

การแปลภาษาไทย-อีสานโดยใช้ฐานกฎ

Thai - Isan Machine Translation Using Rule-Based Approach

ทัศนวรรณ ศูนย์กลาง^{1*}, สุนีย์ พงษ์พินิจภิญโญ² และวีณาวิ ม่วงอัน³

^{1*,2,3} ภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์ อ.เมือง จ.นครปฐม

โทรศัพท์ (034) 245334 โทรสาร (034) 272923 Email: soonklang_t@su.ac.th^{1*}

บทคัดย่อ

ภาษาอีสานเป็นภาษาพูดท้องถิ่นที่ใช้มากในภาคอีสาน หรือภาคตะวันออกเฉียงเหนือของประเทศไทย ในบางครั้งเป็นการยากที่จะทำความเข้าใจหรือสื่อสารระหว่างคนที่พูดภาษาอีสานกับคนอื่น ๆ ที่ไม่ใช่คนท้องถิ่นนั้น งานวิจัยนี้นำเสนอสถาปัตยกรรมของระบบการแปลภาษา 2 ทางระหว่างภาษาอีสานและภาษาไทยที่เป็นภาษาทางการ ระบบประกอบด้วย 2 ส่วน ส่วนแรกคือส่วนที่สร้างกฎไวยากรณ์จากโครงสร้างประโยคภาษาไทย ส่วนที่สองคือส่วนของการแปลภาษาที่ใช้กฎไวยากรณ์และพจนานุกรม 2 ภาษา กฎไวยากรณ์จำนวน 34 กฎ ได้ถูกนำมาใช้ในการแปลภาษาไทย-อีสาน ซึ่งในงานวิจัยนี้สามารถแปลประโยคสนทนาที่เป็นภาษาไทย-อีสานทั่วไปที่ใช้ในชีวิตประจำวัน ประกอบด้วยประโยค 4 ประเภท ได้แก่ ประเภทประโยคบอกเล่า ประโยคปฏิเสธ ประโยคคำถาม และประโยคขอร้อง/คำสั่ง ผลการทดสอบประสิทธิภาพของระบบการแปลภาษาไทย-ภาษาอีสาน โดยใช้ฐานกฎมีความถูกต้องในการแปลภาษาไทยเป็นภาษาอีสาน และภาษาไทยเป็นภาษาอีสานที่ร้อยละ 62.5 และร้อยละ 70 ตามลำดับ งานวิจัยนี้จะมีประโยชน์สำหรับบุคคลทั่วไปที่ต้องการเรียนรู้ หรือสื่อสารด้วยภาษาอีสาน

คำสำคัญ : โปรแกรมแปลภาษาไทย – อีสาน, พจนานุกรม, การแปลภาษาด้วยเครื่องแบบใช้ฐานกฎ, ระยะเวลาแก้ไข

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. บทนำ

ในปัจจุบันโปรแกรมแปลภาษาด้วยเครื่องสำหรับภาษาไทย มีการแปลไปยังภาษาเป้าหมายซึ่งเป็นภาษาต่างประเทศหลากหลายภาษา แต่ทว่ายังไม่มีโปรแกรมใดที่รองรับการแปลภาษาไทยเป็นภาษาถิ่นตามภาคต่าง ๆ ของไทย เช่น ภาษาเหนือ ภาษาใต้ ภาษาอีสาน ภาษาไทยและภาษาถิ่น โดยทั่วไปจะมีความใกล้เคียงกันในเชิงโครงสร้างในแง่การเรียงลำดับคำ ส่วนที่แตกต่างกันคือในแง่ของคำศัพท์ที่ใช้ เนื่องจากคำศัพท์ในภาษาถิ่นไม่มีพจนานุกรมบัญญัติคำศัพท์ไว้ การสะกดคำศัพท์คำเดียวกันอาจจะสะกดได้หลายแบบ แต่อ่านออกเสียงได้เหมือนกัน (คำพ้องเสียง) หรือใกล้เคียงกัน ตัวอย่างเช่น คำในภาษาอีสานดังต่อไปนี้ พ่อเถ้า พ่อเฒ่า – พ่อตา, หม่อง ม่อง – ที่ แห่ง, ม่อ หม้อ – ไกล, ไผ ไผ – ใคร นอกจากนี้อาจจะมีการผันวรรณยุกต์ที่แตกต่างกันได้บ้าง ขึ้นกับการออกเสียงของแต่ละคน เช่น ฮู้จัก ฮู้จัก – รู้จัก ในงานวิจัยนี้ จึงได้ทดลองพัฒนาโปรแกรมแปลภาษาไทย – ภาษาไทยอีสาน เพื่อเป็นต้นแบบในการพัฒนาโปรแกรมแปลภาษาไทยเป็นภาษาถิ่นอื่น ๆ โดยระบบแปลภาษาที่นำเสนอในงานวิจัยนี้สามารถรองรับการสะกดคำที่แตกต่างกันได้ โดยใช้วิธีการหาค่าระยะห่างน้อยที่สุดมาประยุกต์ใช้เพื่อแก้ปัญหาการสะกดคำศัพท์ที่แตกต่างกันของผู้ใช้งาน

