19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

การแปลภาษาไทย-อีสานโดยใช้ฐานกฎ
Thai - Isan Machine Translation Using Rule-Based Approach

ทัศนวรรณ ศูนย์กลาง[1*] สุนีย์ พงษ์พินิจภิญโญ[2] และ วีณาวดี ม่วงอ้น[3]
[123]ภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร, วิทยาเขตพระราชวังสนามจันทร์, นครปฐม
Email: soonklang_t@su.ac.th[1*], pongpinigpinyo_s@su.ac.th[2], muangon_w@su.ac.th[3]

บทคัดย่อ

ภาษาอีสานเป็นภาษาพูดท้องถิ่นที่ใช้มากในภาคอีสาน หรือภาคตะวันออกเฉียงเหนือของประเทศไทย ในบางครั้งเป็นการยากที่จะทำความเข้าใจหรือสื่อสารระหว่างคนท้องถิ่นที่พูดภาษาอีสานกับคนอื่นๆที่ไม่ใช่คนท้องถิ่นนั้น งานวิจัยนี้นำเสนอสถาปัตยกรรมของระบบการแปลภาษา 2 ทางระหว่างภาษาอีสานและภาษาไทยที่เป็นภาษาทางการ ระบบประกอบด้วย 2 ส่วน ส่วนแรกคือส่วนที่สร้างกฎไวยากรณ์จากโครงสร้างประโยคภาษาไทย ส่วนที่สองคือส่วนของการแปลภาษาที่ใช้กฎไวยากรณ์และพจนานุกรม 2 ภาษา กฎไวยากรณ์จำนวน 34 กฎได้ถูกนำมาใช้ในการแปลภาษาไทย-อีสาน ซึ่งในงานวิจัยนี้สามารถแปลประโยคสนทนาที่เป็นภาษาไทย-อีสานทั่วไปที่ใช้ในชีวิตประจำวัน ประกอบด้วยประโยค 4 ประเภท ได้แก่ ประเภทประโยคบอกเล่า ประโยคปฏิเสธ ประโยคคำถาม และประโยคขอร้อง/คำสั่ง ผลการทดสอบประสิทธิภาพของระบบการแปลภาษาไทย-ภาษาอีสานมีความถูกต้องในการแปลภาษาไทยเป็นภาษาอีสาน และภาษาไทยเป็นภาษาอีสานที่ 62.5% และ 70% ตามลำดับ งานวิจัยนี้จะมีประโยชน์สำหรับบุคคลทั่วไปที่ต้องการเรียนรู้หรือสื่อสารด้วยภาษาอีสาน

คำสำคัญ: โปรแกรมแปลภาษาไทย – อีสาน, พจนานุกรม, การแปลภาษาด้วยเครื่องแบบใช้ฐานกฎ, ระยะการแก้ไข

Abstract

Isan language is dialects spoken in the northeastern Thailand. It is sometimes difficult to understand or communicate between northeastern dialects people and people who are not local northeastern. This research proposes the architecture of online 2-way Isan dialect translation to Thai official language as it is a central language. The system consists of 2 modules. The first module is that grammar rules are created from Thai sentence structure. The last module is the translation system which uses created grammar rules and bilingual dictionary for translation. There are 34 grammar rules that are used in the translation. The proposed machine translation can translate the sentences are used for everyday life conversations. The sentences contain four types: affirmative sentence, negative sentence, interrogative sentence, and imperative sentence. The performance evaluation of translation system shows that the accuracy of the Thai-Isan translation and Isan-Thai translation is 62.5% and 70% respectively. The outcome from this research is that it will help people who would like to learn or communicate Isan language

Keywords : Thai – Isan Machine Translation, Dictionary, Rule-based Machine Translation, Edit Distance

1. คำนำ

ในปัจจุบันโปรแกรมแปลภาษาด้วยเครื่องสำหรับภาษาไทย มีการแปลไปยังภาษาเป้าหมายซึ่งเป็นภาษาต่างประเทศหลากหลายภาษา แต่ทว่ายังไม่มีโปรแกรมใดที่รองรับการแปลภาษาไทยเป็นภาษาถิ่นตามภาคต่างๆ ของไทย เช่น ภาษาเหนือ ภาษาใต้ ภาษาอีสาน ภาษาไทยและภาษาถิ่นโดยทั่วไปจะมีความใกล้เคียงกันในเชิงโครงสร้างในแง่การเรียงลำดับคำ ส่วนที่แตกต่างกันคือในแง่ของคำศัพท์ที่ใช้ เนื่องจากคำศัพท์ในภาษาถิ่นไม่มีพจนานุกรมบัญญัติคำศัพท์ไว้ การสะกดคำศัพท์คำเดียวกันอาจจะสะกดได้หลายแบบ แต่อ่านออกเสียงได้เหมือนกัน (คำพ้องเสียง) หรือใกล้เคียงกัน ตัวอย่างเช่น คำในภาษาอีสานดังต่อไปนี้ พ่อเถ้า พ่อเฒ่า – พ่อตา, หม้อง ม่อง -- ที่ แห่ง, ม่อ หม้อ –ใกล้, ไผ ใผ – ใคร นอกจากนี้อาจจะมีการผันวรรณยุกต์ที่แตกต่างกันได้บ้าง ขึ้นกับการออกเสียงของแต่ละคน เช่น ฮู้จัก ฮู่จั๋ก – รู้จัก ในงานวิจัยนี้ จึงได้ทดลองพัฒนาโปรแกรมแปลภาษาไทย-ภาษาไทยอีสาน เพื่อเป็นต้นแบบในการพัฒนาโปรแกรมแปลภาษาไทยเป็นภาษาถิ่นอื่นๆ โดยระบบแปลภาษาที่นำเสนอในงานวิจัยนี้สามารถรองรับการสะกดคำที่แตกต่างกันได้ โดยใช้วิธีการหาค่าระยะห่างน้อยที่สุดมาประยุกต์ใช้เพื่อแก้ปัญหาการสะกดคำศัพท์ที่แตกต่างกันของผู้ใช้งาน

