

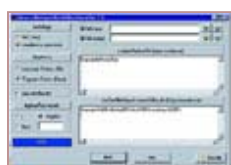


โปรแกรมตัดคำภาษาไทยคืออะไร

เครื่องแปลภาษาคืออะไร

โปรแกรมรู้จำอักขระไทยด้วยแสงคืออะไร

### โปรแกรมตัดคำภาษาไทยคืออะไร



[ ขยายดูภาพใหญ่ ]

ความมุ่งหวังที่จะให้คอมพิวเตอร์สามารถประมวลผลภาษาไทยได้อย่างมีประสิทธิภาพนั้นมีปัญหาเบื้องต้นคือ ลักษณะการเขียนภาษาไทยจะเขียนติดต่อกันเป็นสายอักขระ โดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำดังเช่น ภาษาอังกฤษ ซึ่งเป็นอุปสรรคอย่างหนึ่งที่ต้องการการศึกษาวิจัยและพัฒนา เพื่อให้คอมพิวเตอร์สามารถคำนวณ เพื่อแบ่งสายอักขระไทยออกเป็นคำ ๆ ซึ่งจะส่งผลให้การทำงานของคอมพิวเตอร์ ในการค้นหาคำใด ๆ ทำได้อย่างถูกต้องและแม่นยำรวมถึงการจัดขอบขวาในโปรแกรมประมวลผลคำ (Word processor) เป็นต้น

#### ทำไมคอมพิวเตอร์ต้องรู้จักคำในภาษาไทย

ปัจจุบันนี้ มีการนำคอมพิวเตอร์เข้าไปใช้งานในด้านต่าง ๆ เป็นจำนวนมาก งานทางด้านจัดการกับภาษาไทย ได้แก่ การแปลงข้อความเป็นเสียง (Text - To - Speech) การตรวจตัวสะกด (Spell checking) การแปลภาษาด้วยเครื่องคอมพิวเตอร์ (Machine Translation) ฯลฯ ก็เป็นงานอีกด้านหนึ่งที่ได้นำคอมพิวเตอร์เข้าไปช่วยในการทำงาน สิ่งแรกที่จะต้องจัดการคือ ต้องทำให้คอมพิวเตอร์รู้จักคำในภาษาไทยให้ได้ เช่น ถ้ามีข้อความว่า

"ฉันนั่งตากลมที่หน้าบ้าน" จะต้องทำให้คอมพิวเตอร์รู้ว่า เป็น

"ฉัน นั่ง ตาก ลม ที่ หน้า บ้าน" ไม่ใช่ "ฉัน นั่ง ตา กลม ที่ หน้า บ้าน"

ลักษณะการเขียนภาษาไทยนั้น จะเขียนติดต่อกัน โดยไม่มีเครื่องหมายวรรคตอนคั่นระหว่างคำดังเช่นในภาษาอังกฤษ ซึ่งใช้ช่องว่าง (Space) คั่นระหว่างคำ จึงทำให้การโปรแกรมให้คอมพิวเตอร์รู้จักคำในภาษาไทยนั้นมีความยุ่งยากกว่าในภาษาอังกฤษ ดังนั้น เราจึงได้พัฒนาและค้นหาวิธีการต่าง ๆ ที่จะช่วยให้คอมพิวเตอร์รู้จักคำในภาษาไทย ซึ่งในปัจจุบันก็มีอยู่หลายแนวคิด เช่น

#### ๑. วิธีการตัดคำแบบยาวที่สุด (Longest Matching)

เมื่อเราต้องการโปรแกรมให้คอมพิวเตอร์รู้จักคำในภาษาไทย เราก็จะค้นหาคำ โดยเริ่มจากตัวอักษรซ้ายสุดของข้อความนั้น ไปยังตัวอักษรถัดไป จนกว่าจะพบว่าเป็นคำที่มีอยู่ในพจนานุกรมหลังจากนั้น ก็ค้นหาคำถัดไปจนกว่าจะจบข้อความ ในกรณีที่เราพบว่าเป็นคำในพจนานุกรมจากจุดเริ่มต้นเดียวกัน เราจะเลือกคำที่ยาวที่สุด ตัวอย่างเช่น

