การสกัดคำจากข้อความโดยใช้เครื่องมือประมวลภาษาธรรมชาติ
Word Exception using Natural Language Toolkit

บทคัดย่อ

บทความนี้มีวัตถุประสงค์เพื่อศึกษาเกี่ยวกับกระบวนการการตัดคำหรือประโยค เพื่อหาความหมายที่แท้จริงของคำหรือประโยคนั้น ๆ ลดความกำกวมของภาษาหรือความกำกวมของประโยค และเพิ่มประสิทธิภาพและความแม่นยำให้กับการแปลความหมาย โดยศึกษาวิธีการและกระบวนการทำงานของ Google Translate และ โปรแกรมแปลด้วยภาษิต เพื่อเปรียบเทียบประสิทธิภาพของการแปล และผลลัพธ์ของการแปลจากโปรแกรมทั้งสองโปรแกรมว่ามีความถูกต้องหรือไม่ โดยในบทความนี้ได้นำเครื่องมือในการประมวลผลภาษาทางธรรมชาติเข้ามาช่วยในกระบวนการ การตัดคำ เพิ่มประสิทธิภาพให้กับโปรแกรมแปลภาษาในส่วนของการแปลความหมายให้กับคำหรือประโยค เพื่อให้ได้ผลลัพธ์ที่ดียิ่งขึ้น ถูกต้อง และแม่นยำมากยิ่งขึ้น ทำให้ผู้ใช้สามารถเข้าใจความหมายของคำหรือประโยคได้ง่ายและถูกต้องมากขึ้น

คำสำคัญ: การตัดคำ, การประมวลผลภาษาธรรมปาติ , แอพพลิเคชั่น

ABSTRACT

This article is intended to study the process of word or sentence trimming. To find the true meaning of the word or phrase, reduce the ambiguity of the language or the ambiguity of the sentence. And increase the efficiency and accuracy of interpreting. By studying the methods and processes of Google Translate and the translation program with proverbs. To compare translation performance And the results of the translations from both programs are accurate or not. This article introduces the natural language processing tools to assist in the word wrapping process, increasing the efficiency of translation programs in the interpretation of words or sentences. For better, more accurate, and more accurate results. This allows users to understand the meaning of words or sentences more easily and accurately.

Keyword— Wrapping words, Natural Language Processing, Application

1. บทนำ

เทคโนโลยีในการสื่อสารได้มีการพัฒนาไปอย่างรวดเร็ว ซึ่งในปัจจุบันมีเทคโนโลยีเพื่ออำนวยความสะดวกในการเข้าถึงสื่ออย่างหลากหลาย ผู้บริโภคสามารถเข้าถึงได้ด้วยอุปกรณ์อิเล็กทรอนิกส์ เช่น คอมพิวเตอร์ แท็บเล็ตหรือสมาร์ตโฟน เพื่อใช้รับข่าวสารต่าง ๆ จากทั้งในและต่างประเทศ โดยส่วนใหญ่เป็นภาษาอังกฤษ ซึ่งปัญหาของคือ ความรู้ทางด้านภาษาของผู้บริโภคบางกลุ่มที่อาจไม่มีความรู้ในด้านภาษา จึงส่งผลให้การบริโภคสื่อไม่อ่านและเข้าใจถึงข่าวสารเหล่านั้นได้ จึงได้มีการนำเทคโนโลยีการแปลภาษา เข้ามาช่วยเพื่อให้สามารถรับรู้ถึงสารเหล่านั้นได้

แต่การแปลภาษานั้นยังมีปัจจัยอยู่หลายอย่างซึ่งทำให้การแปลมีความหมายที่ผิดเพี้ยนไป เช่น ส่วนขยายของประโยค คำเชื่อมประโยค หรือ คำเชื่อมต่าง ๆ ถ้าหากสามารถลดปัจจัยเหล่านี้ลดลงได้ การแปลภาษาอาจมีประสิทธิภาพมากขึ้น

ดังนั้นหากเรานำการตัดคำมาใช้ช่วยในการแปลภาษาแล้ว จะสามารถลดความกำกวมของข้อความหรือประโยคนั้น ๆ ได้เพื่อให้ความหมายของการแปลภาษามีประสิทธิภาพมากขึ้น