2. วัตถุประสงค์

เพื่อทดสอบประสิทธิภาพของระบบการแปลภาษาไทย-ภาษาอีสาน

3. งานวิจัยที่เกี่ยวข้อง

ในการแปลภาษาด้วยเครื่อง (Machine Translation) เป็นการใช้ซอฟต์แวร์คอมพิวเตอร์เพื่อแปลภาษาที่ใช้ในการสื่อสารของมนุษย์ระหว่างภาษาหนึ่งไปยังอีกภาษาหนึ่งโดยมีความหมายที่ตรงกันและเข้าใจกัน ภาษาที่ใช้สื่อสารอยู่ในรูปของข้อความ หรือคำพูดภาษาธรรมชาติ เช่น การแปลภาษาสื่อสารจากภาษาไทยไปเป็นภาษาญี่ปุ่น หรือการแปลภาษาสื่อสารจากภาษาท้องถิ่นในแต่ละภูมิภาคของประเทสนั้น ๆ เช่น การแปลภาษาสื่อสารระหว่างภาษาของภาคอีสานและภาษาภาคกลางในประเทศไทย การแปลภาษาด้วยเครื่องมี 3 ระบบคือ 1.ระบบการแปลโดยตรง (Direct Machine Translation) ใช้พจนานุกรม 2 ภาษาที่เป็นภาษาต้นฉบับ และภาษาเป้าหมาย 2.ระบบการแปลแบบถ่ายทอด (Transfer Machine Translation) แบ่งเป็น 3 ขั้นตอนคือขั้นตอนการเปลี่ยนรูปแบบภาษาต้นฉบับไปเป็นรูปแบบการแทนภาษา นำรูปแบบการแทนภาษาไปเปลี่ยนเป็นภาษาเป้าหมาย และการวิเคราะห์สร้างภาษาเป้าหมายตามลักษณะที่เหมาะสม 3.ระบบการแปลโดยใช้ภาษากลาง (Interlingua Machine Translation) แบ่งเป็น 2 ขั้นตอน คือขั้นตอนการวิเคราะห์รูปลักษณะของภาษาต้นฉบับเพื่อแทนค่าด้วยภาษากลางที่เป็นตัวแทนของความหมาย และขั้นตอนการนำภาษากลางไปสร้างเป็นภาษาเป้าหมายโดยใช้ความสัมพันธ์ของรูปแบบคำต่าง ๆ

ที่ไม่ขึ้นอยู่กับภาษาใดภาษาหนึ่งโดยเฉพาะ (Cheragui, 2012) ในงานวิจัยนี้ นำระบบการแปลแบบถ่ายทอดมาใช้ในการแปลระหว่างภาษาไทย-อีสาน โดยใช้กฎการถ่ายทอดที่กำหนดไว้ (Rule-based method) และพจนานุกรม เพื่อนำมาใช้วิเคราะห์ประโยคต้นฉบับในส่วนลักษณะของคำ (Morphology) วากยสัมพันธ์ (Syntactical) และความหมาย (Semantic) แล้วนำไปสร้างเป็นภาษาเป้าหมายตามโครงสร้างภาษา