การแบ่งคำในประโยค "ฉันนั่งตากลมที่หน้าบ้าน" จะเริ่มจากตัวอักษร ฉ และคำแรกที่แบ่งได้คือ "ฉัน" หลังจากนั้น ก็ค้นหาตัวอักษรถัดไปและนำมาเปรียบเทียบกับคำในพจนานุกรม ก็จะแบ่งคำว่า "นั่ง" เป็นคำต่อไป ตัวอักษรถัดไป คือ ต จากตัวอักษรนี้ เราจะได้คำว่า "ตาก" กับคำว่า "ตาก" แนวคิดนี้ให้เลือกคำที่ยาวที่สุดที่ค้นพบจึงเลือกคำว่า "ตาก" หลังจากนั้น ก็จะค้นหาและเปรียบเทียบต่อไป ซึ่งจะได้ผลลัพธ์ออกมาคือ "ฉัน นั่ง ตาก ลม ที่ หน้า บ้าน"

#### ๒. วิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching)

วิธีการตัดคำแบบนี้เป็นการหาวิธีในการตัดคำที่สามารถจะเป็นไปได้ทั้งหมด เช่น เมื่อมีข้อความว่า "ไปหามเหสี" ก็จะตัดคำได้ ๒ แบบ คือ

๑. ไป หาม เห สี

๒. ไป หาม หเห สี

วิธีการนี้จะให้เลือกข้อความที่แบ่งแล้วมีจำนวนคำน้อยที่สุดคือ แบบที่ ๒ ซึ่งมีจำนวนคำที่ตัดได้ ๓ คำ ในขณะที่แบบที่ ๑ มี ๔ คำ ส่วนในกรณีที่จำนวนคำที่เท่ากัน เราก็จะใช้วิธีการตัดคำแบบยาวที่สุด ((Longest Matching) เข้ามาช่วย เช่น ข้อความว่า "ฉันนั่งตากลมที่หน้าบ้าน" ซึ่งจะสามารถแบ่งคำได้ ๒ แบบ คือ

๑. ฉัน นั่ง ตาก ลม ที่ หน้า บ้าน

๒. ฉัน นั่ง ตา กลม ที่ หน้า บ้าน

ทั้ง ๒ แบบมีจำนวนคำที่เท่ากัน จึงเลือกแบบที่ ๑ โดยเปรียบเทียบจากคำที่ต่างกันที่ติดกันนั้น ( ตา / ตาก) จะเห็นได้ว่า ตาก มีตัวอักษรมากกว่า

### ๓. วิธีการตัดคำแบบคำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (probabilistic Model)

วิธีการนี้นำเอาค่าสถิติการเกิดของคำและลำดับของหน้าที่ของคำ (part of speech) เข้ามาช่วยในการคำนวณหาความน่าจะเป็น เพื่อที่จะใช้เลือกแบบที่มีโอกาสเกิดมากที่สุด วิธีการนี้สามารถจะตัดคำได้ดีกว่า ๒ แบบแรก แต่ข้อจำกัดของวิธีการนี้คือ จะต้องมีความรู้ข้อมูลที่มีการตัดคำที่ถูกต้อง และกำหนดหน้าที่ของคำให้ เพื่อที่จะได้นำไปใช้ในการสร้างสถิติ

### ๔. วิธีการตัดคำแบบคุณลักษณะ (Feature - Based Approach)

วิธีการนี้จะพิจารณาจากบริบท (context words) และการเกิดร่วมกันของคำ หรือหน้าที่ของคำ (collocation) เข้ามาช่วยในการตัดคำตัวอย่างเช่น

"ตากลม" ถ้าพบคำว่า "เป่า" ในบริบทก็จะสามารถตัดคำได้ว่า "ตา" "กลม"

"มากกว่า" ถ้าในบริบทที่ตามมาเป็นตัวเลขก็สามารถตัดคำได้ว่า "มา" "กว่า"

วิธีการนี้จำเป็นต้องมีฐานข้อมูลเป็นจำนวนมาก และจะต้องมีการเรียนรู้การสร้างคำในบริบท หรือการเกิดร่วมกันของคำแต่ละคำ เพื่อให้มีข้อมูลที่จะนำมาใช้ในการตัดคำ

