หอสมุดกลาง สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ได้นำความรู้ทางด้านภาษาศาสตร์ รวมไปถึงงานวิจัยที่เกี่ยวข้องกับวิธีการเข้ารหัสเสียง และโดยเฉพาะอย่างยิ่งวิธีการตัดพยางค์ เพื่อนำมาปรับใช้ในงานวิจัยนี้ โดยมีรายละเอียดดังต่อไปนี้

2.1 หลักเกณฑ์การออกเสียง

ระบบเสียงในทุกภาษาย่อมประกอบด้วยหน่วยเสียง (Phoneme) [8] จำนวนหนึ่งหน่วยเสียงนี้หมายถึง เสียงสำคัญในภาษาที่มีหน้าที่แยกความหมายของคำในภาษา หน่วยเสียงยังสามารถแบ่งประเภทได้เป็น หน่วยเสียงพยัญชนะ หน่วยเสียงสระ และสำหรับภาษาไทยนั้น จะมีหน่วยเสียงวรรณยุกต์ด้วย และเพื่อให้การพิจารณาศึกษาเรื่องเสียงที่ปรากฏในภาษาพูดได้สะดวก นักภาษาศาสตร์จึงได้กำหนดอักษรแทนเสียงขึ้นชุดหนึ่งเรียกว่า สัทอักษร (Phonetic Alphabet)

สัทอักษร คือ อักษรที่ใช้เฉพาะในทางสัทศาสตร์ กำหนดขึ้นเพื่อแสดงลักษณะในการออกเสียงของภาษาอย่างเป็นสากล จึงไม่มุ่งแสดงลักษณะเฉพาะของภาษาใดภาษาหนึ่ง โดยมีสมาคมสัทศาสตร์ระหว่างชาติ (the International Phonetic Association - IPA) เป็นผู้กำหนดสัทอักษรขึ้นใช้แทนเสียงพูดของมนุษย์ทั่วโลก สัทอักษรนี้ส่วนใหญ่เป็นอักษรโรมัน และมีเครื่องหมายประกอบบ้าง เพื่อให้มีอักษรเพียงพอที่จะบันทึกเสียงของภาษาต่างๆ ได้ทั่วโลก และแม้ว่าหน่วยเสียงแต่ละหน่วยในทุก ๆ ภาษา แต่ละคนอาจออกเสียงแตกต่างกันไปเล็กน้อย แต่นักภาษาศาสตร์ยังคงถือว่าเป็นหน่วยเสียงเดียวกัน เมื่อนักภาษาศาสตร์จะทำการวิเคราะห์ภาษาใดก็สามารถเลือกสัทอักษรที่มีลักษณะของเสียงตรงกันไปใช้ สัทอักษรที่ใช้สำหรับภาษาไทยก็นำมาจากสัทอักษรของ IPA นี้เช่นกัน

2.2 ระบบเสียงในภาษาไทยและระบบเสียงในภาษาอังกฤษ

ระบบเสียงในภาษาไทย และระบบเสียงในภาษาอังกฤษ [9] นั้นจะมีความแตกต่างกันบ้าง เนื่องจากจำนวนของหน่วยเสียงพื้นฐานที่ไม่เท่ากัน รวมไปถึงระบบเสียงในภาษาอังกฤษนั้นไม่มีหน่วยเสียงวรรณยุกต์เหมือนในระบบเสียงภาษาไทย แต่เนื่องจากในงานวิจัยนี้ ต้องการจะเข้ารหัสให้เป็นหน่วยเสียงที่เหมือนกัน หรือใกล้เคียงกันมากที่สุด จึงจำเป็นต้องมีกระบวนการในการประมวลผลตัวอักษรเบื้องต้น สำหรับกรณีของคำภาษาไทยด้วย

2.2.1 ระบบเสียงในภาษาไทย

หน่วยเสียงภาษาไทย มี 47 หน่วยเสียง ดังนี้

1.หน่วยเสียงพยัญชนะ มี 21 หน่วยเสียง ซึ่งหน่วยเสียงพยัญชนะทั้ง 21 หน่วยเสียงมีการปรากฏในพยางค์ดังนี้

- เป็นพยัญชนะต้นได้ 21 หน่วยเสียง
- เป็นพยัญชนะต้นควบได้ 9 หน่วยเสียง เรียงต่อกันได้ 12 แบบ
- เป็นพยัญชนะท้าย หรือตัวสะกดได้ 9 หน่วยเสียง

2.หน่วยเสียงสระ มี 21 หน่วยเสียง ซึ่งจำแนกเป็น 2 ชนิด ดังนี้

- สระเดี่ยว มี 18 หน่วยเสียง
- สระประสม มี 3 หน่วยเสียง

3.หน่วยเสียงวรรณยุกต์ มี 5 หน่วยเสียง ดังนี้

- หน่วยเสียงวรรณยุกต์สามัญ
- หน่วยเสียงวรรณยุกต์เอก
- หน่วยเสียงวรรณยุกต์โท
- หน่วยเสียงวรรณยุกต์ตรี
- หน่วยเสียงวรรณยุกต์จัตวา

2.2.2 ระบบเสียงในภาษาอังกฤษ

หน่วยเสียงภาษาอังกฤษ มี 44 หน่วยเสียง ดังนี้

1.หน่วยเสียงพยัญชนะ มี 24 หน่วยเสียง

2.หน่วยเสียงสระ มี 20 หน่วยเสียง

2.3 ความหมายของพยางค์

พยางค์ในภาษาไทยนั้นเกิดจากการเปล่งเสียงพยัญชนะ เสียงสระ และเสียงวรรณยุกต์ตามกันออกมาอย่างกระชั้นชิด จนฟังดูเหมือนกับเปล่งเสียงออกมาในครั้งเดียวกัน ซึ่งเรียกว่า การประสมเสียงในภาษา เสียงที่เกิดจากการประสมเสียง จึงเรียกว่า พยางค์ แต่ในภาษาอังกฤษนั้น พยางค์ (Syllable) นั้นจะแตกต่างกับพยางค์ในภาษาไทย ตรงที่ในภาษาอังกฤษนั้นไม่มีหน่วยเสียงของวรรณยุกต์