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการแปลภาษาด้วยเครื่อง (Machine Translation) เป็นการใช้ซอฟต์แวร์คอมพิวเตอร์เพื่อแปลภาษาที่ใช้ในการสื่อสารของมนุษย์ระหว่างภาษาหนึ่งไปยังอีกภาษาหนึ่งโดยมีความหมายที่ตรงกันและเข้าใจกัน ภาษาที่ใช้สื่อสารอยู่ในรูปของข้อความ หรือคำพูดภาษาธรรมชาติ เช่น การแปลภาษาสื่อสารจากภาษาไทยไปเป็นภาษาญี่ปุ่น หรือการแปลภาษาสื่อสารจากภาษาท้องถิ่นในแต่ละภูมิภาคของประเทศนั้นๆ เช่น การแปลภาษาสื่อสารระหว่างภาษาของภาคอีสานและภาษาภาคกลางในประเทศไทย การแปลภาษาด้วยเครื่องมี 3 ระบบคือ 1.ระบบการแปลโดยตรง (Direct Machine Translation) ใช้พจนานุกรม 2 ภาษาที่เป็นภาษาต้นฉบับ และภาษาเป้าหมาย 2. ระบบการแปลแบบถ่ายทอด (Transfer Machine Translation) แบ่งเป็น 3 ขั้นตอนคือขั้นตอนการเปลี่ยนรูปแบบภาษาต้นฉบับไปเป็นรูปแบบการแทนภาษา นำรูปแบบการ

แทนภาษาไปเปลี่ยนเป็นภาษาเป้าหมาย และการวิเคราะห์สร้างภาษาเป้าหมายตามลักษณะที่เหมาะสม 3. ระบบการแปลโดยใช้ภาษากลาง (Interlingua Machine Translation) แบ่งเป็น 2 ขั้นตอนคือ ขั้นตอนการวิเคราะห์รูปลักษณะของภาษาต้นฉบับเพื่อแทนค่าด้วยภาษากลางที่เป็นตัวแทนของความหมาย และขั้นตอนการนำภาษากลางไปสร้างเป็นภาษาเป้าหมายโดยใช้ความสัมพันธ์ของรูปแบบคำต่างๆที่ไม่ขึ้นอยู่กับภาษาใดภาษาหนึ่งโดยเฉพาะ [1] ในงานวิจัยนี้นำระบบการแปลแบบถ่ายทอดมาใช้กับการแปลระหว่างภาษาไทย-อีสาน โดยใช้กฎการถ่ายทอดที่กำหนดไว้ (Rule-base method) และพจนานุกรม เพื่อนำมาใช้วิเคราะห์ประโยคต้นฉบับในส่วนลักษณะของคำ (Morphology), วากยสัมพันธ์ (Syntactical), และความหมาย (Semantic) แล้วนำไปสร้างเป็นภาษาเป้าหมายตามโครงสร้างภาษา

Mouiad Fadiel Alawneh et al. [2] นำเสนองานวิจัยที่ใช้กฎการถ่ายทอด (Rule-Base method) ประยุกต์ใช้กับการแปลภาษาระหว่างภาษาอังกฤษและภาษาอารบิค (หรือภาษาอาหรับ) โดยงานวิจัยนี้ออกแบบผสมการทำงานรวมกันระหว่าง Rule-Based และ Example-Based มาประยุกต์ใช้เพื่อให้เกิดความสมดุลย์ของทั้งสองวิธีที่นำมาใช้ในการแปลข้อความจากเครื่องแปลภาษาและการจัดการของปัญหาเกี่ยวกับข้อตกลงคำและเนื้อเรื่องในการแปลประโยคจากภาษาอังกฤษเป็นภาษาอารบิค อย่างไรก็ตามจำนวนกฎที่ถูกสร้างยังมีน้อยมาก ทำให้ประสิทธิภาพยังไม่เพียงพอ

