่เกิดขึ้นมาจากการใช้ใหม่ๆ หรือมีการบัญญัติขึ้นมาใหม่ ซึ่งไม่สามารถจะคาดการล่วงหน้าได้ ฉะนั้นการตัดคำถึงแม้ว่าจะอาศัยการเปรียบเทียบคำจากพจนานุกรมก็ตาม ก็จำเป็นที่จะต้องยอมให้มีคำที่ไม่ได้บรรจุไว้ในพจนานุกรมเกิดขึ้นได้เช่นกัน

คำที่จะบรรจุอยู่ในพจนานุกรมไม่จำเป็นที่จะต้องเป็นหน่วยคำที่ย่อยที่สุดที่คงความหมายไว้เสมอไป อาจเป็นคำประสม (แม่น้ำ, ดูแล, ช่างทอง, เป็นต้น) หรือวลี (แสงอาทิตย์---รัศมี+ดวงตะวัน,การแปลภาษาด้วยเครื่องคอมพิวเตอร์, หนีเสือปะจระเข้, เป็นต้น) ก็ได้ ทั้งนี้ก็ขึ้นอยู่กับว่าตำแหน่งของคำในวลีและโครงสร้างทางวากยสัมพันธ์นั้นๆคงที่แน่นอนหรือไม่ และจะสามารถกำหนดความหมายที่ชัดเจนให้กับวลีนั้นๆได้หรือไม่

3.1 กฎทางอักขรวิธี แม้ว่าการเขียนข้อความในภาษาไทยจะไม่มีการเว้นวรรคระหว่างคำ ไม่มีการใช้เครื่องหมายวรรคตอนที่ชัดเจน ไม่มีตัวชี้บ่งหน้าที่ทางไวยากรณ์ (Syntactic Marker---Postposition) ไม่มีการแปรรูป (Inflection) ไม่มีการใช้ตัวอักษรใหญ่-เล็ก เป็นต้น แต่ภาษาไทยก็มีกฎทางอักขรวิธีที่กำหนดลักษณะของการประสมอักษร การเว้นวรรคดังได้กล่าวไว้ในหัวข้อที่ 2. และการขึ้นย่อหน้า ซึ่งทั้ง 3 ลักษณะนี้

จะเป็นตัวช่วยในการชี้บ่งขอบเขตของการพิจารณาในการเปรียบเทียบกับคำในพจนานุกรม

การขึ้นย่อหน้าเป็นตัวชี้บ่งได้ดีถึงการสิ้นสุดของความ

การเว้นวรรคเป็นตัวชี้บ่งถึงความเป็นไปได้ของการสิ้นสุดของคำหรือประโยค

กฎทางอักขรวิธีเป็นตัวชี้บ่งถึงความเป็นไปได้ของการพิจารณาในการที่จะแยกสายอักขระออกเพื่อการพิจารณาคำหรือไม่ ฉะนั้นจึงมีการแบ่งอักขระออกใหม่ดังนี้

อักขระกลุ่มที่ 1 กลุ่ม Non-spacing Character คือกลุ่มของรูปสระ วรรณยุกต์ และเครื่องหมายพิเศษประกอบการเขียน ที่เมื่อประสมเข้ากับพยัญชนะใดๆแล้วไม่ทำให้มีการเคลื่อนขวาของตำแหน่งที่จะเขียนต่อไป อักขระในกลุ่มนี้จะไม่สามารถอยู่เดี่ยวๆ ได้ มีดังนี้คือ ◌ั, ◌ิ, ◌ี, ◌ึ, ◌ื, ◌ุ, ◌ู, ◌็, ◌่, ◌้, ◌๊, ◌๋, ◌์, ◌ำ

อักขระกลุ่มที่ 2 กลุ่มอักขระที่จำเป็นที่จะต้องมีรูปพยัญชนะตามเสมอ มีดังนี้คือ เ◌, แ◌, โ◌, ใ◌, ไ◌

อักขระกลุ่มที่ 3 กลุ่มอักขระที่จำเป็นที่จะต้องมีรูปพยัญชนะอยู่หน้าเสมอ มีดังนี้คือ ◌ะ, ◌า, ◌ำ, ◌ๅ สำหรับ ◌ๅ จะมีพยัญชนะนำหน้าได้อยู่เพียง 2 ตัวคือ ฤ กับ ฦ

อักขระกลุ่มที่ 4 กลุ่มของอักษรที่เป็นตัวการันต์ที่มีไม้ทัณฑฆาตบังคับข้างบน เช่น ย์ ใน ทิพย์, น์ ใน กาญจน์ เป็นต้น เนื่องจากว่าตัวการันต์เป็นพยัญชนะสุดท้ายที่ไม่อ่านออกเสียง ฉะนั้นจะไม่มีการพิจารณาให้เป็นอักษรแรกของคำ