(Alawneh และ Sembok, 2011) นำเสนองานวิจัยที่ใช้กฎการถ่ายทอด (Rule-Based method) ประยุกต์ใช้ในการแปลภาษาระหว่างภาษาอังกฤษและภาษาอารบิก (หรือภาษาอาหรับ) โดยงานวิจัยนี้ออกแบบผสมการทำงานร่วมกันระหว่าง Rule-Based และ Example-Based มาประยุกต์ใช้เพื่อให้เกิดความสมดุลย์ของทั้งสองวิธีที่นำมาใช้ในการแปลข้อความจากเครื่องแปลภาษา และการจัดการของปัญหาเกี่ยวกับข้อตกลงคำและเนื้อเรื่องในการแปลประโยคจากภาษาอังกฤษเป็นภาษาอารบิก อย่างไรก็ตามจำนวนกฎที่ถูกสร้างยังมีไม่มาก ทำให้ประสิทธิภาพยังไม่เพียงพอ

(Tan et. al., 2012) นำเสนองานวิจัยใช้ระบบสำหรับการแปลภาษามลายู และระบบการสังเคราะห์ โดยระบบจะแปลงประโยคที่เทียบกับภาษามลายู และสังเคราะห์การพูดที่สอดคล้องระบบแปลงข้อความเป็นหน่วยเสียงโดยใช้กฎ ซึ่งสามารถแบ่งกฎออกเป็น 6 กลุ่ม คือ 1.กฎที่ไม่พิจารณาบริบท การแปลงข้อความโดยไม่พิจารณาบริบท 2.กฎที่ให้ความสำคัญในบริบท 3.กฎที่ให้ความสำคัญในบริบทที่ต้องการข้อมูลพยางค์ 4.กฎที่ให้ความสำคัญในบริบทที่ต้องการข้อมูลคำแสดงวิภัติปัจจัย 5.กฎที่ให้ความสำคัญในบริบทที่ต้องการสองพยางค์และข้อมูลคำแสดงวิภัติปัจจัย และ 6.กฎที่เฉพาะเจาะจงภาษาถิ่น

(ณัฐพงษ์ จุฑาทงกูร, 2553) นำเสนอการพัฒนาระบบแปลภาษาด้วยเครื่องแบบใช้กฎไวยากรณ์สำหรับการแปลภาษาไทยเป็นภาษาญี่ปุ่น และภาษาญี่ปุ่นเป็นภาษาไทยผ่านระบบอินเทอร์เน็ต มีการสร้างพจนานุกรมที่ครอบคลุมคำศัพท์และวลีภาษาไทย และการพัฒนาฟังก์ชันวิเคราะห์คุณสมบัติของคำและประโยค แต่ระบบที่สร้างขึ้นยังไม่ครอบคลุมโครงสร้างทั้งหมดของภาษาไทย

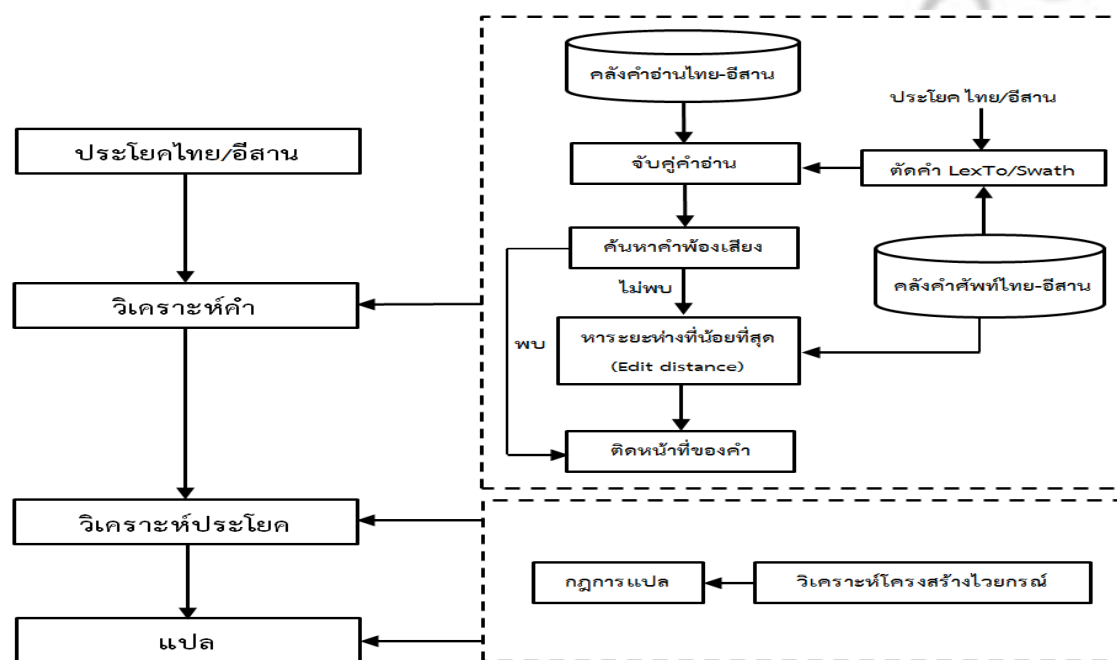
(ณัฐดนัย หอมคง, 2553) นำเสนอการพัฒนาระบบที่สามารถแปลข้อความภาษาไทยเป็นข้อความภาษามือไทยแบบอัตโนมัติโดยเลือกใช้การแปลแบบกฎไวยากรณ์ ข้อความภาษาไทย ที่จะถูกแปลจะถูกตัดคำเป็นคำศัพท์ย่อย ๆ ที่จะถูกนำไปตรวจสอบคำศัพท์ที่ยาวที่สุด มีการระบุหน้าที่ของคำศัพท์ และนำคำศัพท์ทั้งหมดที่ตรวจสอบแล้วสร้างเป็นข้อความใหม่ และนำไปตรวจสอบกับโครงสร้างไวยากรณ์ภาษาไทยที่สอดคล้องและเพียงพอสำหรับข้อความในการสื่อสารระหว่างคนปกติกับผู้พิการทางการได้ยิน หลังจากนั้นข้อความใหม่จะถูกจัดเรียงตามโครงสร้างภาษามือไทย