Tien-Ping Tan et al. [3] นำเสนองานวิจัยใช้ระบบสำหรับการแปลภาษามลายู และระบบการสังเคราะห์ โดยระบบจะแปลงประโยคที่เทียบกับภาษามลายู และสังเคราะห์การพูดที่สอดคล้อง ระบบแปลงข้อความเป็นหน่วยเสียงโดยใช้กฎ ซึ่งสามารถแบ่งกฎออกเป็น 6 กลุ่ม คือ 1. กฎที่ไม่พิจารณาบริบท การแปลงข้อความโดยไม่พิจารณาบริบท 2. กฎที่ให้ความสำคัญในบริบท 3. กฎที่ให้ความสำคัญในบริบทที่ต้องการข้อมูลพยางค์ 4. กฎที่ให้ความสำคัญในบริบทที่ต้องการข้อมูลคำแสดงวิภัตติปัจจัย 5. กฎที่ให้ความสำคัญในบริบทที่ต้องการสองพยางค์และข้อมูลคำแสดงวิภัตติปัจจัย และ 6. กฎที่เฉพาะเจาะจงภาษาถิ่น

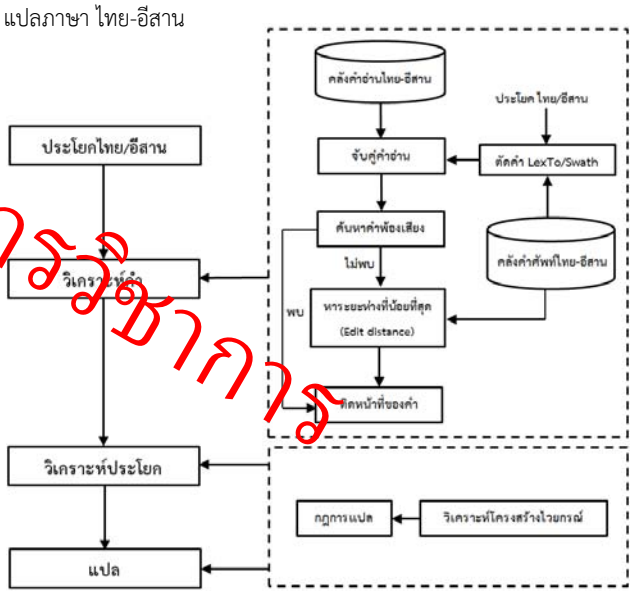
ณัฐพงษ์ จุฑางกูร [4] นำเสนอการพัฒนาระบบแปลภาษาด้วยเครื่องแบบใช้กฎไวยากรณ์สำหรับการแปลภาษาไทยเป็นภาษาญี่ปุ่น และภาษาญี่ปุ่นเป็นภาษาไทยผ่านระบบอินเตอร์เนต มีการสร้างพจนานุกรมที่ครอบคลุมคำศัพท์และวลีภาษาไทย และการพัฒนาฟังก์ชันวิเคราะห์คุณสมบัติของคำและประโยค แต่ระบบที่สร้างขึ้นยังไม่ครอบคลุมโครงสร้างทั้งหมดของภาษาไทย

ณัฐดนัย หอมคง [5] นำเสนอการพัฒนาระบบที่สามารถแปลข้อความภาษาไทยเป็นข้อความภาษามือไทยแบบอัตโนมัติโดยเลือกใช้การแปลแบบกฎไวยากรณ์ ข้อความภาษาไทยที่จะถูกแปลจะถูกตัดคำเป็นคำศัพท์ย่อยๆที่จะถูกนำไปตรวจสอบคำศัพท์ที่ยาวที่สุด มีการระบุหน้าที่ของคำศัพท์ และนำคำศัพท์ทั้งหมดที่ตรวจสอบแล้วสร้างเป็นข้อความใหม่และนำไปตรวจสอบกับโครงสร้างไวยากรณ์ภาษาไทยที่สอดคล้องและเพียงพอสำหรับข้อความในการสื่อสารระหว่างคนปกติกับผู้พิการทางการได้ยิน หลังจากนั้นข้อความใหม่จะถูกจัดเรียงตามโครงสร้างภาษามือไทย

Vincent Berment [6] นำเสนองานวิจัยที่ใช้ระบบการแปลโดยตรง สำหรับการแปลภาษาลาวไปเป็นภาษาฝรั่งเศส และการแปลภาษาฝรั่งเศสไปเป็นภาษาลาว งานวิจัยใช้พจนานุกรมลาวและฝรั่งเศสสำหรับการแปลภาษา