จากที่ได้อธิบายมา ก็คงทำให้ผู้อ่านพอจะเข้าใจถึงแนวคิดในการตัดคำภาษาไทย ซึ่งจากวิธีการข้างต้น เรายังพบว่า แต่ละวิธีการมีข้อจำกัดและไม่สามารถแก้ปัญหาได้ทั้งหมด ในปัจจุบันจึงยังมีการพัฒนาและคิดค้นวิธีการใหม่ ๆ ที่จะช่วยในการแบ่งคำให้มีประสิทธิภาพมากยิ่งขึ้น

## เครื่องแปลภาษาคืออะไร



เครื่องแปลภาษา คือ เครื่องมือที่ใช้สำหรับการแปลงข้อความจำนวนมาก ๆ จากภาษาหนึ่งไปยังอีกภาษาหนึ่ง โดยที่สามารถป้อนข้อมูลภาษาต้นทางเป็นข้อความหรือเสียงก็ได้ และจะได้ภาษาปลายทางเป็นข้อความหรือเสียงก็ได้เช่นกัน ซึ่งจะช่วยให้การแปลสามารถแปลงข้อความได้เป็นจำนวนมากและรวดเร็ว ในยุคที่ต้องการข้อมูลข่าวสารอย่างรวดเร็ว

### ประวัติการทำวิจัยและพัฒนาเครื่องแปลภาษา

การทำวิจัยและพัฒนาเครื่องแปลภาษาเป็นงานแขนงหนึ่งในศาสตร์แห่งการประมวลผล ภาษาธรรมชาติ (Natural Language Processing) เครื่องแปลภาษาเครื่องแรกได้กำเนิดขึ้นประมาณปี ค.ศ. ๑๙๓๐ เครื่องแปลภาษาเป็นซอฟต์แวร์ที่พยายามแปลงข้อความในรูปประโยค โดยพิจารณาเรื่องวากยสัมพันธ์ รวมถึงอรรถศาสตร์ด้วย ไม่ใช่แปลเป็นคำๆ เท่านั้น

### เครื่องแปลภาษาในยุคแรก (ปี ค.ศ. ๑๙๕๐ - ๑๙๖๐)

จากการสำรวจการทำวิจัยและพัฒนาเครื่องแปลภาษา (ALPAC 1966 - 672) พบว่า ระบบที่สำคัญที่สุดในยุคนั้น คือ ระบบ Georgetown Automatic Translation (GAT) ส่วนในยุโรปก็ได้แก่ งานของ Kulagina และ Mel'cuk ในยุคแรกนี้ เครื่องแปลภาษาจะแปลแบบตรง (direct approach) คือ แปลแบบพิจารณาเฉพาะที่ โดยแปลจากภาษาต้นทางเป็นภาษาปลายทางโดยตรงที่ละคู่ภาษาเท่านั้น และโปรแกรมจะพิจารณาไปที่ละคำ ไม่ได้ดูเป็นประโยค เป็นการแปลคำต่อคำแล้วใช้การเปลี่ยนการเรียงลำดับคำให้สอดคล้องกับภาษาเป้าหมาย ส่วนการเขียนกฎไวยากรณ์หรือการใช้ความรู้ทางภาษาศาสตร์เพื่อมาช่วยในการวิเคราะห์นั้น ในยุคแรกนี้ จะเขียนเป็นรหัสโดยตรง จึงใช้งานได้ยาก และมีลักษณะเป็นกระบวนการ (procedural) กล่าวคือ กระบวนการแปลทั้งหมดคือการทำตามคำสั่งต่อ ๆ กัน

แต่เป็นที่น่าเสียดายที่รายงาน ALPAC ดังกล่าว แสดงผลว่า การทำวิจัยและพัฒนาเครื่องแปลภาษาไม่ประสบความสำเร็จ รัฐบาลสหรัฐอเมริกาจึงไม่สนับสนุนงบประมาณในการทำวิจัยเรื่องนี้ ซึ่งผลให้เกิดยุคมืดของการทำวิจัยและพัฒนาแบบแปลภาษาขึ้น อย่างไรก็ตาม การทำวิจัยและพัฒนาในประเทศอื่น ๆ ก็ยังดำเนินต่อไป