พระยาอุปกิตศิลปสาร [10] กล่าวว่า "ถ้อยคำที่เราใช้พูดกันนั้น บางทีก็เปล่งเสียงออกครั้งเดียว บางทีก็หลายครั้ง เสียงที่เปล่งออกมาครั้งหนึ่ง ๆ นั้น ท่านเรียกว่า "พยางค์" คือ ส่วนของคำพูด"

กาญจนา นาคสกุล [11] กล่าวว่า "พยางค์ จึงหมายถึง จำนวนเสียงที่ดังเด่น ซึ่งปรากฏในกลุ่มเสียงที่เรียงเป็นคำพูด เสียงอื่น ๆ ที่อยู่ข้างเคียงก็จะประกอบเข้าเป็นส่วนของพยางค์ โดยปกติเสียงสระจะเป็นเสียงที่มีลักษณะประจำตัว เป็นเสียงก้องที่ดังกว่าเสียงอื่น ฉะนั้นเสียงสระจึงมักจะเป็นเสียงที่ทำให้เกิดพยางค์"

จากคำอธิบายความหมายของพยางค์ดังกล่าว สรุปได้ว่าพยางค์คือ เสียงที่เปล่งออกมาครั้งหนึ่ง ๆ ซึ่งมีเสียงสระเป็นเสียงที่ดังเด่นอยู่จำนวน 1 เสียง และเสียงที่อยู่ข้างเคียงอีกอย่างน้อย 2 เสียง ได้แก่ เสียงของพยัญชนะ และเสียงของวรรณยุกต์สำหรับระบบเสียงภาษาไทย (ในภาษาอังกฤษจะไม่มีหน่วยเสียงวรรณยุกต์) ซึ่งเสียงสระนี้จะเป็นเสียงที่ทำให้เกิดพยางค์ขึ้นในภาษา พยางค์อาจจะเป็นคำก็ได้ถ้าพยางค์นั้นมีความหมาย เช่น

นา	เป็น 1 พยางค์ 1 คำ
สาหร่าย	เป็น 2 พยางค์ 1 คำ
อภิเษก	เป็น 3 พยางค์ 1 คำ

ซึ่งในงานวิจัยนี้ได้ใช้นิยามความหมายของพยางค์ เพื่อจำแนกและตัดพยางค์ เนื่องจากถ้าเราทราบจำนวนของสระที่ปรากฏอยู่ในคำใดๆ แล้ว เราจะสามารถทราบได้ถึงจำนวนของพยางค์ ซึ่งจำนวนของพยางค์จะเป็นตัวบ่งบอกถึงการจะแบ่งคำว่าจะสามารถแบ่งคำได้ออกเป็นกี่พยางค์

2.4 การถอดอักษร

การถอดอักษร (Transliteration) หมายถึง การนำคำในภาษาหนึ่งมาเขียนด้วยตัวอักษรอีกภาษาหนึ่งแบบอักษรต่ออักษร โดยพยายามใช้หน่วยเสียงของอักษรทั้งสองภาษาใกล้เคียงกันมากที่สุด [12] ตัวอย่างเช่น คำว่า "OXFORD" ในภาษาอังกฤษถอดอักษรเป็น "ออกซ์ฟอร์ด" ในภาษาไทย เป็นต้น การถอดอักษรแบ่งเป็น 3 ขั้นตอนหลัก ๆ ดังนี้

1. ถอดหน่วยอักษรในภาษาต้นแบบ (Source Language) เป็นหน่วยเสียงในภาษาต้นแบบ เช่น ถอดหน่วยอักษร "B" ในภาษาอังกฤษเป็นหน่วยเสียง /b/ ในภาษาอังกฤษ เป็นต้น
2. แทนหน่วยเสียงในภาษาต้นแบบ ด้วยหน่วยเสียงในภาษาเป้าหมาย (Target Language) โดยพยายามใช้หน่วยเสียงที่ใกล้เคียงกันมากที่สุด เช่น แทนหน่วยเสียง /b/ ในภาษาอังกฤษเป็นหน่วยเสียง /b/ ในภาษาไทย เป็นต้น
3. ถอดหน่วยเสียงในภาษาเป้าหมาย เป็นหน่วยอักษรในภาษาเป้าหมาย เช่น ถอดหน่วยเสียง /b/ ในภาษาไทยเป็นหน่วยอักษร "บ" ในภาษาไทย เป็นต้น

ปัญหาต่าง ๆ ในการถอดอักษรได้แก่

1. ความสัมพันธ์ของหน่วยอักษรและหน่วยเสียง มีความสัมพันธ์แบบหนึ่งตัวอักษรแทนหลายหน่วยเสียง เช่น ในภาษาอังกฤษ "C" แทนด้วย /k/ หรือ /s/ เป็นต้น และมี

ความสัมพันธ์แบบหลายหน่วยอักษรแทนหนึ่งหน่วยเสียง เช่น ในภาษาอังกฤษ "N, TN, GN, PN" แทนด้วย /n/ ในภาษาไทย "ร, ฤ, หร" แทนด้วย /r/ และ "ฉ, ช, ฌ" แทนด้วย /ch/ เป็นต้น

2. การแบ่งพยางค์ในภาษาต้นแบบ เมื่อมีพยัญชนะตัวเดียวอยู่ระหว่างสระ เช่น คำว่า money ในภาษาอังกฤษ จะแบ่งพยางค์อย่างไร จะถอดพยัญชนะซ้ำสองตัวเพื่อให้อ่านได้สะดวกเป็น มัน-นีย์ หรือจะถอดอักษรเพียงตัวเดียวตามที่ปรากฏในภาษาอังกฤษเป็น มะ-นีย์ หรือ มัน-อีย์
3. ปัญหาอันเนื่องมาจากช่วงเวลาของการยืมคำทับศัพท์ คำทับศัพท์บางคำยืมมาเป็นเวลานาน ซึ่งในอดีตมีหลักเกณฑ์การทับศัพท์ไม่ตรงกับหลักเกณฑ์ในปัจจุบัน เช่น "C" ที่แทน /k/ ในอดีตนิยมถอดเป็นอักษร "ก" เช่น กุ๊ก (Cook) กัปตัน (Captain) กะรัต(Carat) แก๊ป(Cap) เป็นต้น แต่ปัจจุบัน "C" ที่แทน /k/ มักจะถอดเป็น "ค" ในตำแหน่งพยัญชนะต้น เช่น คอนโดมิเนียม (Condominium) แคปซูล (Capsule) แคร็อต (Carrot) เป็นต้น