3. การออกแบบและวิเคราะห์ระบบแปลภาษาไทย-อีสานโดยใช้ฐานกฎ

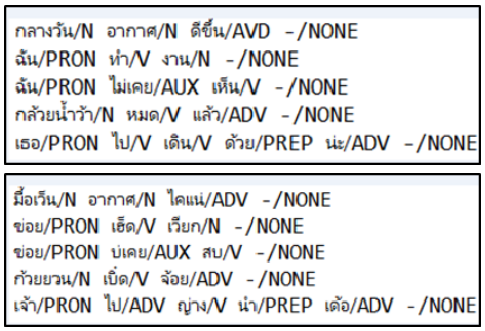
การทำงานโดยรวมของระบบจะเริ่มจากส่วนติดต่อกับผู้ใช้งานซึ่งมีหน้าที่รับความต้องการของผู้ใช้ ซึ่งก็คือ ประโยคภาษาไทย หรือ ภาษาอีสานที่จะถูกแปลไปยังภาษาเป้าหมาย และแสดงผลลัพธ์ของระบบ โดยผลลัพธ์จะมี 2 ส่วน คือ ประโยคที่แปลเป็นภาษาเป้าหมาย และคำศัพท์ที่ใกล้เคียงกันกับคำศัพท์ในข้อความประโยคที่รับเข้าเพื่อช่วยผู้ใช้ในการแปลเมื่อเกิดกรณีการสะกดคำศัพท์ไม่ถูกต้อง ในส่วนของค้นหาคำใกล้เคียงและระบบแปลภาษาจะประมวลผลข้อความประโยคที่รับเข้า เพื่อแสดงผลให้ผู้ใช้ โดยการสร้างระบบแปลภาษานี้จะประกอบไปด้วยส่วนสำคัญคือ การสร้างคลัง การวิเคราะห์คำ ค้นหาคำใกล้เคียง การวิเคราะห์ประโยค และการแปล รูปที่ 3.1 คือแบบจำลองโครงสร้างการทำงานของระบบการแปลภาษา ไทย-อีสาน



รูปที่ 3.1 แบบจำลองโครงสร้างการทำงานของระบบการแปลภาษา ไทย-อีสาน

3.1 คลังสำหรับโปรแกรมแปลภาษา ไทย-อีสาน

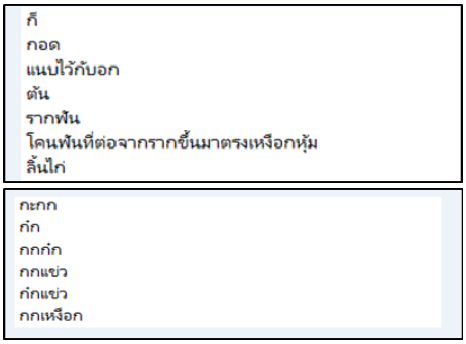
คลังประโยคภาษาไทย-อีสาน ในส่วนคลังประโยคภาษาไทยจะประกอบด้วยประโยคภาษาไทยจำนวน 100 ประโยค และตัวอย่างประโยคภาษาอีสาน [7][8][9] จำนวน 100 ประโยค โดยแต่ละประโยคจะติดหน้าที่ของคำแต่ละคำในประโยค คลังนี้จะถูกใช้ในขั้นตอนการค้นหาหน้าที่ของคำ มีโครงสร้างดังนี้ คำศัพท์/หน้าที่ของคำ คำศัพท์/หน้าที่ของคำ -/NONE (เมื่อจบประโยค) และขึ้นบรรทัดใหม่เมื่อเริ่มประโยคใหม่ ซึ่งตัวอย่างคลังประโยคภาษาไทย-อีสาน รูปที่ 3.2 ตัวอย่างการเก็บ คลังประโยคภาษาไทย-อีสาน



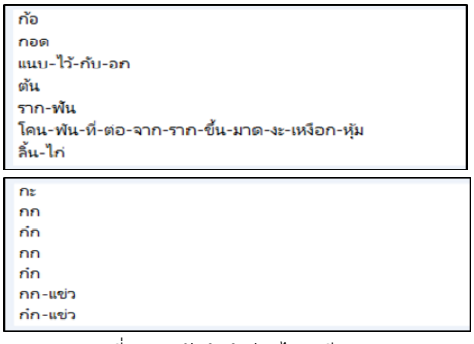
รูปที่ 3.2 คลังประโยค ไทย-อีสาน

คลังคำศัพท์ภาษาไทยและอีสานที่ใช้ในงานวิจัยนี้สร้างมาจากคำศัพท์ภาษาไทยจำนวน 1,895 คำ และคลังคำศัพท์ภาษาอีสานประกอบด้วยคำศัพท์ภาษาอีสานจำนวน 1,895 คำ คลังคำศัพท์เหล่านี้จะถูกนำไปใช้ในขั้นตอนการหาคำศัพท์ใกล้เคียง ซึ่งใช้วิธีการหาค่าระยะการแก้ไขที่น้อยที่สุด (Minimal Edit Distance) โดยโครงสร้างคลังคำศัพท์ทั้งสองภาษาแต่ละบรรทัดจะประกอบไปด้วยคำศัพท์เพียงแค่บรรทัดละ 1 คำศัพท์เท่านั้น รูปที่ 3.3 ตัวอย่างคลังคำศัพท์ไทย-อีสานที่ใช้ในงานวิจัย

คลังคำอ่านภาษาไทย-อีสาน คำอ่านในคลังนี้จะถูกนำไปใช้ในขั้นตอนการเปรียบเทียบคำอ่านของคำศัพท์ที่รับเข้า เพื่อวิเคราะห์คำพ้องเสียงโดยคลังคำอ่านภาษาไทยประกอบด้วย คำอ่านของคำศัพท์ภาษาไทยจำนวน 1,895 คำ และคลังคำอ่านภาษาอีสานประกอบด้วย คำศัพท์ภาษาอีสานจำนวน 1,895 คำ ซึ่งแต่ละบรรทัดจะประกอบไปด้วยคำอ่านภาษาไทยเพียงแค่บรรทัดละ 1 คำ ในกรณีที่คำศัพท์นั้นมีมากกว่า 1 พยางค์ จะมีโครงสร้างในรูปแบบ คำอ่านต่อคำอ่าน (คำอ่าน-คำอ่าน) ใน 1 บรรทัด ดังรูปที่ 3.4