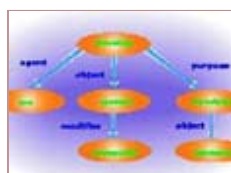
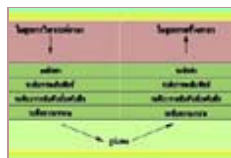
ในยุคนี้ได้มีการนำเครื่องแปลภาษามาใช้งานจริงในปี ค.ศ. ๑๙๗๖ ซึ่งทาง Commission of the European Communities ได้ตัดสินใจสั่งซื้อและพัฒนาระบบ Systran ซึ่งเป็นเครื่องแปลภาษาอังกฤษ - ฝรั่งเศส เพื่อใช้สำหรับบริการการกระจายข่าวของแคนาดา และหลังจากนั้นมาอีก ๕ ปี บริษัทต่าง ๆ ที่พัฒนาเครื่องแปลภาษาก็สามารถนำผลงานออกสู่ตลาดได้เป็นผลสำเร็จ

### เครื่องแปลภาษาในยุคปี ค.ศ. ๑๙๘๐

งานวิจัยและพัฒนาเครื่องแปลภาษาในยุคนี้ยังให้ความสำคัญกับวากยสัมพันธ์ของภาษาและใช้วิธีการ "เปลี่ยน" (transfer approach) ตัวอย่างเช่น ระบบเอเรียน (Ariane) ของมหาวิทยาลัยเกรตโนบลล์ ประเทศฝรั่งเศส ระบบเมทล ของเท็กซัส ระบบซูซี ของ Sarbrucken ระบบ MU ของมหาวิทยาลัยเกียวโต รวมถึงโครงการ Multilingual Eurotra ของกลุ่มประชาคมยุโรป

ต่อมาได้มีการนำวิธีการใช้ภาษากลาง (interlingual approach) เข้ามาใช้ เพื่อที่จะทำการแปลแบบหลายภาษา (multilingual machine translation) ให้ได้ผล เช่น โครงการ DLT และ โรเซตต้า (Rosetta) ประเทศเนเธอร์แลนด์ รวมทั้งโครงการระบบเครื่องแปลหลายภาษาสำหรับภาษาอาเซีย ๕ ภาษา คือ ภาษาญี่ปุ่น ภาษาจีน ภาษามาลาเซีย ภาษาอินโดนีเซีย และภาษาไทย

การพัฒนาเครื่องแปลภาษาในยุคนี้มีแนวคิดเปลี่ยนไปจากเดิมคือ เริ่มมองว่าเครื่องแปลภาษาเป็นเครื่องมือที่จะช่วยแปลภาษาเท่านั้น แต่ไม่สามารถนำมาทดแทนนักแปลภาษามืออาชีพได้ แนวคิดเช่นนี้ก่อให้เกิดการรวมตัวของนักวิจัย เพื่อหาแนวทางใหม่ๆ ในการทำวิจัยและพัฒนา ซึ่งสามารถสรุปแนวทางที่เกิดขึ้นในยุคนี้ได้ดังนี้



[ ขยายดูภาพใหญ่ ]

### ๑. เครื่องแปลภาษาแบบใช้กฎไวยากรณ์

สมมติฐานของการพัฒนาเครื่องแปลภาษาแบบใช้กฎไวยากรณ์ก็คือ การมีกระบวนการวิเคราะห์และการสร้างรูปแบบ (representation) ความหมายของภาษาต้นทาง และสร้างภาษาปลายทางจากรูปแบบนั้น โดนที่รูปแบบจะต้องไม่มีความกำกวมทั้งในระดับคำและโครงสร้าง โดยมีการวิเคราะห์ภาษาต้นทางด้วยความรู้ทางภาษาศาสตร์ ซึ่งประมวลผลออกมาเป็นกฎไวยากรณ์และมีวิธีใช้เพื่อให้สามารถไปถึงจุดหมายนั้นได้หลายวิธี ดังนี้

#### ๑.๑ เครื่องแปลภาษาแบบ "เปลี่ยน"

แนวทางการใช้วิธีการ "เปลี่ยน" นั้น ได้แก่ ระบบเอเรียน และระบบยูโรทรา ฯลฯ ซึ่งเป็นระบบที่ให้ความสำคัญกับวากยสัมพันธ์ของภาษาและมีการทำงานเป็นขั้นตอน ดังนี้