2.5 การถ่ายเสียงด้วยตัวอักษรโรมัน

การถ่ายเสียงด้วยตัวอักษรโรมัน (Romanization) คือการถ่ายเสียงตัวอักษรของภาษาอื่นที่ไม่ใช่อักษรโรมัน เช่น ไทย จีน ญี่ปุ่น ฯลฯ ให้เป็นตัวอักษรโรมัน เพื่อให้ผู้ที่ไม่รู้จักภาษานั้น ๆ สามารถอ่านออกเสียงได้ ทางราชบัณฑิตยสถานจึงได้กำหนดระบบการใช้ตัวอักษรโรมันเพื่อการถ่ายเสียงสำหรับตัวอักษรไทยออกเป็น 2 ระบบ คือระบบทั่วไปและระบบพิสดาร โดยระบบทั่วไปจะใช้สำหรับกรณีที่การออกเสียงสำคัญกว่าการเขียนตัวสะกด ซึ่งจะอาศัยหลักการออกเสียงเป็นสำคัญ ต้องสอดคล้องกับไวยากรณ์ของไทย และสามารถขยายเป็นระบบเฉพาะได้ เช่น คำว่า "กษัตริย์" ถ่ายเสียงเป็น Kasat ส่วนระบบพิสดารจะใช้ในกรณีที่จะแสดงตัวอักษรให้ละเอียดแม่นยำ เพื่อให้คงความหมายของคำนั้นไว้ เช่น คำว่า "กษัตริย์" ถ่ายเสียงเป็น Kasatriy

2.6 การตัดพยางค์

การตัดพยางค์ [13] นั้นจะกระทำโดยการหาขอบเขตความยาวของแต่ละพยางค์ที่ปรากฏในคำนั้นๆ ซึ่งจะแบ่งการพิจารณาออกเป็น 2 ชนิด คือ

- **การหาขอบเขตหน้า** (Front boundary) จะใช้สำหรับการกำหนดการเริ่มต้นของพยางค์
- **การหาขอบเขตหลัง** (Tail boundary) จะใช้สำหรับการกำหนดตำแหน่งท้ายสุดของพยางค์

เนื่องจากการเขียนประโยคภาษาไทย เป็นการเขียนคำต่อเนื่องโดยไม่มีการเว้นวรรคช่องว่างระหว่างคำ ดังนั้นจำเป็นต้องมีการตัดพยางค์ก่อน ซึ่งวิธีการตัดพยางค์นั้นจะใช้การอ้างอิงกับกฎที่สร้างขึ้นจากโครงสร้างของแต่ละพยางค์ที่เป็นไปได้จากหลักภาษาไทยที่ถูกต้อง ทั้งนี้เนื่องจากแต่ละพยางค์ในภาษาไทยนั้น มีโครงสร้างตายตัว กฎที่ใช้ภายในบทความนี้ แบ่งออกเป็น 2 หมวด

1.**หมวดกฎที่สร้างจากสระ** ตัวอย่างกฎจากสระที่ใช้หาขอบเขตหน้า เช่น สระ อา อะ อิ อี อึ อื อุ อู อำ อั ต้องมีอักษรข้างหน้าอย่างน้อย 1 ตัวเสมอ ตัวอย่างกฎจากสระที่ใช้หาขอบเขตหลัง เช่น สระ อี ถ้ามีวรรณยุกต์ ไม่มีตัวสะกดเสมอ สำหรับสระที่มี หรือ ไม่มีตัวสะกดก็ได้ให้ตัดแบบไม่มีตัวสะกดก่อน ผลลัพธ์ที่ได้จากการผ่านกฎจากสระ แบ่งออกเป็น 2 แบบ

- กลุ่มอักษรที่มีสระหรือวรรณยุกต์ประกอบร่วมอยู่ เช่น พ่อ, ข้าว เป็นต้น
- กลุ่มอักษรที่มีแต่พยัญชนะประกอบอย่างเดียว เช่น ขวด, ครอบครอง เป็นต้น

2.**หมวดกฎที่สร้างจากพยัญชนะ** ใช้กับผลลัพธ์ที่ผ่านกฎจากสระมาแล้ว โดยพิจารณาเฉพาะกลุ่มอักษรที่มีแต่พยัญชนะประกอบอย่างเดียว เพื่อใช้หาขอบเขตหน้า และขอบเขตหลังที่แท้จริงของพยางค์ เนื่องจากขอบเขตของพยางค์อาจขยายออกไปอีกได้ เช่น

- พยัญชนะต้นของพยางค์มีได้มากกว่า 1 ตัว (คำควบกล้ำหรืออักษรนำ) เช่น "กลอง"
- ตัวสะกดของพยางค์สามารถมีได้มากกว่า 1 ตัว เช่น "จิตร"

2.7 นิวรอลเน็ตเวิร์ก (Neural Networks)

นิวรอลเน็ตเวิร์ก (Neural Networks) เป็นวิทยาการแขนงหนึ่งทางปัญญาประดิษฐ์ (Artificial Intelligence) ที่พยายามลอกเลียนแบบการทำงานของเซลล์ประสาทในสมองมนุษย์ เรียกว่า เซลล์ประสาทเทียม (Artificial Neural) หรือยูนิต (Unit) ซึ่งเซลล์ประสาทเทียมที่จำลองขึ้นนั้นจะมีความเร็วในการประมวลผลที่สูงกว่าเซลล์ประสาทจริงในสมองมนุษย์ แต่เซลล์ประสาทในสมองมนุษย์สามารถทำงานได้หลากหลายและมีความสลับซับซ้อนได้ดีและเร็วกว่าเซลล์ประสาทเทียมในคอมพิวเตอร์ เนื่องจากโครงสร้างของเซลล์ประสาทมนุษย์นั้นมีโครงสร้างที่ขนานกันอย่างมาก (Massively Parallel Structure) ซึ่งการจำลองการทำงานของเซลล์ประสาทเทียมนั้น ได้อาศัยโครงสร้างที่ประกอบไปด้วยหน่วยประมวลผล (Processing Element) ที่เรียกว่า นิวรอล (Neural) จำนวนมากเชื่อมต่อกัน (Connection) ทำให้การทำงานของนิวรอลเน็ตเวิร์กนั้นเป็นไปในรูปแบบของการประมวลผลพร้อม ๆ กันจำนวนมาก (Massively Parallel Processing) โดยการเชื่อมต่อถึงกันของหน่วยประมวลผลนั้นจะมีค่าน้ำหนัก (Weight) ของแต่ละการเชื่อมต่อและเมื่อมีการให้ตัวอย่างที่ใช้ในการเรียนรู้ นิวรอลเน็ตเวิร์กก็จะทำการปรับค่าน้ำหนักให้เหมาะสม จนได้