(Bermert, 2005) นำเสนองานวิจัยที่ใช้ระบบการแปลโดยตรง สำหรับการแปลภาษาลาวไปเป็นภาษาฝรั่งเศส และการแปลภาษาฝรั่งเศสไปเป็นภาษาลาว งานวิจัยใช้พจนานุกรมลาวและฝรั่งเศสสำหรับการแปลภาษา

4. การออกแบบและวิเคราะห์ระบบแปลภาษาไทย-อีสานโดยใช้ฐานกฎ

การทำงานโดยรวมของระบบจะเริ่มจากส่วนติดต่อกับผู้ใช้งาน ซึ่งมีหน้าที่รับความต้องการของผู้ใช้ ซึ่งก็คือ ประโยคภาษาไทย หรือ ภาษาอีสานที่จะถูกแปลไปยังภาษาเป้าหมาย และแสดงผลลัพธ์ของระบบ โดยผลลัพธ์จะมี 2 ส่วน คือ ประโยคที่แปลเป็นภาษาเป้าหมาย และคำศัพท์

ที่ใกล้เคียงกันกับคำศัพท์ในข้อความประโยคที่รับเข้าเพื่อช่วยผู้ใช้ในการแปลเมื่อเกิดกรณีการสะกดคำศัพท์ไม่ถูกต้อง ในส่วนของค้นหาคำใกล้เคียงและระบบแปลภาษาจะประมวลผลข้อความประโยคที่รับเข้าเพื่อแสดงผลให้ผู้ใช้งาน โดยการสร้างระบบแปลภาษานี้จะประกอบไปด้วยส่วนสำคัญคือการสร้างคลัง การวิเคราะห์คำ ค้นหาคำใกล้เคียง การวิเคราะห์ประโยค และการแปล ภาพที่ 1 แสดงแบบจำลองโครงสร้างการทำงานของระบบการแปลภาษา ไทย-อีสาน



ภาพที่ 1 แบบจำลองโครงสร้างการทำงานของระบบการ

4.1 คลังสำหรับโปรแกรมแปลภาษาไทย-อีสาน

คลังประโยคภาษาไทย-อีสาน ในส่วนคลังประโยคภาษาไทยจะประกอบด้วยประโยคภาษาไทยจำนวน 100 ประโยค และตัวอย่างประโยคภาษาอีสาน (สำลี รักสุทธี, 2554), (Cyberlab, 2555), (ชมรมศิลปวัฒนธรรมอีสาน, 2554) จำนวน 100 ประโยค โดยแต่ละประโยคจะตีความหน้าที่ของคำแต่ละคำในประโยค คลังนี้จะถูกใช้ในขั้นตอนการค้นหาหน้าที่ของคำ มีโครงสร้างดังนี้ คำศัพท์/หน้าที่ของคำ คำศัพท์/หน้าที่ของคำ -/NONE (เมื่อจบประโยค) และขึ้นบรรทัดใหม่เมื่อเริ่มประโยคใหม่ ซึ่งตัวอย่างคลังประโยคภาษาไทย-อีสาน ภาพที่ 2 แสดงตัวอย่างการเก็บคลังประโยคภาษาไทย-อีสาน