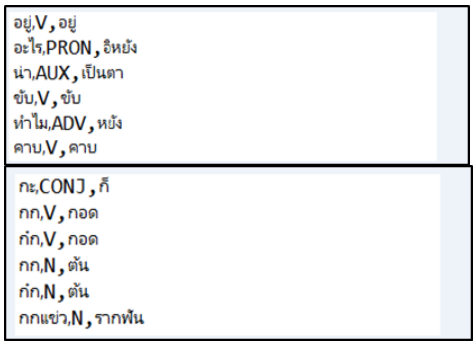


รูปที่ 3.3 คลังคำศัพท์ไทย-อีสาน



รูปที่ 3.4 คลังคำคำอ่านไทย-อีสาน

พจนานุกรมภาษาไทย และภาษาอีสาน โครงสร้างการเก็บของพจนานุกรมทั้งภาษาไทย และภาษาอีสานนี้จะประกอบด้วยการระบุหน้าที่ของคำ และความหมาย พจนานุกรมทั้งสองนี้จะถูกนำไปใช้ในขั้นตอนการหาความหมายของคำศัพท์ภาษาอีสานแปลไปเป็นภาษาไทย มีจำนวนภาษาไทยทั้งหมด 1895 คำ และภาษาอีสาน 1895 คำ มีโครงสร้างภายในพจนานุกรม 1 บรรทัดจะมีเก็บคำศัพท์ หน้าที่ของคำ ความหมาย ของคำศัพท์ 1 คำเท่านั้น ดังรูปที่ 3.5



รูปที่ 3.5 พจนานุกรมภาษาไทย-อีสาน

3.2 การวิเคราะห์คำ และออกแบบระบบค้นหาคำใกล้เคียง

เนื่องจากภาษาอีสานเป็นภาษาพูดท้องถิ่น ดังนั้น ปัญหาของภาษาอีสานคือ ไม่มีพจนานุกรมทางการที่กำหนดคำภาษาอีสานไว้ ดังนั้นคำของภาษาภาษาอีสานของแต่ละคนจึงอาจสะกดไม่เหมือนกัน แต่อย่างไรก็ตามก็จะมีคำพ้องเสียงเดียวกัน ดังนั้นในงานวิจัยนี้จึงนำเอาหลักการแปลงคำเป็นคำอ่านมาเพื่อแก้ไขปัญหาการสะกดไม่ตรงกัน ในกรณีที่ไม่สามารถหาคำพ้องเสียงในคลังคำศัพท์ได้ ระบบจะนำคำอ่านไปเข้าหลักการการหาระยะการแก้ไขที่น้อยที่สุด ซึ่งขั้นตอนนี้เป็นการวัดหาค่าความต่างของสายอักขระสองชุด โดยจะเลือกคำที่มีค่าความต่างน้อยที่สุดเป็นผลลัพธ์ จากนั้นนำผลลัพธ์ไปหาหน้าที่ของคำ โดยขั้นตอนนี้จะประกอบด้วยขั้นตอนต่อไปนี้

3.2.1 ขั้นตอนวิเคราะห์คำ

ในระบบการแปลภาษาไทย-อีสาน เริ่มแรกระบบจะดำเนินการตัดคำจากประโยคที่รับเข้ามา โดยประโยคภาษาไทยจะตัดคำโดยใช้โปรแกรม SWATH [10] และประโยคภาษาอีสานจะตัดคำโดยใช้โปรแกรม LexTo [11] เหตุผลที่เลือกใช้โปรแกรม LexTo ในการตัดคำภาษาอีสานเพราะสามารถปรับปรุงพจนานุกรมได้ ทำให้การตัดคำนั้นมีประสิทธิภาพมากขึ้น เช่น ฉันกินข้าว ตัดคำ ฉัน|กิน|ข้าว

3.2.2 จับคู่คำอ่าน และหาค้นหาคำพ้องเสียง

เมื่อตัดคำเสร็จเรียบร้อยแล้ว จากนั้นก็จะนำคำที่ได้จากการตัดคำนำมาเปลี่ยนเป็นคำอ่าน จะได้คำอ่านของคำแต่ละคำในประโยคที่ตัดคำได้นำคำอ่านของคำศัพท์ที่รับเข้ามา ไปเปรียบเทียบกับไฟล์คำอ่าน เพื่อนำมาวิเคราะห์คำพ้องเสียงเนื่องจากภาษาอีสานมักจะพบคำที่อ่านออกเสียงเหมือนกันแต่จะเขียนต่างกัน โดยใช้หลักการเปรียบเทียบคำอ่าน และ

ใช้การจับคู่ (Matching) มาช่วยในการกรองข้อมูลเพื่อให้ได้คำศัพท์ที่มีคำอ่านตรงกันกับคำที่รับเข้ามา เช่น