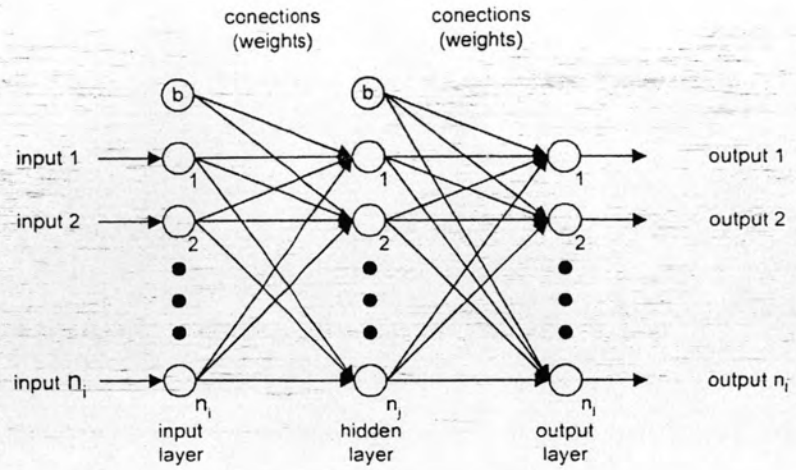
ผลลัพธ์ที่ถูกต้องหรือมีข้อผิดพลาดน้อยที่สุด และสามารถนำค่าน้ำหนักที่ถูกปรับนี้ไปใช้ในการเรียนรู้ครั้งต่อไปได้

นิวรอลเน็ตเวิร์กเป็นวิธีการเรียนรู้ของเครื่อง (Machine Learning) วิธีหนึ่ง ซึ่งมักใช้ในปัญหาการจำแนกประเภท (Classification) หรือปัญหาการจำแนกรูปแบบ (Pattern Classification) นิวรอลเน็ตเวิร์กประกอบด้วยนิวรอน (Neuron) จำนวนมากเชื่อมต่อกัน นิวรอนที่ใช้นั้นมีได้หลายรูปแบบ แบบที่นิยมใช้ได้แก่นิวรอนแบบซิกมอยด์ (Sigmoid unit) ซึ่งคำนวณผลลัพธ์จากอินพุตได้ตามสมการ (2.1) โดยกำหนดให้ W คือเวกเตอร์ของค่าน้ำหนัก (Weight) X คือเวกเตอร์ของค่าอินพุต และ σ คือฟังก์ชันกระตุ้น (Activation function) ประเภทฟังก์ชันซิกมอยด์ (Sigmoid function)

$$\mathbf{O} = \sigma(\mathbf{W} \cdot \mathbf{X}) \tag{2.1}$$

โดย $$\sigma(y) = \frac{1}{1+e^{-y}} \tag{2.2}$$

รูปที่ 2.1 แสดงนิวรอลเน็ตเวิร์กที่มีโครงสร้าง 3 ชั้นคือชั้นอินพุต (Input layer) ชั้นซ่อน (Hidden layer) และชั้นเอาท์พุต (Output layer) แต่ละการเชื่อมต่อของนิวรอนมีค่าน้ำหนักกำหนดไว้ ค่าน้ำหนักเหล่านี้จะถูกปรับค่าให้เหมาะสมในระหว่างการเรียนรู้ จนกระทั่งได้ผลการเรียนรู้ถูกต้องที่สุดหรือมีความผิดพลาดน้อยที่สุด วิธีการเรียนรู้ของนิวรอลเน็ตเวิร์กแบบหนึ่งคือการเรียนรู้แบบแพร่กระจายย้อนกลับ (Backpropagation Learning Algorithm) ซึ่งทำโดยการผ่านข้อมูลฝึกสอนเข้าไปยังนิวรอลเน็ตเวิร์ก แล้วนำผลลัพธ์ไปเปรียบเทียบกับค่าเป้าหมาย (Target) ผลต่างที่ได้คือค่าผิดพลาด ค่าผิดพลาดนี้จะนำไปใช้ในการคำนวณปริมาณการปรับค่าน้ำหนักเพื่อให้ค่าน้ำหนักของการเชื่อมต่อมีค่าที่เหมาะสม กระบวนการนี้จะทำซ้ำจนกระทั่งค่าผิดพลาดมีค่าน้อยในระดับที่ยอมรับได้หรือจนกระทั่งครบจำนวนรอบที่กำหนด การเรียนรู้แบบแพร่กระจายย้อนกลับสามารถอธิบายได้ดังรูปที่ 2.2



รูปที่ 2.1 นิวรอลเน็ตเวิร์กที่มี 3 ชั้น

X : vector of network input values

T : vector of target network output values

η : learning rate

x_{ji} : input from unit i to unit j

w_{ji} : weight from unit i to unit j

Create feed-forward network with n_{in} inputs, n_{hidden} hidden units, and n_{out} output units

Initialize all network weights to small random numbers

Until the termination condition is met, Do

For each training examples (X, T) Do

Input training instance X to the network and compute output O

For each network output unit k, calculate its error term δ_k

$$\delta_k \leftarrow O_k (1-O_k)(T_k-O_k)$$

For each hidden unit h, calculate its error term δ_h

$$\delta_h \leftarrow O_h (1-O_h) \Sigma_{k \in outputs} w_{kh} \delta_k$$

Update each network weight w_{ji}

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji} \quad \Delta w_{ji} = \eta \delta_k x_{ji}$$