กลางวัน/N อาภาศ/N ดีขึ้น/ADV -/NONE	มือเร็น/N อาภาศ/N ไคแม่/ADV -/NONE
ฉัน/PRON ทำ/V งาน/N -/NONE	ข่อย/PRON เอ็ด/V เวียก/N -/NONE
ฉัน/PRON ไม่เคย/AUX เห็น/V -/NONE	ข่อย/PRON บ่เคย/AUX สบ/V -/NONE
กล้วยน้ำว้า/N หมด/V แล้ว/ADV -/NONE	ก้วยยวน/N เบ็ด/V จ้อย/ADV -/NONE
เธอ/PRON ไป/V เดิน/V ด้วย/REP น๊ะ/ADV -/NONE	เจ้า/PRON ไป/ADV ญ่าง/V นำ/REP เต๋อ/ADV -/NONE

ภาพที่ 2 ตัวอย่างคลังประโยคภาษาไทย-อีสาน

คลังคำศัพท์ภาษาไทยและอีสานที่ใช้ในงานวิจัยนี้สร้างมาจากคำศัพท์ภาษาไทยจำนวน 1,895 คำ และคลังคำศัพท์ภาษาอีสานประกอบด้วยคำศัพท์ภาษาอีสานจำนวน 1,895 คำ คลังคำศัพท์เหล่านี้จะถูกนำไปใช้ในขั้นตอนการหาคำศัพท์ใกล้เคียง ซึ่งใช้วิธีการหาค่าระยะการแก้ไขที่น้อยที่สุด (Minimal Edit Distance) โดยโครงสร้างคลังคำศัพท์ทั้งสองภาษาแต่ละบรรทัดจะประกอบไปด้วยคำศัพท์เพียงแค่บรรทัดละ 1 คำศัพท์เท่านั้น ภาพที่ 3 แสดงตัวอย่างคลังไฟล์คำศัพท์ภาษาไทย และภาพที่ 4 แสดงตัวอย่างคลังคำศัพท์ภาษาอีสานที่ใช้ในงานวิจัย

ก
กอด
แนบไว้กับอก
ตัน
รากฟัน
โคนฟันที่ต่อจากรากขึ้นมาตรงเหงือกหุ้ม
ลิ้นไก่

ภาพที่ 3 ตัวอย่างคลังไฟล์คำศัพท์ภาษาไทย

กะก
กัก
กักกัก
กกแซว
กักแซว
กกเหวี่ยง

ภาพที่ 4 ตัวอย่างคลังไฟล์คำศัพท์ภาษาอีสาน

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 แสดงภาพที่ 5

ก้อ กอด แบน-ไว้-กับ-อก ต้น ราก-ฟัน โคน-ฟัน-ที่-ต่อ-จาก-ราก-ขึ้น-มาด-จะ-เหวี่ยง-หัก ลื่น-ไถ่
กะ กก กัก กก กัก กก-เข้า กัก-เข้า

ภาพที่ 5 คลังคำคำอ่านไทย-อีสาน

พจนานุกรมภาษาไทย และภาษาอีสาน โครงสร้างการเก็บของพจนานุกรมทั้งภาษาไทย และภาษาอีสานนี้จะประกอบด้วยภาระหน้าที่ของคำ และความหมาย พจนานุกรมทั้งสองนี้จะถูกนำไปใช้ในขั้นตอนการหาความหมายของคำศัพท์ภาษาอีสานแปลไปเป็นภาษาไทย มีจำนวนภาษาไทย ทั้งหมด 1,895 คำ และภาษาอีสาน 1,895 คำ มีโครงสร้างภายในพจนานุกรม 1 บรรทัด จะมีเก็บ คำศัพท์ หน้าที่ของคำ ความหมาย ของคำศัพท์ 1 คำเท่านั้น ดังภาพที่ 6 แสดงตัวอย่างพจนานุกรม ภาษาไทย และภาพที่ 7 แสดงตัวอย่างพจนานุกรมภาษาอีสาน