2. งานวิจัยและทฤษฎีที่เกี่ยวข้อง

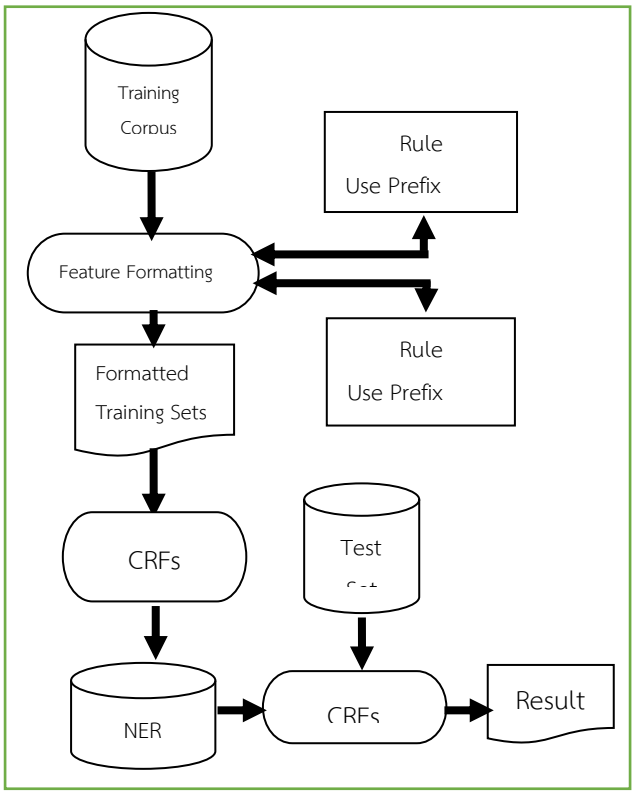
2.1 การศึกษาเปรียบเทียบการใช้และการไม่ใช้คำนำหน้าชื่อคำต่อท้ายในการจดจำนิพจน์ระบุนาม

การจดจำนิพจน์ระบุนามใน [1] ภาษาไทย หมายถึง สิ่งที่ใช้เรียกชื่อบุคคลชื่อองค์กรหรือชื่อสถานที่ แต่ปัญหาของการจดจำนิพจน์ระบุนามในบทความภาษาไทยพบว่าคือการเขียนติดกันทั้งประโยคและความไม่มีขอบเขตที่แน่นอน ไม่มีตัวอักษรพิมพ์ใหญ่กับพิมพ์ตัวเล็กในการใช้นิพจน์ระบุนามเหมือนกับภาษาอังกฤษ และสามารถเกิดขึ้นใหม่ได้ตลอดเวลา เช่น เกิดขึ้นจากคำเดี่ยว เกิดขึ้นจากการนำคำมาผสมกัน หรือเกิดขึ้นจากการนำภาษาต่างประเทศมาใช้ ซึ่งก็เป็นนิพจน์ระบุนามได้แล้ว ปัญหาเหล่านี้ที่กล่าวมาทำให้ยากในการจดจำนิพจน์ระบุนามในภาษาไทย

การจดจำนิพจน์ระบุนาม คือ การค้นหาและสกัดคำที่เป็นนิพจน์ระบุนาม จากข้อความโดยทั่วไปนิพจน์ระบุนามสามารถแบ่งออกได้เป็นหมวดหมู่หลัก ได้แก่ ชื่อบุคคล เช่น “อภิสิทธิ์ เวชชาชีวะ” “ไทเกอร์ วูดส์” ชื่อองค์กร เช่น “ธนาคารกรุงไทย” “ศูนย์ข้อมูลคนหาย” ชื่อสถานที่ เช่น “เชียงใหม่” “สยามเซ็นเตอร์”นิพจน์ระบุนามเป็นหน่วยหนึ่งซึ่งมีความสำคัญต่อการประมวลผลทางภาษาธรรมชาติที่นำไปใช้ในการพัฒนางานด้านต่าง ๆ เช่น การตัดคำ การสืบค้นข้อมูล หรือการกรองข้อมูล การย่อความ เป็นต้น จากปัญหาที่กล่าวมาว่าของนิพจน์ระบุนาม