#### ๑.๒ เครื่องแปลภาษาแบบ "ภาษากลาง" (Interlingual Approach)

การทำวิจัยเครื่องแปลภาษาด้วยการใช้ภาษากลางนั้นเริ่มต้นที่มหาวิทยาลัย Carnegie - Melon ซึ่งทำวิจัยและพัฒนา Knowledge - based MT system โดยมีแนวคิดที่ว่า นอกเหนือจากการใช้ความรู้ทางภาษาศาสตร์แล้ว ในการแปลภาษาจะต้องมีความเข้าใจเรื่องหลักภาษาด้วย จึงมีการแปลงรูปแบบทางอรรถศาสตร์ โดยใช้ความรู้จากฐานความรู้ในโดเมนใด ๆ ซึ่งแสดงเป็นโครงข่าย (network) ที่แสดง actual events

### ๒. เครื่องแปลภาษาแบบใช้คลังข้อความ (Corpus - based Approach)

แนวคิดการทำวิจัยและพัฒนาเครื่องแปลภาษาได้เริ่มเข้าสู่ยุคใหม่ ซึ่งมีการเก็บรวบรวมข้อความจำนวนมากที่เรียกว่า คลังข้อความ (Corpus base) เพื่อนำไปศึกษาวิจัย โดยอ้างอิงข้อมูลขนาดใหญ่ที่ใช้จริงในภาษา และมีการดึงข้อมูลจากคลังข้อความไปใช้ในการวิเคราะห์ภาษาหรือการสร้างภาษาโดยตรง จึงก่อให้เกิดวิธีการนำไปใช้ดังต่อไปนี้

#### ๒.๑ วิธีอาศัยค่าสถิติ (Statistic Approach)

ในช่วงปลายคริสต์ทศวรรษ ๑๙๘๐ โครงการ IBM candid Research (Brown et all. 1980, 1990) ได้อาศัยวิธีการทางสถิติในการวิเคราะห์และการสร้างภาษา โดยอาศัยคลังข้อความขนาดใหญ่ที่มีชื่อว่า Canadian Hansard ซึ่งได้เก็บบันทึกการอภิปรายในสภาไว้เป็น ๒ ภาษาคือ ภาษาอังกฤษ และภาษาฝรั่งเศส วิธีการที่นำมาใช้คือ การหาขอบเขตของประโยค ๒ ภาษาที่ตรงกันจากคลังข้อความที่เป็นคู่ภาษา จากนั้นนำมาคำนวณหาค่าความเป็นไปได้ของคำในภาษาต้นทางว่าตรงกับภาษาปลายทางที่ค่า โดยคำนวณค่าความเป็นไปได้จากการจับคู่คำที่อยู่ติดกัน (bigram) ของภาษาอังกฤษ กับคู่คำที่อยู่ติดกัน (bigram) ของภาษาฝรั่งเศส ซึ่งผลสำเร็จที่สามารถแปลได้มีความถูกต้องประมาณ ๔๘ %

ต่อมาได้มีการปรับปรุงการทำวิจัยและพัฒนาในเรื่องนี้ โดยปรับการคำนวณค่าความเป็นไปได้ให้ถูกต้องมากยิ่งขึ้น นอกจากนั้น นักวิจัยบางคนก็นำความรู้ทางภาษาศาสตร์เรื่องคำและไวยากรณ์มาใช้ร่วมกัน เพื่อให้งานนั้นถูกต้องยิ่งขึ้น