รูปที่ 2.2 ขั้นตอนวิธีการเรียนรู้แบบแพร่กระจายย้อนกลับ

2.8 การวัดผลการค้นคืน

มาตรวัดที่ใช้วัดประสิทธิภาพของการค้นคืนได้แก่ ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) [14] และตัววัด F1 (F1 Measurement) [15] ซึ่งมีวิธีคำนวณจากการนับจำนวนข้อมูลที่เกี่ยวข้อง (Relevant Data) และข้อมูลที่ระบบค้นคืนกลับมา (Retrieved Data) ได้ดังนี้

$$\text{ค่าความเที่ยง} = \frac{\text{จำนวนคำที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำที่คืนกลับมาทั้งหมด}}$$

$$\text{ค่าเรียกคืน} = \frac{\text{จำนวนคำที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำที่เกี่ยวข้องทั้งหมด}}$$

$$F1 = \frac{2 \times \text{ค่าความเที่ยง} \times \text{ค่าเรียกคืน}}{\text{ค่าความเที่ยง} + \text{ค่าเรียกคืน}}$$

คำที่เกี่ยวข้อง หมายถึง คำที่มีเสียงอ่านตรงกันจริง ส่วนคำที่คืนกลับมาหมายถึง คำที่ผ่านเกณฑ์การเปรียบเทียบรหัสคำ โดยที่ในความเป็นจริงแล้วอาจมีเสียงอ่านตรงกันหรือไม่ก็ได้

2.9 ขั้นตอนวิธีระยะแก้ไขสั้นที่สุด (Minimum Edit Distance)

ระยะแก้ไขสั้นที่สุด [6] เป็นเทคนิคหนึ่งที่ถูกนำมาใช้ในการวัดความคล้ายคลึงกันระหว่าง 2 สายอักขระใดๆ ซึ่งจะทำการคำนวณหาจำนวนคำสั่งที่น้อยที่สุดที่จะใช้ในการเพิ่ม การลบ และการแทนที่แต่ละตัวอักขระ เพื่อทำให้สายอักขระทั้งสองสายเหมือนกัน ตัวอย่างเช่น ระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE เท่ากับ 3 ซึ่งมีวิธีการคำนวณดังนี้

ตัวอย่างที่ 2.1 การคำนวณระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE

1. การลบตัวอักษร S EXSAMBL → EXAMBL
2. การแทนที่ตัวอักษร B ด้วย P EXAMBL → EXAMPL
3. การเพิ่มตัวอักษร E EXAMPL → EXAMPLE

ดังนั้นระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE มีค่าเท่ากับ 3

จากวิธีการคำนวณข้างต้นสามารถเขียนในอยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด Edit (P_j, W_k) ได้ดังนี้

$$\text{Edit}(P_0, W_0) = 0$$

$$\text{Edit}(P_j, W_0) = j$$

$$\text{Edit}(P_0, W_k) = k$$

$$\text{Edit}(P_j, W_k) = \min[\ \text{Edit}(P_{j-1}, W_k) + 1,$$

$$\text{Edit}(P_j, W_{k-1}) + 1,$$

$$Edit(P_{j-1}, W_{k-1}) + r(p_j, w_k)\ ]$$

โดยที่ $P_j = p_1 p_2 p_3 \ldots p_j$ เป็นสายอักขระต้นแบบ มีความยาว j ตัวอักษร

$W_k = w_1 w_2 w_3 \ldots w_k$ เป็นสายอักขระเป้าหมาย มีความยาว k ตัวอักษร

$r(p_j, w_k) = 0$ ถ้า p_j เท่ากับ w_k

1 ถ้า p_j ไม่เท่ากับ w_k

2.10 งานวิจัยของ วรรณี อุดมพาณิชย์

งานวิจัย [16] นำแนวคิดของการเข้ารหัสซาวด์เด็กซ์ภาษาอังกฤษ (ดูตารางที่ 2.1) มาพัฒนา และเสนอกฎเกณฑ์การสร้างรหัสคำได้แก่ (1) ไม่ใช้สระ วรรณยุกต์ และไม้ไต่คู้ มาสร้างรหัส ยกเว้นสระที่ให้เสียงตัวสะกดเป็นพยัญชนะ ได้แก่ ไ- ใ- -ำ (2) เปลี่ยน ไ- ใ- ไ-ย -ย เป็น -ัย (3) เปลี่ยน รร เป็น -ัน (4) ตัดการันต์ พยัญชนะที่มีการันต์กำกับ รวมทั้งสระและอักษรที่ควบการันต์ทิ้ง (5) ใช้รหัสความยาว 7 หลัก รหัสตัวแรกเป็นตัวอักษรซึ่งใช้ตารางเทียบรหัสดังแสดงในตารางที่ 2.2 ส่วนรหัสตัวที่เหลือเป็นตัวเลขซึ่งใช้ตารางที่ 2.3 ในการเทียบรหัส ถ้ารหัสที่ได้มีความยาวน้อยกว่า 7 หลัก ให้เติม 0 จนครบ นอกจากนี้ยังได้เพิ่มกฎเกณฑ์เพื่อให้เหมาะสมกับภาษาไทย ได้แก่

- กรณีพบ ใ- ไ- ไ-ย และ -ัย จะเปลี่ยนให้อยู่ในรูปแบบเดียวกันคือ -ัย ก่อนทำการเข้ารหัส เนื่องจากสระดังกล่าวอ่านออกเสียงเหมือนกัน เช่น ใท ไท ไทย และ ทัย จะเปลี่ยนเป็น ทัย
- กรณีพบ รร จะเปลี่ยนเป็น -ัน ในกรณีที่ไม่มีตัวสะกดตามหลัง และเปลี่ยนเป็น -ั กรณีที่มีตัวสะกดตามหลัง เช่น เปลี่ยนคำว่า สรรเพชร รังสรรค์ พรรณนที และ ธรรมรัตน์ เป็น สันเพชร รังสัน พัณนที และ ธัมรัตน์ ตามลำดับ
- กรณีพบการันต์ จะตัดการันต์และพยัญชนะที่มีตัวการันต์กำกับรวมทั้งสระและอักษรควบการันต์ทิ้ง เช่น คำว่า จันทร์ ศักดิ์ และ พันธุ์ เปลี่ยนเป็น จัน ศัก และ พัน ตามลำดับ