ตัดคำ ฉัน | ไป | โรงเรียน

คำอ่าน ฉัน | ไป | โรง – เรียน

ในกรณีคำอ่านของคำที่รับเข้ามาไม่ตรงกับคำอ่านของไฟล์คำอ่าน จะนำคำไปเข้าหลักการการหาค่าระยะการแก้ไขที่น้อยที่สุด โดยเลือกใช้วิธีการ Levenshtein edit distance [12] เป็นขั้นตอนการวัดหาค่าความต่างของสายอักขระสองชุด โดยที่ค่าความต่างกันจะวัดจากจำนวนของการที่จะต้องทำการตัดออก แทรก และแทนที่ จนกระทั่งอักขระมีลักษณะเหมือนกันทุกประการ โดยจะเลือกคำที่มีค่าระยะการแก้ไขที่น้อยที่สุดเป็นผลลัพธ์ จากนั้นนำผลลัพธ์ไปติดหน้าที่ของคำต่อไป เช่น

ข่าว เป็น ข้าว edit distance = 1 (แทนที่ ไม้เอก ด้วย ไม้โท 1 ครั้ง)

ข้อย เป็น ข่อย edit distance = 1 (แทนที่ ไม้โท ด้วย ไม้เอก 1 ครั้ง)

3.2.3 การติดหน้าที่ของคำ

ขั้นตอนของการติดหน้าที่ของคำเริ่มต้นจากนำผลลัพธ์ที่ได้มาจากการเปรียบเทียบคำอ่านและการหาระยะห่างที่น้อยที่สุด นำคำไปเทียบกับข้อมูลในคลังประโยค โดยคำที่นำไปเทียบนั้นจะพิจารณาบริบทรอบข้าง (Bigram Tagger) ว่าถ้าตามด้วยอีกคำแล้วคำที่เราต้องการติดหน้าที่ของคำด้วยชนิดใด เช่น คำนาม คำกิริยา ฯลฯ ถ้าการเทียบโดยสนใจบริบทรอบข้างแล้วไม่พบ ก็จะนำคำนั้นไปเทียบในคลังประโยคอีกครั้งโดยไม่สนใจบริบทรอบข้าง (Lookup Tagger) โดยหาว่าคำนั้นในคลังประโยคมีหน้าที่ของคำว่าอะไรมากที่สุด และถ้ากรณีที่เทียบโดยไม่สนใจบริบทรอบข้างแล้วยังไม่พบอีก ก็ติดหน้าที่ของคำเป็นค่าเริ่มต้น (default) โดยที่ค่าเริ่มต้นนี้ได้มาจากหน้าที่ของคำที่ถูกใช้มากที่สุดในคลังประโยค เมื่อได้ผลลัพธ์เป็นคำพร้อมติดหน้าที่ของคำแล้วก็จะนำไปเข้ากฎในการแปลเพื่อหาความหมายต่อไป

3.3 การวิเคราะห์ประโยค และการแปล

เมื่อได้โครงสร้างประโยคพร้อมกับหน้าที่ของคำแล้ว ก็จะนำโครงสร้างไวยากรณ์ของประโยคที่ได้มานั้นเข้าสู่กระบวนการแปลภาษาเพื่อแปลไปยังภาษาเป้าหมาย และดึงความหมายมาจากพจนานุกรม โดยอาศัยกฎไวยากรณ์ (Grammars) ของภาษา และการตัดคำ เพื่อตรวจสอบความถูกต้องของโครงสร้างไวยากรณ์ในประโยค โดยมีขั้นตอนดังนี้

3.3.1 การวิเคราะห์ไวยากรณ์ และความหมาย

การวิเคราะห์ไวยากรณ์ [13] คือ และความสัมพันธ์ระหว่างคำในประโยคเพื่อนำไปประมวลผลหาผลลัพธ์มาแสดงให้กับผู้ใช้ ซึ่งงานวิจัยนี้ใช้เทคนิคการเครื่องแปลภาษาโดยใช้ฐานกฎ (Rule-based machine translation) [14] ตามหลักภาษาไทยในการเทียบโครงสร้างไวยากรณ์เพื่อนำไปหาความหมาย และใช้พจนานุกรมสองภาษา (bilingual dictionary) ซึ่งผู้พัฒนาได้สร้างกฎไวยากรณ์เพื่อใช้ในการแปลความหมายจากตัวอย่างประโยค 80 ประโยค โดยจะแบ่งประโยคเป็น 5 ประเภท ได้แก่ ประโยคบอกเล่า 25 ประโยค ประโยคปฏิเสธ 20 ประโยค ประโยคคำถาม 15 ประโยค ประโยคขอร้อง 10 ประโยค และประโยคคำสั่ง 10 ประโยค ซึ่งกฎที่ได้ในงานวิจัยนี้มีทั้งหมด 34 กฎ ซึ่งมีตัวอย่างกฎ ดังนี้

กฎข้อที่ 1 ประโยคต้องแบ่งเป็นสองส่วนคือ ส่วนประธานและส่วนกรรม

(S -> NP + VP)

กฎข้อที่ 2 ส่วนประธาน ประกอบด้วย คำนามตามด้วยคำสรรพนาม

(NP -> N + PRON)

กฎข้อที่ 3 ส่วนประธาน ประกอบด้วย คำนามตามด้วยคำวิเศษณ์

(NP -> N + ADV)