การเขียนติดกันทั้งประโยคและไม่มีขอบเขตที่แน่นอน คือปัญหาของภาษาไทย เช่นข้อความต่อไปนี้ “นที คงสุข ผู้สื่อข่าวกีฬาไทยรัฐรายงานจากกรุงโตเกียวประเทศญี่ปุ่นถึงความเคลื่อนไหวของขุนพลนักเตะทีมชาติไทยชุดใหญ่ที่มีโปรแกรมจะลงฟาดแข้งศึกฟุตบอลรอบคัดเลือกโซนเอเชียรอบ 3 กับทีมชาติญี่ปุ่นในเย็นวันนี้” จากข้อความจะพบนิพจน์ระบุนามดังนี้ นที คงสุข ,ไทยรัฐ, กรุงโตเกียว ประเทศญี่ปุ่น,ทีมชาติไทย, เอเชีย,ทีมชาติญี่ปุ่น เป็นต้น

โมเดลมีกระบวนการทำงานอยู่ 3 กระบวนคือ การคลังข้อมูล (Corpus) เครื่องมือที่ใช้ (Tool) การสร้างกฎใน การเรียนรู้ และ ทำการเรียนรู้ระบบและทำการทดสอบระบบซึ่งจะแสดงกระบวนการทดลอง



ภาพที่ 1: โมเดลกระบวนการทำงาน

การสร้างโมเดลการคำนำหน้าชื่อและคำลงท้าย (Rule used Prefix Suffix) และจะมีคำนำหน้านิพจน์ระบุ นามดังนี้บุคคลมีประมาณ 194 คำ สถานที่และองค์กรรวมกันมี ประมาณ 15 คำ ส่วนคำ ที่ปรากฏหลัง นิพจน์ระบุนามมีประมาณ 7 คำ โดยรายละเอียดตามตัวอย่างดังตารางที่ 1

ตารางที่ 1 คำนำหน้าชื่อและคำลงท้าย (Rule used Prefix Suffix)

คำ	ประเภท	TAG	LABLE
นาย , นาง , ดร. , คณะ , กรุง , ประเทศ ฯลฯ	**เป็นคำอยู่หน้าชื่อ**	**P**	**B-NE**
แห่งชาติ,โพลล์,จำกัด (มหาชน),ฯลฯ	**เป็นคำอยู่ต่อท้าย**	**S**	**I-NE**
กิน,นอน,เที่ยว,ไป , ฯลฯ	**เป็นคำทั่วไป**	**O**	**I**

การทดสอบประสิทธิภาพจะมีขนาด 50,000 คำจากบทความข่าวเพื่อวัดประสิทธิภาพการของโมเดลทั้งสองโมเดลด้วยวิธีการวัดค่า F-measure ซึ่งจะหาได้จากสูตรในการหาค่า F1 คือ หาค่าความแม่นยา (Precision) และค่าหาความครบถ้วน (Recall) ดังสมการ (1)

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \quad (1)$$

2.2 ไวยากรณ์และการตีความประโยค

การสื่อสารทางภาษาเป็นกระบวนการโดยคนสองคนหรือมากกว่า บุคคลทั่วไปรับและส่งข้อความที่อยู่ในรูปแบบของภาษามนุษย์ โครงสร้างของภาษาเหล่านี้เป็นประโยค ทฤษฎีของการตีความประโยคพยายามที่จะอธิบายทักษะการฟังเป็นการตีความของคำพูดที่หลากหลายที่พบในภาษาของบุคคลทั่วไปในชีวิตประจำวัน