ตัวอย่างการเข้ารหัสคำ เช่น คำ "อัมพร" หรือ "อำภรณ์" จะได้รหัสเป็น "อ059000" คำ "พรรณศักดิ์" หรือ "พันธุ์ศักดิ์" จะได้รหัสเป็น "พ341000" คำ "เนืองนิตย์" หรือ "เนืองนิจ" จะได้รหัสเป็น "น623400"

ตารางที่ 2.1 การกำหนดรหัสซาวด์เด็กซ์ภาษาอังกฤษของ Odell และ Russel

ตัวอักษร	รหัสตัวเลข
A E I O U H W Y	0
B F P V	1
C G J K Q S X Z	2
D T	3
L	4
M N	5
R	6

ตารางที่ 2.2 การกำหนดรหัสตัวอักษรของรหัสซาวด์เด็กซ์ภาษาไทย
จากงานวิจัยของวรรณี อุดมพาณิชย์

รหัสตัวอักษร	ตัวอักษร
ก	ก
ข	ข ฃ ค ฅ ฆ
ง	ง
จ	จ
ช	ช ฌ ฉ
ส	ซ ศ ษ ส
ด	ด ฎ
ต	ต ฏ
ท	ฐ ฑ ฒ ณ ท ธ
น	ณ น

รหัสตัวอักษร	ตัวอักษร
บ	บ
ป	ป
พ	พ ภ ผ
ฟ	ฝ ฟ
ม	ม
ย	ญ ย
ร	ร ล ฬ ฤ ฦ
ว	ว
อ	อ
ฮ	ห ฮ

ตารางที่ 2.3 การกำหนดรหัสตัวเลขของรหัสซาวด์เด็กซ์ภาษาไทย
จากงานวิจัยของของวรรณี อุดมพาณิชย์

รหัสตัวเลข	ตัวอักษร
0	ม ว –ำ
1	ก ข ฃ ค ฅ ฆ
2	ง ย
3	ญ ณ น
4	ฎ ฏ ด ต ศ ษ ส
5	บ ป พ ภ
6	ผ ฝ ฟ ห อ ฮ
7	จ ฉ ช ฌ ฒ
8	ฐ ฑ ฒ ถ ท ธ
9	ร ฤ ล ฦ ฬ

2.11 งานวิจัยของ นิลเนตร อรุณวงศ์ ณ อยุธยา

งานวิจัย [17] มีการปรับวิธีการจาก [16] ได้แก่ (1) ใช้อักขระไทยทั้งหมดในการเข้ารหัส โดยใช้ตารางที่ 2.4 เป็นตารางกำหนดรหัสสำหรับอักขระตัวแรกของรหัส และตารางที่ 2.5 สำหรับอักขระตัวที่เหลือของรหัส (2) ตัดตัวควบกล้ำทิ้ง เช่น ปรับ เป็น ปับ หรือ คลอง เป็น คอง (3) อักษรนำเสียงสนิท (ได้แก่ อย, หง, หญ, หน, หม, หย, หร, หล, หว ให้ตัด "อ" หรือ "ห" ทิ้ง (4) เปลี่ยนตำแหน่งสระหน้า (เช่น เ- แ- โ- ไ- ใ-) ไปไว้หลังสุด (5) ตัวอักษรติดกันและเหมือนกันให้ยุบเหลือรหัสตัวเดียว (6) ไม่นำสระ -ะ และ ิ- มาเข้ารหัส ตัวอย่างการเข้ารหัสคำ เช่น "กิตติพรรณ" แทนด้วย "กตบน" ซึ่งได้จากการตัด ิ- , ตัว "ต" ติดกัน 2 ตัว ตัดเหลือตัวเดียว, ตัว "พ" แทนด้วยรหัส "บ", แทน "รร" ด้วย ั-น แล้วตัดตัว ั- ออก

ตัวอย่างการเข้ารหัสคำ เช่น คำ "อัมพร" หรือ "อำภรณ์" จะได้รหัสเป็น "อมบน" คำ "พรรณศักดิ์" หรือ "พันธุ์ศักดิ์" จะได้รหัสเป็น "พนดก" คำ "เนืองนิตย์" หรือ "เนืองนิจ" จะได้รหัสเป็น "นืองนดเ"

ตารางที่ 2.4 การกำหนดรหัสสำหรับอักขระตัวแรก
จากงานวิจัยของนิลเนตร อรุณวงศ์ ณ อยุธยา

รหัสตัวอักษร	ตัวอักษร
ก	ก
ข	ข ฃ ค ฅ ฆ
ง	ง
จ	จ
ช	ช ฌ ฉ
ซ	ซ ศ ษ ส
ด	ด ฎ
ต	ต ฏ
ท	ฐ ฑ ฒ ณ ถ ท ธ
น	ณ น

รหัสตัวอักษร	ตัวอักษร
บ	บ
ป	ป
พ	พ ภ ผ
ฟ	ฝ ฟ
ม	ม
ย	ญ ย
ร	ร ล ฬ ฤ ฦ
ว	ว
อ	อ
ฮ	ห ฮ

ตารางที่ 2.5 การกำหนดรหัสสำหรับอักขระถัดจากตัวแรก
จากงานวิจัยของนิลเนตร อรุณวงศ์ ณ อยุธยา

รหัสตัวอักษร	ตัวอักษร
ก	ก ข ค ฆ
ง	ง
ด	จ ฉ ช ซ ฌ ฎ ฏ ฐ ฑ ฒ ฌ ด ต ถ ท ธ ศ ส ษ
น	ญ ณ น ร ล ฬ
บ	บ ป พ ฟ ภ ผ ฝ
ม	ม ำ
ย	ย
ว	ว
ไ	ไ-ย -ย ใ-

2.12 งานวิจัยของ ประยุทธ สุวรรณวิสาท และ สมชาย ประสิทธิ์จูตระกูล

สำหรับงานวิจัย [3] และ [4] ได้แบ่งการทดลองออกเป็น 2 ส่วนหลักๆ คือ การค้นคืนข้ามภาษาไทยทับศัพท์ภาษาอังกฤษ และ การค้นคืนข้ามภาษาอังกฤษทับศัพท์ภาษาไทย โดยที่กรณี