อยู่,V, อยู่ อะไร,PRON, อีหยัง น้า,AUX, เป็นตา ชับ,V, ชับ ทำไม,ADV, หยั่ง คบ,V, คบ

ภาพที่ 6 ตัวอย่างพจนานุกรมภาษาไทย

กะ,CONJ, ก็ กก,V, กอด กัก,V, กอด กก,N, ต้น กัก,N, ต้น กกเข้า,V, รากฟัน

ภาพที่ 7 ตัวอย่างพจนานุกรมภาษาอีสาน

4.2 การวิเคราะห์คำ และออกแบบระบบค้นหาคำใกล้เคียง

เนื่องจากภาษาอีสานเป็นภาษาพูดท้องถิ่น ดังนั้นปัญหาของภาษาอีสานคือไม่มีพจนานุกรมทางการที่กำหนดคำภาษาอีสานไว้ ดังนั้นคำของภาษาอีสานของแต่ละคนจึงอาจสะกดไม่เหมือนกัน แต่อย่างไรก็ตามก็จะมีคำพ้องเสียงเดียวกัน ดังนั้นในงานวิจัยนี้จึงนำเอาหลักการแปลงคำเป็นคำอ่านมาเพื่อแก้ไขปัญหาคำสะกดไม่ตรงกัน ในกรณีที่ไม่สามารถหาคำพ้องเสียงในคลังคำศัพท์ได้ ระบบจะนำคำอ่านไปเข้าหลักการการหาคำระยะการแก้ไขที่น้อยที่สุด ซึ่งขั้นตอนนี้เป็น การวัดหาค่าความต่างของสายอักขระสองชุด โดยจะเลือกคำที่มีค่าความต่างน้อยที่สุดเป็นผลลัพธ์ จากนั้นนำผลลัพธ์ไปติดหน้าที่ของคำ โดยขั้นตอนนี้จะประกอบด้วยขั้นตอนต่อไปนี้

1. ขั้นตอนวิเคราะห์คำ

ในระบบการแปลภาษาไทย-อีสาน เริ่มแรกระบบจะดำเนินการตัดคำจากประโยคที่รับเข้ามา โดยประโยคภาษาไทยจะตัดคำโดยใช้โปรแกรม SWATH (Charoenpornasawat, 1999) และประโยคภาษาอีสานจะตัดคำโดยใช้โปรแกรม LexTo (สรสาร, 2554) เหตุผลที่เลือกใช้โปรแกรม LexTo ในการตัดคำภาษาอีสานเพราะสามารถปรับปรุงพจนานุกรมได้ ทำให้การตัดคำนั้นมีประสิทธิภาพ มากขึ้น เช่น ฉันทินข้าว ตัดคำ ฉันทิน|ข้าว

2. จับคู่คำอ่าน และหาค้นหาคำพ้องเสียง

เมื่อตัดคำเสร็จเรียบร้อยแล้ว จากนั้นก็จะนำคำที่ได้จากการตัดคำนำมาเปลี่ยนเป็นคำอ่าน จะได้คำอ่านของคำแต่ละคำในประโยคที่ตัดคำได้นำคำอ่านของคำศัพท์ที่รับเข้ามาไปเปรียบเทียบกับไฟล์คำอ่าน เพื่อนำมาวิเคราะห์คำพ้องเสียงเนื่องจากภาษาอีสานมักจะพบคำที่อ่านออกเสียงเหมือนกันแต่จะเขียนต่างกัน โดยใช้หลักการเปรียบเทียบคำอ่าน และใช้การจับคู่ (Matching) มาช่วยในการกรองข้อมูลเพื่อให้ได้คำศัพท์ที่มีคำอ่านตรงกันกับคำที่รับเข้ามา เช่น