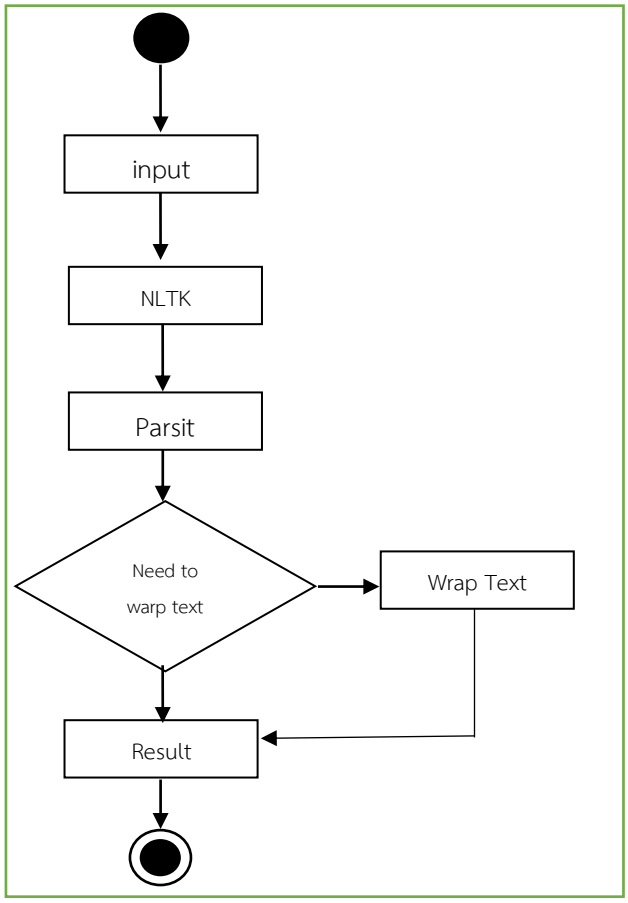
ธรรมชาติของความเข้าใจ ปัญหาของการพยายามที่จะพัฒนาทฤษฎีของการตีความประโยคที่มีระดับของความเข้าใจหรือ ไม่เข้าใจ ความล้มเหลวที่จะเข้าใจคำสั่งในภาษาอังกฤษที่เกี่ยวกับบางหลักการของกลศาสตร์ ความเข้าใจเป็นปลายเปิดมิติที่ปลายด้านหนึ่งมันจะติดแน่นอยู่กับข้อมูลทางภาษาและทักษะเฉพาะ แต่มันแผ่ออกไปอย่างรวดเร็วในเว็บที่ซับซ้อนของการไม่ทราบความจริง

เพื่อมุ่งเน้นถึงบทบาทของไวยากรณ์ในกระบวนการแปลความหมาย ดังนั้นจึงมีความจำเป็นต้องแยกความหมายรวมประโยคบนพื้นฐานของรูปแบบทางภาษานั้นและการตีความที่ถูกให้แก่ประโยค ที่สำคัญการตีความประโยคในโอกาสใดก็ตามอาจแตกต่างกันตามความหลากหลายของตัวแปรภาษาจึงเป็นเรื่อง

ธรรมดาที่จะแยกแยะความแตกต่างระหว่างความหมายที่เกี่ยวข้องกับประโยคบนพื้นฐาน รูปแบบของภาษาและการตีความที่มีกับประโยคนั้น ๆ เพื่อเสริมสร้างความแตกต่างนี้จะเป็นประโยชน์ที่จะบอกว่าประโยคในภาษาที่มีความหมายเป็นตัวแทนที่ได้รับมอบหมายโดยไวยากรณ์และการตีความความหมายโดยผู้ฟัง.[2]

3. โครงสร้างระบบ

จากรูปภาพจะมีกระบวนการทำงานดังนี้



รูปที่ 2. โครงสร้างระบบ

อินพุท (Input) จะเป็นตัวกรอกข้อความหรือคำที่ต้องการจะแปล โดยรับเข้ามาเป็น String เท่านั้น เครื่องมือประมวลผลภาษาธรรมชาติ (NLTK : Natural Language Toolkit) เป็นเครื่องมือที่ใช้ในการแบ่ง String ที่เป็นข้อความออกเพื่อให้ทราบถึงการสิ้นสุดของประโยคนั้น ๆ โปรแกรมแปลด้วยภาษิต (Parsit) เป็นกระกวนการของการแปลข้อความหรือคำ อินพุท (หากต้องการที่จะแปลอย่างเดียวก็จะแสดงผลลัพธ์นั้น ๆ การตัดคำ (Wrap Text) วิธีการตัดคำได้ใช้วิธีการของ [1] เพื่อช่วยในการทำให้ข้อความนั้น ๆ ลดความกำกวมลง ผลลัพธ์ (Result) ผลลัพธ์ที่ได้จากกระกวนการทำงานทั้งหมด

