

บทที่ ๒
การจัดคำภาษาไทย



ในบทนี้จะได้กล่าวถึงประเภทของวิธีการจัดคำภาษาไทย และลักษณะของงานที่สามารถนำการจัดคำเหล่านี้ไปใช้ได้ โดยจะพิจารณาเฉพาะหน้าที่การจัดคำแต่เพียงอย่างเดียว กล่าวคือ การวิเคราะห์ว่าตำแหน่งใดในเนื้อความเป็นจุดแยกระหว่างคำ (หรือพยางค์) ซึ่งผลลัพท์ที่ได้จะเป็นเพียงตำแหน่งของจุดตัดคำ ทั้งนี้ได้หมายรวมถึงการแทรกอักขระพิเศษใดๆ ลงในเนื้อความเพื่อระบุจุดที่มีการตัดคำ

2.1 ประเภทของขั้นตอนวิธีการจัดคำภาษาไทย

ขั้นตอนวิธีการจัดคำภาษาไทยที่ได้มีการพัฒนาขึ้นมา นั้น จะใช้วิธีการหลักๆ อยู่

2 วิธี คือ (Charnyapornpong, 1983)

1. ใช้พจนานุกรม (dictionary) วิธีนี้จะทำการเก็บคำที่เป็นไปได้ทั้งหมดไว้ การแบ่งแยกคำจะใช้การเปรียบเทียบ คำในข้อความกับคำในพจนานุกรม เพื่อหาจุดตัดคำที่ถูกต้อง

2. ใช้กฎ ซึ่งได้จากการวิเคราะห์โครงสร้างของคำโดยทั่วไปในภาษาไทย

ขั้นตอนวิธีการตัดคำทั้งสองวิธีนี้ทั้งข้อดีและข้อเสียต่างกัน กล่าวคือ การใช้พจนานุกรมจะทำให้ได้เปอร์เซ็นต์ความถูกต้องของการตัดคำ สูงกว่าการใช้กฎ แต่จะต้องใช้พจนานุกรมขนาดใหญ่มาก ซึ่งเป็นการสิ้นเปลืองและทำงานได้ช้ากว่า ขั้นตอนวิธีที่เหมาะสม อาจจะเป็นการผสมผสานการทำงานของทั้งสองวิธีก็ได้

2.2 ประโยชน์ของขั้นตอนวิธีการจัดคำภาษาไทย

ขั้นตอนวิธีการตัดคำภาษาไทย เป็นพื้นฐานที่สำคัญในการพัฒนาโปรแกรมใช้งาน หรือขั้นตอนวิธีที่เกี่ยวข้องกับภาษาไทย อื่นๆ ได้มีผู้วิจัยงานด้านต่างๆ ซึ่งต้องอาศัยขั้นตอนวิธีการตัดคำภาษาไทย ดังนี้

1. การแปลภาษาด้วยเครื่อง (machine translation) โดยเฉพาะ

การแปลจากภาษาไทยไปเป็นภาษาอื่น ต้องทำการแยกคำให้ได้ก่อนจึงจะสามารถหาความหมายของคำได้

2. การทำดัชนี (indexing) ลัดดา ยินดีมาก โกรติ (2531) ได้ทำการวิจัยเกี่ยวกับ การทำดัชนีคำสำคัญจากเอกสาร กล่าวไว้ว่า "การแบ่งแยกคำเป็นขั้นตอนที่สำคัญในการทำดัชนีคำสำคัญเพื่อที่จะสามารถแยกคำสำคัญออกจากข้อความได้"

3. การจัดเรียงคำไทย การจัดเรียงคำไทยตามพจนานุกรมนั้น ต้องมีขั้นตอนวิธีที่ถูกต้อง เนื่องจากไม่สามารถเรียงตามรหัสแอสกี (ASCII) ได้เหมือนในภาษาอังกฤษ ยืน ภู่วรรณ (2530) ได้กล่าวไว้ว่า "ไม่มีอัลกอริทึมการจัดเรียงคำไทยใด ที่เรียงได้ถูกต้องเหมือนพจนานุกรม ถ้าไม่ใช้อัลกอริทึมการแบ่งแยกพยางค์ร่วม"

4. โปรแกรมประมวลผลคำภาษาไทย จะใช้ขั้นตอนวิธีการตัดคำภาษาไทยในการจัดหลังให้ด้านขวาของบรรทัดตรงกัน