การค้นคืนข้ามภาษาไทยทับศัพท์ภาษาอังกฤษ นั้นใช้รหัสคำเป็นตัวเลขทั้งหมดโดยไม่จำกัดความยาวของรหัสคำ ตารางที่ 2.6 แสดงการกำหนดรหัสสำหรับคำไทยทับศัพท์คำอังกฤษ สำหรับการค้นคืนนั้นใช้การเปรียบเทียบรหัสคำแบบเหมือนกันทุกประการ นั่นคือรหัสคำต้องตรงกันทุกตัวจึงจะถือว่าสองคำนั้นมีเสียงอ่านตรงกัน จากผลการทดลองพบว่าเมื่อใช้รหัสคำความยาวมากกว่า 4 หลักขึ้นไป ได้ค่าความเที่ยง 78% และค่าเรียกคืน 90% สำหรับกรณีการค้นคืนข้ามภาษาอังกฤษทับศัพท์ภาษาไทย ใช้รหัสคำเป็นตัวอักษรไทยผสมกับสัญลักษณ์เสียงสากล

ตารางที่ 2.7 และตารางที่ 2.8 แสดงรหัสสำหรับพยัญชนะและสระตามลำดับ สำหรับคำไทยมีการประมวลผลเบื้องต้นก่อนการเข้ารหัสคำ ซึ่งได้แก่ การลดรูป การตัดวรรณยุกต์และตัวการันต์ การแทนที่ ไ- ใ- ไ-ย -ย ด้วย -ัย แทนที่ รร เป็น -ัน เป็นต้น ในการเปรียบเทียบรหัสคำนั้นใช้วิธีการเปรียบเทียบเชิงประมาณด้วยเทคนิคระยะแก้ไขสั้นสุด และมีการกำหนดต้นทุนในการแทนที่อักขระสำหรับแต่ละคู่อักขระตามกฎเกณฑ์ที่ได้สร้างขึ้น ค่าของต้นทุนมี 4 ระดับคือ C_1 C_2 C_3 และ C_4 ซึ่งมีค่า 0 1 4 และ 7 ตามลำดับ ดังตัวอย่างในรูปที่ 2.3 และรูปที่ 2.4 จากผลการทดลองพบว่าได้ค่าความเที่ยง 69% และค่าเรียกคืน 73% สำหรับกรณีคำไทยทับศัพท์คำอังกฤษ และได้ค่าความเที่ยง 96% และค่าเรียกคืน 75% สำหรับกรณีคำอังกฤษทับศัพท์คำไทย

2.13 งานวิจัยของ ทัศนวรรณ ศูนย์กลาง สมชาย ประสิทธิ์จูตระกูล และบุญเสริม กิจศิริกุล

งานวิจัย [5] เสนอการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กแบบแพร่กระจายย้อนกลับ โดยการพิจารณาตัวอักษรที่อยู่ข้างเคียงกับตัวที่กำลังพิจารณา ในงานวิจัยนี้ใช้ตัวอักษรที่อยู่ข้างเคียงข้างหน้าและข้างหลังอย่างละ 4 ตัว ดังนั้นเมื่อรวมตัวอักษรที่กำลังพิจารณาแล้ว ข้อมูลเข้าของนิวรอลเน็ตเวิร์กจึงมี 9 ตัวอักษร ส่วนข้อมูลออกคือรหัสเสียงจากข้อมูลขาเข้า ดังแสดงในรูปที่ 2.5 และรูปที่ 2.6 นอกจากนี้การเข้ารหัสคำของคำไทยยังมีการใช้การประมวลผลเบื้องต้นด้วยเช่นเดียวกับในงานวิจัย [12] ในการค้นคืนใช้วิธีการเปรียบเทียบรหัสคำแบบประมาณด้วยวิธีระยะแก้ไขสั้นที่สุด โดยยอมให้ความต่างของรหัสคำมีได้ไม่เกิน 1 และกำหนดค่าต้นทุนการแก้ไขอักขระทั้งการเพิ่ม ลบ และแทนที่ ให้มีค่าเท่ากับ 1 จากผลการทดลองพบว่าได้ค่าความเที่ยง 87.28% และค่าเรียกคืน 77.19% สำหรับกรณีคำไทยทับศัพท์คำอังกฤษ และได้ค่าความเที่ยง 96.34% และค่าเรียกคืน 75.15% สำหรับกรณีคำอังกฤษทับศัพท์คำไทย

ตารางที่ 2.6 การกำหนดรหัสสำหรับคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ จากงานวิจัยของ ประยุทธ สุวรรณวิสาทร

ภาษาอังกฤษ	ภาษาไทย	รหัส
AEIOUHWY[2]	อ ห ฮ ย ญ	0
BFPV	บ ฝ ฟ ป ผ พ ภ ว	1
CGJKQSXZ	ข ฃ ค ฅ ฆ ฉ ช ฌ ก จ ซ ศ ษ ส	2
DT	ฎ ด ฏ ต ฐ ฑ ฒ ฌ ถ ท ธ	3
L	ล ฬ	4
MN	ม ณ น	5
R	ร	6
AEIOU[1]	อ	7
H[1]	ห ฮ	8
W[1]	ว	1
Y[1]	ย ญ	9
	ง	52

1 : สำหรับตัวอักษรแรกของคำ

2 : สำหรับตัวอักษรตั้งแต่ตัวที่ 2 เป็นต้นไปของคำ

ตารางที่ 2.7 การกำหนดรหัสของพยัญชนะสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของ ประยุทธ สุวรรณวิสาทร

อักษรอังกฤษ	อักษรไทย	รหัส
b	บ	บ
bh	พ	พ
c	ซ	ซ
ch	ช ฉ ฌ	ช
ck	ก	ก
d	ด ฎ	ด
dh	ท	ท
f	ฟ ฝ	ฟ

อักษรอังกฤษ	อักษรไทย	รหัส
n	น ณ	น
ng	ง	ง
p	ป	ป
ph	พ ผ ภ	พ
q	ค	ค
r	ร ฤ	ร
s	ส ซ ศ ษ	ส
t	ต ฏ	ต