อกรรมกริยา	สกรรมกริยา

รูปที่ 3. ส่วนคำที่ตัด

อกรรมคือประโยคที่ไม่มีกรรมแล้วเป็นประโยคอ่านแล้วเข้าใจหรืออาจกล่าวได้ว่าเป็นประโยคที่สมบูรณ์โดยไม่มีกรรมมาต่อท้าย

สกรรมคือประโยคที่มีกรรมมาต่อท้ายจึงจะเป็นประโยคที่สมบูรณ์

2ส่วนนี้เป็นส่วนของกระกวนการที่ใช้ในการตัดคำของประโยคเพื่อลดความกำกวม

4.ผลการดำเนินงาน

ผลของการทดสอบการแปลระหว่าง โปรแกรมแปลด้วยภาษิต และ Google Translate.

การทดสอบประสิทธิภาพจะวัดจากวิธีการแปลคำจากประโยคหรือ ข้อความ เพื่อวัดประสิทธิภาพของโปรแกรมทั้งสอง ด้วยวิธีการเทียบกับหลักไวยากรณ์ในภาษาอังกฤษ คือ ประธาน (Subject) + กริยา (Verb) + กรรม (Object) โดยประโยคที่ใช้มีดังในตารางที่ 2

ตารางการที่ 2 ตัวอย่างประโยคที่ใช้ในการแปล

ประโยคที่ใช้ในการแปล
May I talk to Ms. Brown?
I want Melissa to go there with me instead of Jennifer.
She writes her friend a letter.
My father bought me a car.
Your problem is similar to mine.
A man sit on the table.

ตารางที่ 3 ผลลัพธ์ที่ได้จากการแปล

โปรแกรมแปลด้วยภาษิต	Google Translate
ฉันคุยกับคุณบราวน์ได้ไหม?	ผมอาจจะพูดคุยกับนางสาวสีน้ำตาล?

ฉันอยากให้เมลิสซาไปที่ตรงนั้นกับฉันแทนที่เจนนิเฟอร์	เมลิสสาฉันต้องการไปที่นั่นกับฉันแทนเจนนิเฟอร์
เธอเขียนจดหมายถึงเพื่อนของเธอ	เธอเขียนเพื่อนของเธอเป็นตัวอักษร
พ่อของฉันซื้อรถให้ฉัน.	พ่อของฉันซื้อฉันรถ
ปัญหาของคุณเหมือนกับว่าของฉัน.	ปัญหาของคุณคือคล้ายกับระเบิด
นั่งผู้ชายที่โต๊ะ.	มีชายคนหนึ่งนั่งอยู่บนโต๊ะ

โดยจากตัวอย่างในตารางที่ 3 นั้นจะยกตัวอย่างเพื่อเห็นข้อแตกต่างระหว่างโปรแกรมทั้งสองในประโยคที่มีความกำกวมและประโยคที่ไม่สมบูรณ์

ตัวอย่างที่ 1 **I want Melissa to go there with me instead of Jennifer.**

ซึ่งประโยคข้างต้นสามารถแปลความหมายได้ว่า **ฉันต้องการให้ Melissa ไปกับฉันแทนที่ Melissa จะไปกับ Jennifer** หรือ **ฉันต้องการให้ Melissa ไปกับฉัน ไม่ใช่ Jennifer ไปกับฉัน** โดยประโยคข้างต้นนี้มีความหมายมากกว่าหนึ่งความหมาย หรือประโยคกำกวม โดยผลการแปลมีดังนี้

โปรแกรมแปลด้วยภาษิต : ฉันอยากให้เมลิสซาไปที่ตรงนั้นกับฉันแทนที่เจนนิเฟอร์

Google Translate.: เมลิสสาฉันต้องการไปที่นั่นกับฉันแทนเจนนิเฟอร์

โปรแกรมแปลด้วยภาษิตนั้นสามารถแปลความหมายออกมาได้เข้าใจมากกว่าทาง Google Translate.