อักขระกลุ่มที่ 5 กลุ่มของอักษรที่เหลือทั้งหมด

เนื่องจากว่าข้อมูลสายอักขระถูกเก็บไว้ในระบบ X-TIS (รหัส 2 ไบท์สำหรับอักษรไทย) ดังนั้นอักขระในกลุ่มที่ 1 จึงไม่ต้องคำนึงถึง เพราะกลายเป็นส่วนหนึ่งของรหัสผสมไปแล้ว

อักขระในกลุ่มที่ 2, 3 และ 4 จะเป็นตัวจำกัดที่ดีที่จะไม่ให้มีการพิจารณาการตัดคำในระหว่างตัวอักษรนั้นๆ

3.2 Longest Matching วิธีการนี้ถือได้ว่าเป็นวิธีทาง Heuristic อันหนึ่ง โดยจะทำการตรวจสอบคำในพจนานุกรมแล้วแยกเป็นคำๆจากซ้ายไปขวา การตรวจสอบจะเริ่มตั้งแต่หน่วยของข้อความที่สั้นที่สุดที่จะสามารถคัดมาได้ เนื่องจากว่าภาษาไทยไม่มีเครื่องหมายบอกจุดสิ้นสุดของประโยค ดังนั้นจากเกณฑ์ทางอักขระดังในข้อที่ 3.1 จะเป็นตัว

บอกถึงขอบเขตของหน่วยข้อความได้ เมื่อตรวจสอบกับคำในพจนานุกรมแล้วถ้าไม่พบก็จะทำการลดความยาวของข้อความลงทีละตัวไปตามเกณฑ์ทางอักขรวิธี เช่น ในข้อความ "ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคัญ" เมื่อไม่ปรากฏในพจนานุกรมแล้วข้อความนี้ก็จะถูกลดลงเหลือ "ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคั" แล้วก็เป็น "ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำ" จนในที่สุดก็จะได้คำว่า "ความก้าวหน้า" เป็นคำแรก ผลของการตัดคำทั้งหมดจะเป็นดังกล่าวในรูปที่ 1.

ส่วนของคำที่ยาวที่สุด	ส่วนที่เหลือ
ความก้าวหน้า	ทางด้านวิทยาศาสตร์มีบทบาทสำคัญ
ทาง	ด้านวิทยาศาสตร์มีบทบาทสำคัญ
ด้าน	วิทยาศาสตร์มีบทบาทสำคัญ
วิทยาศาสตร์	มีบทบาทสำคัญ
มี	บทบาทสำคัญ
บทบาท	สำคัญ
สำคัญ	

รูปที่ 1. แสดงผลของการตัดคำโดยวิธี Longest Matching ภายใต้เงื่อนไขของอักขรวิธี

วิธีการนี้จะให้ความถูกต้องในการตัดคำได้ประมาณ 80% ข้อบกพร่องที่เห็นได้ชัดเจนคือการเลือกขอบเขตของคำที่ยาวเกินไปตั้งแต่แรก จึงทำให้คำที่ตามมาผิดเพี้ยนไป เช่น ข้อความ "กีฬาเป็นการออกกำลังกายอย่างหนึ่ง" จะถูกแบ่งออกเป็น "กีฬา/เป็นการ/ออกกำลัง/กาย/อย่างหนึ่ง." เสมอ เพราะคำที่ยาวที่สุดที่จะพบได้ในพจนานุกรมสำหรับคำที่สองหลังจากที่ได้คำว่า "กีฬา" แล้ว จะเป็นคำว่า "เป็นการ" เสมอ ฉะนั้นจึงไม่สามารถจะแบ่งให้ได้คำว่า "การออกกำลังกาย"ที่มีความหมายที่ถูกต้องได้ในข้อความ

3.3 วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรม (Unknown Word) น้อยที่สุด นี่เป็นวิธีการทาง Heuristic อีกวิธีหนึ่ง และจะช่วยได้มากถ้าข้อความนั้น ประกอบด้วยคำจำนวนมากหรือมีความยาวของข้อความมากพอสมควร วิธีนี้จะทำบน อัลกอริธึม

ของ Longest Matching ในหัวข้อที่ 3.2 อีกทีหนึ่ง โดยจะทำการหาความเป็น ไปได้ทั้งหมดในการตัดคำในข้อความนั้นๆ