ตัดคำ ฉันทิน | ไป | โรงเรียน

คำอ่าน ฉันทิน | ไป | โรง - เรียน

ในกรณีคำอ่านของคำที่รับเข้ามาไม่ตรงกับคำอ่านของไฟล์คำอ่าน จะนำคำไปเข้าหลักการการหาคำระยะการแก้ไขที่น้อยที่สุด โดยเลือกใช้วิธีการ Levenshtein edit distance (Piotrowski, 2012) เป็นขั้นตอนการวัดหาค่าความต่างของสายอักขระสองชุด โดยที่ค่าความต่างกันจะวัดจากจำนวนของการที่จะต้องทำการตัดออก แทรก และแทนที่ จนกระทั่งอักขระมีลักษณะเหมือนกันทุกประการ โดยจะเลือกคำที่มีค่าระยะการแก้ไขที่น้อยที่สุดเป็นผลลัพธ์ จากนั้นนำผลลัพธ์ไปติดหน้าที่ของคำต่อไป เช่น

ข้าว เป็น ข้าว edit distance = 1 (แทนที่ ไหมเอก ด้วย ไหมโท 1 ครั้ง)

ช้อย เป็น ช้อย edit distance = 1 (แทนที่ ไหมโท ด้วย ไหมเอก 1 ครั้ง)

3. การติดหน้าที่ของคำ

ขั้นตอนของการติดหน้าที่ของคำเริ่มต้นจากนำผลลัพธ์ที่ได้มาจากการเปรียบเทียบคำอ่านและการหาระยะห่างที่น้อยที่สุด นำคำไปเทียบกับข้อมูลในคลังประโยค โดยคำที่นำไปเทียบนั้นจะพิจารณาบริบทรอบข้าง (Bigram Tagger) ว่าถ้าตามด้วยอีกคำแล้วคำที่เราต้องการติดหน้าที่ของคำด้วยชนิดใด เช่น คำนาม คำกริยา ฯลฯ ถ้ากรณีที่เทียบโดยสนใจบริบทรอบข้างแล้วไม่พบก็จะนำคำนั้นไปเทียบในคลังประโยคอีกครั้งโดยไม่สนใจบริบทรอบข้าง (Lookup Tagger) โดยหาว่า

คำนั้นในคลังประโยคมีหน้าที่ของคำว่าอะไรมากที่สุด และถ้ากรณีที่เทียบโดยไม่สนใจบริบทรอบข้างแล้วยังไม่พบอีก ก็ติดหน้าที่ของคำเป็นค่าเริ่มต้น (default) โดยที่ค่าเริ่มต้นนี้ได้มาจากหน้าที่ของคำที่ถูกใช้มากที่สุดในการคลังประโยค เมื่อได้ผลลัพธ์เป็นคำพร้อมติดหน้าที่ของคำแล้วก็จะนำไปเข้ากฎในการแปลเพื่อหาความหมายต่อไป

4.3 การวิเคราะห์ประโยค และการแปล

เมื่อได้โครงสร้างประโยคพร้อมกับหน้าที่ของคำแล้ว ก็จะนำโครงสร้างไวยากรณ์ของประโยคที่ได้มานั้นเข้าสู่กระบวนการแปลภาษาเพื่อแปลไปยังภาษาเป้าหมาย และดึงความหมายมาจากพจนานุกรม โดยอาศัยกฎไวยากรณ์ (Grammars) ของภาษา และการตัดคำ เพื่อตรวจสอบความถูกต้องของโครงสร้างไวยากรณ์ในประโยค โดยมีขั้นตอนดังนี้

1. การวิเคราะห์ไวยากรณ์ และความหมาย

การวิเคราะห์ไวยากรณ์ (Danthanavanich, 2008) คือ และความสัมพันธ์ระหว่างคำในประโยคเพื่อนำไปประมวลผลหาผลลัพธ์มาแสดงให้กับผู้ใช้ ซึ่งงานวิจัยนี้ใช้เทคนิคการเครื่องแปลภาษาโดยใช้ฐานกฎ (Rule-Based Machine Translation) (Dangsaart et. al., 2007) ตามหลักภาษาไทยในการเทียบโครงสร้างไวยากรณ์ เพื่อนำไปหาความหมาย 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 คำนามตามด้วยกริยาช่วย

(NP -> N + AUX)

กฎข้อที่ 5 ส่วนประธาน ประกอบด้วย คำนามตามด้วยคำนาม

(NP -> N + N)

กฎข้อที่ 6 ส่วนประธาน ประกอบด้วย คำนาม

(NP -> N)

กฎข้อที่ 7 ส่วนประธาน ประกอบด้วย คำกริยาตามด้วยคำบุพบท

(NP -> V + PREP)