กฎข้อที่ 4 ส่วนประธาน ประกอบด้วย คำนามตามด้วยกิริยาช่วย

(NP -> N + AUX)

กฎข้อที่ 5 ส่วนประธาน ประกอบด้วย คำนามตามด้วยคำนาม

(NP -> N + N)

กฎข้อที่ 6 ส่วนประธาน ประกอบด้วย คำนาม

(NP -> N)

กฎข้อที่ 7 ส่วนประธาน ประกอบด้วย คำกิริยาตามด้วยคำบุพบท

(NP -> V + PREP)

กฎไวยกรณ์ที่ได้จะถูกนำมาใช้ตรวจสอบความถูกต้องของไวยากรณ์ในประโยคเมื่อประโยคที่รับเข้ามานั้นผ่านการการตัดคำมาแล้ว

เช่น	“ฉันกินข้าว”	ตัดคำ	ฉัน|กิน|ข้าว
	ติดหน้าที่ของคำ	N | V | N	
	S		สัญลักษณ์เริ่มต้น
	-> NP + VP		ใช้กฎข้อที่ 1
	-> N + VP		ใช้กฎข้อที่ 6
	-> N(ฉัน) + V(กิน) + N(ข้าว)		ใช้กฎข้อที่ 19

3.3.2 การแปลภาษา

เมื่อผ่านกระบวนการตรวจสอบไวยากรณ์จากกฎไวยากรณ์ว่าประโยคนั้นมีการติดหน้าที่ของคำแล้วมีโครงสร้างกฎไวยากรณ์ที่ถูกต้อง ก็นำเข้าสู่กระบวนการแปล โดยนำคำที่ติดหน้าที่ของคำไปเทียบหาความหมายของคำหรือประโยคจากพจนานุกรม

เช่น	Input	ฉันกินข้าว	(ภาษาไทย)
	Output	ข่อยแดกเข่า	(ภาษาอีสาน)

4. การทดสอบ

ในงานวิจัยนี้กฎไวยากรณ์ที่ถูกสร้างขึ้นมาเพื่อใช้ในการแปลความหมายจากตัวอย่างประโยค 80 ประโยค โดยจะแบ่งประโยคเป็น 5 ประเภท ได้แก่ ประโยคบอกเล่า 25 ประโยค ประโยคปฏิเสธ 20 ประโยค ประโยคคำถาม 15 ประโยค ประโยคขอร้อง 10 ประโยค และประโยคคำสั่ง 10 ประโยคให้ผู้เชี่ยวชาญที่เป็นบุคคลผู้มีถิ่นฐานที่อยู่ภาคอีสาน จำนวน 2 คนเป็นผู้ทดสอบระบบแปลภาษาไทย-อีสาน โดยใช้แบบทดสอบวัดความถูกต้องในส่วนของการแปลภาษา โดยแบบทดสอบจะประกอบไปด้วยตัวอย่างประโยคทั้งภาษาไทยและภาษาอีสานทั้งหมดจำนวน 20 ประโยค โดยแบ่งรูปแบบของประโยคออกเป็น 3 รูปแบบ คือ ประโยคที่มีคำทุกคำอยู่ในพจนานุกรมของระบบจำนวน 10 ประโยค ประโยคที่ไม่มีคำ

ในพจนานุกรมจำนวน 5 ประโยค และประโยคที่มีคำที่เขียนผิดจำนวน 5 ประโยค และในการทดสอบระบบจะให้ผู้เชี่ยวชาญกรอกข้อความประโยคแล้วนำผลลัพธ์ที่ได้จากระบบมาเปรียบเทียบกับผลลัพธ์ที่ผู้เชี่ยวชาญคิดว่าถูกต้อง ถ้าผู้เชี่ยวชาญคิดว่าระบบแปลความหมายถูกต้องให้คิดเป็น 1 และถ้าผู้เชี่ยวชาญคิดว่าระบบแปลความหมายไม่ถูกต้องให้คิดเป็น 0 จากการเปรียบผลการทดสอบความถูกต้องของระบบในแต่ละประโยค ซึ่งสามารถหาได้จากสมการ $\bar{X} = \frac{\sum x}{N}$ โดยที่ X คือคะแนน และ N คือจำนวนผู้ประเมิน จากนั้นหาค่าเฉลี่ยจากการเปรียบเทียบผลการทดสอบประโยคตัวอย่างทั้ง 20 ประโยค จะได้เปอร์เซ็นต์ความถูกต้อง ในขั้นตอนการทดสอบความถูกต้องของระบบจะแบ่งการทดลองออกเป็น 2 รูปแบบ คือ รูปแบบที่ 1 การแปลภาษาจาก ภาษาไทยไปยังภาษาอีสาน มีความถูกต้องของการแปลที่ 62.5% รูปแบบที่ 2 การแปลภาษาจาก ภาษาอีสานไปยังภาษาไทย ได้ผลความถูกต้องที่ 70%