#### ๒.๒ วิธีการแปลแบบใช้ประโยคตัวอย่าง (Example base Approach)

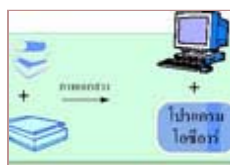
ในกลางคริสต์ทศวรรษ ๑๙๘๐ นากาโอะ (Nagao) ได้เสนอวิธีการแปลโดยใช้ประโยคตัวอย่างจากคลังข้อความคู่ภาษาขนาดใหญ่ ซึ่งโปรแกรมจะคำนวณว่า ประโยคที่มีปรากฏว่าแปลไว้ในคลังตัวอย่าง และดึงขึ้นมาใช้ โครงการ ATR ซึ่งมีเครื่องแปลภาษาแบบใช้วิธี "เปลี่ยน" ก็ได้นำวิธีนี้มาใช้ โดยเลือกเก็บประโยคตัวอย่าง ซึ่งได้มาจากการคำนวณค่าระยะห่างของคำที่มีความหมายใกล้เคียงกันในอภินิหารศัพท์ (Thesaurus) และจะใช้การ "เปลี่ยน" ในระดับต่าง ๆ เช่น ระดับคำ ระดับรูปแบบ และระดับวากยสัมพันธ์ เช่น ในภาษาญี่ปุ่น โครงสร้าง N1 no N2 จะเปลี่ยนไปให้ตรงกับ N2 of N1 แต่ก็ไม่เสมอไป เช่น จะใช้สำนวนว่า fee for the conference มากกว่า fee of the conference และจะใช้ conference in Tokyo มากกว่า conference of Tokyo ดังนั้น จึงมีการเก็บประโยคตัวอย่างเหล่านี้ไว้

### เครื่องแปลภาษาไทยเป็นภาษาอื่น ๆ

การทำวิจัยและพัฒนาเครื่องแปลภาษาในประเทศไทยเริ่มต้นในปี พ.ศ. ๒๕๒๔ โดยจุฬาลงกรณ์มหาวิทยาลัยร่วมกับมหาวิทยาลัยเกรอนอบล์ (Grenoble) แห่งประเทศฝรั่งเศสได้ร่วมโครงการทำวิจัยและแปลภาษาเป็นไทยด้วยคอมพิวเตอร์ ซึ่งเป็นโครงการของทบวงมหาวิทยาลัย (พ.ศ. ๒๕๒๔ - ๒๕๓๐)

ต่อมาเกิดโครงการความร่วมมือในการพัฒนาระบบแปลหลากหลายภาษาสำหรับภาษาในเอเชีย ซึ่งได้แก่ ภาษาจีน ญี่ปุ่น มาเลเซีย อินโดนีเซีย และไทย โครงการนี้เป็นโครงการของกระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม ซึ่งดำเนินการโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (พ.ศ. ๒๕๓๐ - ๒๕๓๗)

### โปรแกรมรู้จำอักขระไทยด้วยแสงคืออะไร



โอซีอาร์เป็นคำย่อของภาษาอังกฤษ คือ "Optical Character Recognition : OCR" แปลเป็นภาษาไทยได้ว่า "การรู้จำอักขระด้วยแสง" ซึ่งเป็นงานประยุกต์งานหนึ่งของสาขาวิทยาการคอมพิวเตอร์ ที่ได้รับความสนใจและพัฒนามานานกว่า ๗๐ ปีแล้ว โอซีอาร์เป็นการรู้จำรูปแบบตัวอักษร ซึ่งเป็นงานวิจัยในสาขารูจำรูปแบบ (Pattern Recognition) เป็นเทคโนโลยีที่ส่งผลให้ระบบคอมพิวเตอร์สามารถระบุรูปแบบได้อย่างถูกต้อง เช่น สามารถบอกได้ว่า ภาพนั้นคือภาพอะไร ตัวอักษรนั้นคือตัวอักษรอะไร หรือเสียงนั้นคือเสียงของคำสั่งอะไร เป็นต้น

นักวิจัยเป็นจำนวนมากมีความสนใจงานโอซีอาร์ เพราะเป็นงานที่เกี่ยวข้องกับเอกสาร ซึ่งมีปริมาณมากมายมหาศาล การ



[ ขยายดูภาพใหญ่ ]