ตัวอย่างที่ 2 **A man sit on the table.**

ตามหลักไวยากรณ์ในภาษาอังกฤษที่ถูกต้อง ประโยคนี้ถือว่าเป็นประโยคที่ผิดหลักไวยากรณ์เพราะว่า ตามหลักของภาษาอังกฤษนั้น ประธานที่เป็นเอกพจน์นั้นกริยาจะต้องเติม s แต่ในประโยคข้างต้นนั้นไม่มีการเติม s จึงมีผลการแปลเป็นดังนี้

โปรแกรมแปลด้วยภาษิต : นั่งผู้ชายที่โต๊ะ.

Google Translate: มีชายคนหนึ่งนั่งอยู่บนโต๊ะ

ตัวอย่างดังกล่าวโปรแกรมกูเกิ้ลนั้นสามารถแปลออกมาได้ใจความสำคัญมากกว่าโปรแกรมแปลด้วยภาษิตถึงแม้จะผิดหลักไวยากรณ์ก็ตามแต่ก็สามารถแปลออกมาได้

5. สรุป

ข้อแตกต่างระหว่าง Google Translate และโปรแกรมแปลด้วยภาษิต Google Translate [3,4] นั้นเกิดจากการประมวลผลของคอมพิวเตอร์หลาย ๆ เครื่อง โดยคอมพิวเตอร์เหล่านี้จะใช้กระบวนการแปลที่เรียกว่า ระบบการแปลภาษาเชิงสถิติ (Statistical Machine Translation) ซึ่งเป็นวิธีการที่ให้คอมพิวเตอร์สร้างหลักการแปลด้วยการวิเคราะห์จากเอกสารที่ถูกแปลเสร็จแล้วเพื่อวิเคราะห์และหาหลักการแปลที่เป็นวิธีการของตนเอง และเมื่อมีการใช้งานโปรแกรมเกิดขึ้น กูเกิ้ลจะนำเอาข้อความของผู้ใช้ไปเปรียบเทียบกับคำแปลที่ได้ทดสอบ และตัดสินใจด้วยคอมพิวเตอร์เพื่อนำเอาความหมายที่ใกล้เคียงกับข้อความนั้น ๆ ออกมา จึงเป็นเหตุผลที่ Google Translate นั้นได้ผลลัพธ์ที่ไม่ค่อยน่าพอใจนัก แต่โปรแกรมแปลด้วยภาษิตนั้นสร้างขึ้นจากนักภาษาศาสตร์ โดยจะมีการอิงตามหลักไวยากรณ์ซึ่งเป็นข้อดีอย่างหนึ่ง เพราะจะให้ผลลัพธ์ที่สามารถเข้าใจได้ง่าย และการสร้างจากนักภาษาศาสตร์ จึงทำให้ข้อความหรือถ้อยคำมีภาษาที่สละสลวยมากกว่า

อย่างไรก็ตามไม่ว่าจะเป็น โปรแกรมด้วยภาษิตเอง หรือ Google Translate ผลลัพธ์ที่ได้จากการแปลก็ขึ้นอยู่กับผู้ใช้ในส่วนหนึ่งซึ่งเป็นคนตัดสินใจและพิจารณาว่าความหมายที่ได้จากการแปลนั้นมีความเหมาะกับการนำไปใช้งานมากน้อยเพียงใด

เอกสารอ้างอิง

[1] นายสายัณห์ เทพแดง. “การศึกษาเปรียบเทียบการใช้และการไม่ใช้คำนำหน้าชื่อคำต่อท้ายในการจดจำนิพจน์ระบุนาม” การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ (National Conference on Information Technology: NCIT) ครั้ งที่ 7, 2559 หน้า 358-362

[2] Limber, J. Syntax and sentence interpretation. In R. J. Wales & E. Walker (Eds.), New Approaches to language mechanisms. Amsterdam: North Holland, 1976(a).

[3] Carlos Alberto Gómez Grajales, The statistics behind Google Translate แหล่งที่มา : http://www.statisticsviews.com/details/feature/8065581/The-statistics-behind-Google-Translate.html http://www.hutchinsweb.me.uk/Nutshell-2005.pdf

[4] How Google Translate Works – The Independent (September, 2011) http://www.independent.co.uk/life-style/gadgets-and-tech/features/how-google-translate-works-2353594.html