ตารางที่ 2.7 (ต่อ) การกำหนดรหัสของพยัญชนะสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของ ประยุทธ สุวรรณวิสาทร

อักษรอังกฤษ	อักษรไทย	รหัส
g	ก	ก
h	ห ฮ	ห
j	จ	จ
k	ก	ก
kh	ข ฃ ค ฅ ฆ	ข
l	ล ฦ ฬ	ล
m	ม	ม

อักษรอังกฤษ	อักษรไทย	รหัส
th	ท ฐ ฑ ฒ ถ ธ	ท
v	ว	ว
w	ว	ว
x	ก	ก
y	ย ญ	ย
z	ซ	ซ

ตารางที่ 2.8 การกำหนดรหัสของสระสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของ ประยุทธ สุวรรณวิสาทร

ตัวอักษรอังกฤษ	ตัวอักษรไทย	รหัส
-a	ะ	ะ
-aa	า	า
-ae	แ-ะ แ-	x
-ai	-ัย	-ัย
-ao	เ-า	@
-aiu	เ-ียว	I
-arn	าน	าน
-art	าท	าท
-e	เ-ะ เ-	เ
-ee	-ี	-ี
-eo	แ-ว	แ-ว
-er	เ-อ เ-ิ	q

ตัวอักษรอังกฤษ	ตัวอักษรไทย	รหัส
-eu	-ึ	-ึ
-i	-ิ	-ิ
-ia	เ-ียะ เ-ีย	I
-ie	เ-ียะ เ-ีย	I
-o	-อ	อ
-oe	เ-อ เ-	Q
-oi	อย	อย
-oo	-ู	-ู
-orn	-อน	ร
-u	-ึ -ื -ุ -ู	-ุ
-ua	เ-ือะ เ-ือ -ัวะ -ัว	U
-ue	-ื	-ื

	ก	ข	ค	...	ท	ธ	น	บ	...
ก	C_1	C_2	C_2	...	C_4	C_4	C_4	C_4	...
ข	C_2	C_1	C_1	...	C_4	C_4	C_4	C_4	...
ค	C_2	C_1	C_1	...	C_4	C_4	C_4	C_4	...
...	...	...	...	...	...	...	...	...	...
ท	C_4	C_4	C_4	...	C_1	C_1	C_4	C_4	...
ธ	C_4	C_4	C_4	...	C_1	C_1	C_4	C_4	...
น	C_4	C_4	C_4	...	C_4	C_4	C_1	C_4	...
บ	C_4	C_4	C_4	...	C_4	C_4	C_4	C_1	...
...	...	...	...	...	...	...	...	...	...

รูปที่ 2.3 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับพยัญชนะ
จากงานวิจัยของประยุทธ สุวรรณวิสาทร

	ะ	ั	า	ิ	ี	ึ	ื	ุ	...
ะ	C_1	C_1	C_1	C_4	C_4	C_4	C_4	C_4	...
ั	C_1	C_1	C_2	C_4	C_4	C_4	C_4	C_2	...
า	C_1	C_2	C_1	C_4	C_4	C_4	C_4	C_4	...
ิ	C_4	C_4	C_4	C_1	C_1	C_3	C_4	C_4	...
ี	C_4	C_4	C_4	C_1	C_1	C_4	C_4	C_4	...
ึ	C_4	C_4	C_4	C_3	C_4	C_1	C_1	C_2	...
ื	C_4	C_4	C_4	C_4	C_4	C_1	C_1	C_2	...
ุ	C_4	C_2	C_4	C_4	C_4	C_2	C_2	C_1	...
...	...	...	...	...	...	...	...	...	...

รูปที่ 2.4 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับสระ
จากงานวิจัยของประยุทธ สุวรรณวิสาทร

หอสมุดกลาง สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ลำดับตัวอักษร	รหัส
_ , _ , _ , _ , ก , น , ก , พ , ี	k
_ , _ , _ , ก , น , ก , พ , ี , ร	a, n
_ , _ , ก , น , ก , พ , ี , ร , ะ	o, k
_ , ก , น , ก , พ , ี , ร , ะ , ว	p
ก , น , ก , พ , ี , ร , ะ , ว , ุ	i
น , ก , พ , ี , ร , ะ , ว , ุ , ฒ	r
ก , พ , ี , ร , ะ , ว , ุ , ฒ , ิ	a
พ , ี , ร , ะ , ว , ุ , ฒ , ิ , _	v
ี , ร , ะ , ว , ุ , ฒ , ิ , _ , _	u
ร , ะ , ว , ุ , ฒ , ิ , _ , _ , _	t
ะ , ว , ุ , ฒ , ิ , _ , _ , _ , _	_

รูปที่ 2.5 ตัวอย่างการเข้ารหัสคำไทยของทัศนวรรณ ศูนย์กลาง และคณะ

ลำดับตัวอักษร	รหัส
_ , _ , _ , _ , R , H , O , D , I	r
_ , _ , _ , R , H , O , D , I , U	_
_ , _ , R , H , O , D , I , U , M	o
_ , R , H , O , D , I , U , M , _	d
R , H , O , D , I , U , M , _ , _	l
H , O , D , I , U , M , _ , _ , _	_
O , D , I , U , M , _ , _ , _ , _	m

รูปที่ 2.6 ตัวอย่างการเข้ารหัสคำอังกฤษของทัศนวรรณ ศูนย์กลาง และคณะ

2.14 สรุป

ในบทนี้ได้กล่าวถึงทฤษฎีที่เกี่ยวข้องกับงานวิจัยนี้ ได้แก่ การถอดอักษร การถ่ายเสียง นิวรอลเน็ตเวิร์ก การวัดผลการค้นคืน และขั้นตอนวิธีระยะแก้ไขสั้นสุด จากนั้นได้กล่าวถึงงานวิจัยในอดีตที่เกี่ยวข้องกับการเข้ารหัสคำ และการค้นคืนข้ามภาษาไทย-อังกฤษ สำหรับในบทต่อๆ ไป จะกล่าวถึงการนำทฤษฎี และงานวิจัยที่เกี่ยวข้องมาประยุกต์ใช้ในการทำงานวิจัยนี้