เก็บข้อมูลเหล่านี้ให้เป็นแฟ้มข้อความ (Text File) ไว้ในระบบคอมพิวเตอร์นั้น ต้องใช้บุคลากรในการจัดพิมพ์เอกสารนั้น ๆ โดยใช้โปรแกรมพิมพ์และประมวลผลเอกสาร (Word Processing Program) ถึงแม้ว่าโปรแกรมประเภทนี้จะมีความสามารถและเป็นเครื่องมือที่ดี แต่ก็ต้องใช้บุคลากรในการพิมพ์งานซึ่งใช้เวลามากพอสมควร และยังเป็นงานที่จำเจสำหรับบุคลากรอีกด้วย ถ้าไอซีอาร์ประสบความสำเร็จ งานพิมพ์เอกสารต่าง ๆ เพื่อเก็บเป็นแฟ้มข้อความก็จะกลายเป็นหน้าที่ของระบบคอมพิวเตอร์แทน โดยทั่วไปแล้ว เวลาในการประมวลผลของไอซีอาร์จะเร็วกว่าการพิมพ์ด้วยมนุษย์โดยเฉลี่ยประมาณ ๕ เท่า และในบางระบบงานที่ได้จากไอซีอาร์จะมีความถูกต้องมากกว่า งานที่ได้จากการพิมพ์ของมนุษย์อีกด้วย จึงทำให้งานวิจัยด้านไอซีอาร์ได้รับความนิยมเป็นอย่างมาก

ประเทศไทยเริ่มจะมีงานวิจัยเรื่องไอซีอาร์ สำหรับใช้งานกับเอกสารภาษาไทยในระยะเวลาประมาณ ๑๐ ปีที่ผ่านมาในทศวรรษนี้ และเพื่อความสะดวกในการกล่าวถึงงานวิจัยไอซีอาร์สำหรับภาษาไทย จึงเป็นที่นิยมโดยทั่วไปที่จะเรียกงานวิจัยในสาขานี้ว่า "ไทยไอซีอาร์" ซึ่งมีการทำวิจัยทั้งในหน่วยงานภาครัฐและภาคเอกชน ตัวอย่างการทำวิจัยในสาขานี้ของทางภาครัฐ เช่น สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง โดย ดร. ชม กัมปาล สสถาบันบัณฑิตพัฒนบริหารศาสตร์ โดย ดร. พิพัฒน์ หิรัญนิชกร จุฬาลงกรณ์มหาวิทยาลัย โดย ดร. ชิตชนก เหลือสินทรัพย์ และ ดร. สมชาย จิตตะพันธ์กุล ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ โดย ดร. ทวีศักดิ์ กอนันตกุล และ ดร. จุฬารัตน์ ต้นประเสริฐ เป็นต้น ส่วนทางภาคเอกชน ได้แก่ บริษัทเทไรม เทคโนโลยี จำกัด และบริษัทเอ็นเอสที อิเลคทรอนิกส์ พับลิชชิ่ง จำกัด

ซอฟต์แวร์ไทยไอซีอาร์เริ่มออกสู่ตลาดในปี พ.ศ. ๒๕๓๙ จนถึงปี พ.ศ. ๒๕๔๑ ในท้องตลาดมีซอฟต์แวร์ไทยไอซีอาร์อยู่ ๓ ซอฟต์แวร์ด้วยกัน ได้แก่ ซอฟต์แวร์ไทยไอซีอาร์ ของบริษัทเทไรม เทคโนโลยี จำกัด ซอฟต์แวร์อ่านไทย โดยความร่วมมือระหว่าง ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ กับบริษัท ไทยซอฟท์ จำกัด และซอฟต์แวร์ทีเรค (T - rec) โดยบริษัทเอ็นเอสที อิเลคทรอนิกส์ พับลิชชิ่ง จำกัด ความสามารถในการรู้จำของทั้ง ๓ ซอฟต์แวร์ มีความแตกต่างกันไปตามลักษณะของฟอนต์ตัวอักษรไทย และ / หรือลักษณะของเอกสาร ซึ่งในปัจจุบันนี้ถือว่า ความสามารถในการรู้จำของไอซีอาร์อยู่ในระดับพอใช้งานได้ แต่ยังคงต้องการพัฒนาปรับปรุงเพิ่มประสิทธิภาพต่อไปอีกในอนาคต

## ดูเพิ่มเติมเรื่อง การพัฒนาอักษรไทยในเครื่องคอมพิวเตอร์ เล่ม ๑๑ โครงข่ายประสาทเทียม และการพัฒนาซอฟต์แวร์ เล่ม ๒๕



[ จอหลัก | รูป | เล่าให้ฟัง | วิดีโอ | ลับสมอง | ค้นข้อมูล | บทความ | ย้อนกลับ ]

Copyright © 1997 Thai Junior Encyclopedia Project Webmaster. All rights reserved.

Reproduction of the information contain in this web site without permission is prohibited.