กฎไวยากรณ์ที่ได้จะถูกนำมาใช้ตรวจสอบความถูกต้องของไวยากรณ์ในประโยคเมื่อประโยคที่รับเข้ามานั้นผ่านการการตัดคำมาแล้ว เช่น “ฉันกินข้าว” ตัดคำ ฉัน|กิน|ข้าว

ติดหน้าที่ของคำ N | V | N

S

จากภาษาไทยเป็นภาษาอีสาน ซึ่งส่วนหนึ่งเนื่องมาจากระบบได้มีการออกแบบเพื่อรองรับการสะกดผิดเนื่องจากภาษาอีสานเป็นภาษาพูดจึงยากต่อการสะกดให้ถูกต้องตามพจนานุกรมได้ จึงทำให้ไม่สามารถแปลคำศัพท์นั้นได้ถูกต้อง ส่วนการแปลภาษาไทย-อีสาน ความผิดพลาดส่วนใหญ่เกิดจากขั้นตอนการตัดคำภาษาไทยอัตโนมัติยังคงให้ผลลัพธ์ที่ไม่ถูกต้อง เป็นผลให้ขั้นตอนที่เป็นส่วนของการแปลไม่ถูกต้อง

7. เอกสารอ้างอิง

- ณัฐพงษ์ จุฑาทู. (2553). การแปลภาษาไทยเป็นภาษาญี่ปุ่นแบบใช้ฐานกฎด้วยเครื่องคอมพิวเตอร์. วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์. มหาวิทยาลัยเชียงใหม่.
- ณัฐดนัย หอมคง. (2553). การแปลภาษาด้วยเครื่องสำหรับข้อความภาษาไทยเป็นข้อความภาษามือไทย. วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์. มหาวิทยาลัยเชียงใหม่.
- สำลี รักสุทธี. (2554). พจนานุกรม ภาษาอีสาน – ไทยกลาง. กรุงเทพฯ. สำนักพิมพ์พัฒนาศึกษา. ชมรมศิลปวัฒนธรรมอีสานจุฬาลงกรณ์มหาวิทยาลัย. (2554). คลังประโยค, [ระบบออนไลน์], แหล่งที่มา http://www.isan.clubs.chula.ac.th/lang/index.php?transaction=search_word.php, เข้าดูเมื่อวันที่ 1/12/2557.
- สรสรสาร. (2554). Thai Lexeme Tokenizer (LexTo), [Online] Available HTTP:<http://www.sansarn.com/lexto/download-lexto.php>, on 1/12/2557.
- Alawneh, M.F. and Sembok, T.M. (2011). Rule-Based and Example-Based Machine Translation from English to Arabic, paper presented in Bio-Inspired Computing: Theories and Applications (BIC-TA) 2011, Penang. Malaysia.
- Berment, V. (2005). Online Translation Services for the Lao Language. paper presented in The First International Conference on Lao Studies, USA.
- Charoenpornasawat, P. (1999). Feature-based Thai Word Segmentation. Master's Thesis. Computer Engineering, Chulalongkorn University, Bangkok, Thailand.
- Cheragui, M.A. (2012). Theoretical Overview of Machine translation, paper presented in the 4th International Conference on Web and Information Technologies 2012, Sidi Bel Abbès, Algeria.
- Cyberlab. (2555). ภาษาถิ่นไทย. [ระบบออนไลน์], แหล่งที่มา <http://cyberlab.lh1.ku.ac.th/elearn/faculty/human/hm19/lesson1.htm>, เข้าดูเมื่อวันที่ 1 ธันวาคม 2557.
- Dangsaart, S., Naruedomkul, K., Cerccone, N., Sirinaovakul, B. (2007). Bridging the Gap: Thai- Thai Sign Machine Translation, paper presented in The 10th Conference of the Pacific Association for Computational Linguistics, University of Melbourne, Australia.

- Danthanavanich, S. (2008). **A Grammer of Thai Sign Language**. Ph.D. Dissertation, Mahidol University, Bangkok, Thailand.
- Piotrowski, M. (2012). **Natural Language Processing for Historical Texts**, Morgan & Claypool Publishers.
- Tan, T.P., Goh, S.S. and Khaw, Y.M. (2012). A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System, **paper presented in Asian Language Processing (IALP) 2012**, Hanoi, Vietnam.

Industrial Technology
มหาวิทยาลัยราชภัฏลำปาง