Backtracking จะเริ่มกระทำหลังจากที่ได้คำตอบจากวิธี Longest Matching แล้ว โดยจะกระทำไปทีละคำจากซ้ายไปขวา เช่น "กีฬา/เป็นการ/ออกกำลัง/กาย/อย่างหนึ่ง." ซึ่งเป็นผลที่ได้จากการตัดแบบ Longest Matching จะถูก Backtrack ที่คำว่า "กีฬา" แต่เนื่องจากว่าจะไม่เกิดประโยชน์อันใดในการที่จะยอมให้เกิดคำที่ไม่มีในพจนานุกรมอีก ฉะนั้นการ Backtracking จึงสิ้นสุดลงตรงคำว่า "กีฬา" ต่อไปก็ทำการ Backtracking ที่คำว่า "เป็นการ" เพื่อตรวจดูความเป็นไปได้ทั้งหมดในการแบ่งคำ ผลที่ได้คือจะสามารถแบ่งได้เป็น "เป็น/การ...." และการ Backtracking ก็สิ้นสุดลงตรงคำว่า "เป็น"

เมื่อทำการ Backtracking สิ้นสุดลงกับทุกคำแล้ว ก็จะทำการคำนวณหา Cost ให้กับแต่ละ Path ที่เป็นไปได้ โดยบังคับให้มีการเกิด Unknown Word น้อยที่สุด แล้วเรียงผลของการตัดคำที่ได้ใหม่โดยให้คำตอบที่น่าจะเป็นไปได้มากที่สุดหรือถูกต้องมากที่สุดมาในอันดับแรก ผลของการตัดคำและ Cost ที่คำนวณได้แสดงไว้ในรูปที่ 2.

ผลจากการ Backtracking	Cost
กีฬา/เป็น/การออกกำลังกาย/อย่างหนึ่ง.	4
กีฬา/เป็นการ/ออกกำลัง/กาย/อย่างหนึ่ง.	5
กีฬา/เป็นการ/ออก/กำลังกาย/อย่างหนึ่ง.	5
กีฬา/เป็นการ/ออกกำลัง/กาย/อย่าง/หนึ่ง.	6
กีฬา/เป็นการ/ออกกำลัง/กา/ยอ/ย่าง/หนึ่ง.	7
กีฬา/เป็นการ/ออ/[ก]/กำลังกาย/อย่างหนึ่ง.	11
กีฬา/เป็นการ/ออกกำลัง/กาย/อย่า/[ง]/หนึ่ง.	12
กีฬา/เป็นการ/ออ/[ก]/กำลังกาย/อย่าง/หนึ่ง.	12
กีฬา/เป็นการ/ออ/[ก]/กำลังกาย/อย่า/[ง]/หนึ่ง.	18

รูปที่ 2. แสดงผลของการตัดคำโดยเรียงตาม Cost ที่คำนวณได้

4. บทสรุป การตัดคำเป็นกระบวนการเริ่มแรกที่จะนำข้อมูลเข้าสู่ระบบการแปลภาษาด้วยเครื่องคอมพิวเตอร์ และต้องการการประมวลผลที่มีความเร็วสูง วิธีดังกล่าวในงาน

วิจัยนี้แม้ว่าจะไม่สมบูรณ์ในตัวมันเอง กล่าวคือ ไม่สามารถที่จะให้ผลของการตัดคำที่ถูกต้องได้ 100% ก็ตาม เนื่องจากว่าในกระบวนการดังกล่าวไม่ได้ทำการวิเคราะห์ให้ลึกไปจนถึงโครงสร้างทางไวยากรณ์หรือความสัมพันธ์ทางความหมายเลย แต่ในเชิงปฏิบัติ เมื่อใช้หลักการทั้ง 3 ประการประกอบเข้าด้วยกันแล้ว ผลที่ได้จะให้ความถูกต้องได้สูงเกิน 90% เลยทีเดียว และในขณะเดียวกันก็ให้ผลของการตัดคำได้เร็วพอสำหรับการใช้งานในขั้นต่อๆไปได้ สำหรับระบบที่ต้องการความถูกต้องในการหาขอบเขตของคำสูงมากๆ ดังเช่น ในระบบการแปลภาษาด้วยเครื่องคอมพิวเตอร์ ระบบการค้นหาคำศัพท์ ระบบKWIC เป็นต้น ก็สามารถที่จะใช้ผลที่ได้ลองลงมาสำหรับการประมวลต่อไปได้

เอกสารอ้างอิง

1. Tanaka Hozumi, 1989, Fundamental of Natural Language Analysis, Sangyotosho, in Japanese.
2. พระยาอุปกิตศิลปสาร, 1990, หลักภาษาไทย, ไทยวัฒนาพานิช