5. สรุป

จากระบบการแปลภาษาไทย-อีสานโดยใช้ฐานกฎและใช้พจนานุกรมสองภาษานั้นพบว่าระบบการแปลภาษาสามารถวิเคราะห์คำพ้องเสียงเนื่องจากภาษาอีสานมักจะพบคำที่อ่านออกเสียงเหมือนกันแต่จะเขียนต่างกัน โดยใช้หลักการเปรียบเทียบมากที่สุด และใช้การจับคู่คำอ่านมาช่วยในการกรองข้อมูลเพื่อให้ได้ข้อมูลที่มีความคล้ายคลึงหรือใกล้เคียงกับคำที่รับเข้ามามากที่สุด อย่างไรก็ตาม จากผลการทดลองการแปลภาษาอีสาน-ภาษาไทยมีความผิดพลาดน้อยกว่าการแปลจากภาษาไทยเป็นภาษาอีสาน ซึ่งส่วนหนึ่งเนื่องมาจากระบบได้มีการออกแบบเพื่อรองรับการสะกดผิดเนื่องจากภาษาอีสานเป็นภาษาพูดจึงยากต่อการสะกดให้ถูกต้องตามพจนานุกรมได้ จึงทำให้ไม่สามารถแปลคำศัพท์นั้นได้ถูกต้อง ส่วนการแปลภาษาไทย-อีสาน ความผิดพลาดส่วนใหญ่เกิดจากขั้นตอนการตัดคำภาษาไทยอัตโนมัติยังคงให้ผลลัพธ์ที่ไม่ถูกต้อง เป็นผลให้ขั้นตอนที่เป็นส่วนของการแปลไม่ถูกต้อง

เอกสารอ้างอิง

[1] Mohamed Amine Chéragui, Theoretical Overview of Machine translation, Proceedings of the 4th International Conference on Web and Information Technologies Sidi Bel Abbes, Algeria, April 29-30, 2012.

[2] Mouiad Fadiel Alawneh, Tengku Mohd Sembok, "Rule-Based and Example-Based Machine Translation from English to Arabic", Bio-Inspired Computing: Theories and Applications (BIC-TA), September 27-29, 2011, Penang, Malaysia, pp. 343 – 347.

[3] Tien-Ping Tan, Sang-Seong Goh,Yen-Min Khaw, "A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System", Asian Language Processing (IALP), November 13-15, 2012, Hanoi, Vietnam, pp. 109 – 112.

[4] ณัฐพงษ์ จุฑางกูร, "การแปลภาษาไทยเป็นภาษาญี่ปุ่นแบบใช้ฐานกฎด้วยเครื่องคอมพิวเตอร์ (Thai to Japanese machine translation using rule-based approach), วิทยานิพนธ์, วิศวกรรมศาสตรมหาบัณฑิต, สาขาวิศวกรรมคอมพิวเตอร์, บัณฑิตวิทยาลัย, มหาวิทยาลัยเชียงใหม่, 2553

[5] ณัฐดนัย หอมคง, การแปลภาษาด้วยเครื่องสำหรับข้อความภาษาไทยเป็นข้อความภาษามือไทย (Machine translation for Thai text - Thai sign language text), วิทยานิพนธ์, วิศวกรรมศาสตรมหาบัณฑิต, สาขาวิศวกรรมคอมพิวเตอร์, บัณฑิตวิทยาลัย, มหาวิทยาลัยเชียงใหม่, 2553

[6] Vincent Bermen, Online Translation Services for the Lao Language, First International Conference on Lao Studies, May 20-22, 2005

[7] สำลี รักสุทธี,2554, "พจนานุกรม ภาษาอีสาน – ไทยกลาง", กรุงเทพฯ, สำนักพิมพ์พัฒนาศึกษา, หน้า6-8

[8] Cyberlab, "ภาษาถิ่นไทย", [ออนไลน์], เข้าถึง: http://cyberlab.lh1.ku.ac.th/elearn/faculty/human/hm19/lesson1.htm สืบค้นวันที่ 1 ธันวาคม 2557

[9] ชมรมศิลปวัฒนธรรมอีสานจุฬาลงกรณ์มหาวิทยาลัย, "คลังประโยค", [ออนไลน์], เข้าถึง: http://www.isan.clubs.chula.ac.th/lang/index.php?transaction=search_word.php สืบค้นวันที่ 1 ธันวาคม 2557

[10] Paisarn Charoenpornsawat, "Feature-based Thai Word Segmentation", Master'sThesis, Computer Engineering, Chulalongkorn University, Bangkok, Thailand, 1999.

[11] Thai Lexeme Tokenizer (LexTo)", [ออนไลน์], เข้าถึง http://www.sansarn.com/lexto/download-lexto.php สืบค้นวันที่ 1 ธันวาคม 2557

[12] Mechael Piotrowski, "Natural Language Processing for Historical Texts", Morgan & Claypool Publishers, 2012.

[13] Suksiri Danthanavanich, "A GRAMMAR OF THAI SIGN LANGUAGE", Ph.D. Dissertation, Mahidol University, Bangkok, Thailand, 2008.

[14] Srisavakon Dangsaart, Kanlaya Naruedomkul, Nick Cercone, Booncharoen Sirinaovakul, "Bridging the Gap: Thai- Thai Sign Machine Translation", In Proceedings of The 10th Conference of the Pacific Association for Computational Linguistics, University of Melbourne, Australia, 2007.