บทที่ 4
การประเมินค่าความถูกต้องของผลการแปล

ในบทนี้จะกล่าวถึงขั้นตอนการประเมินค่าการแปล เพื่อประเมินค่าผลการแปลจากระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรมจากที่ได้อธิบายถึงอัลกอริทึมการทำงานในบทก่อนหน้านี้ และอธิบายถึงขั้นตอนการเตรียมข้อมูลเพื่อนำไปประเมินค่าผลการแปลของระบบ ทั้งนี้ผลการแปลที่เคยพบมาก่อนจะสามารถแปลได้อย่างถูกต้องแม่นยำโดยระบบนี้ ดังนั้นการประเมินค่าผลการแปลของระบบจะประเมินเฉพาะผลการแปลจากประโยคที่ไม่เคยพบมาก่อนเท่านั้น

4.1 การประเมินค่าความถูกต้องของผลการแปลสำหรับระบบแปลภาษาด้วยเครื่อง (Machine Translation Evaluation)

การประเมินค่าผลการแปลของระบบแปลภาษาด้วยเครื่อง (ขอเรียกโดยย่อว่า “การประเมินค่า”) ทำได้ยาก เนื่องจากสามารถประเมินค่าได้จากหลายมุมมอง เช่น

- การประเมินค่าเชิงไวยากรณ์ (Grammatical Evaluation)
- การประเมินค่าเชิงคุณภาพการแปลอย่างเพียงพอ (Adequacy Evaluation)
- การประเมินค่าเชิงคุณภาพการแปล (Quality Assessment)

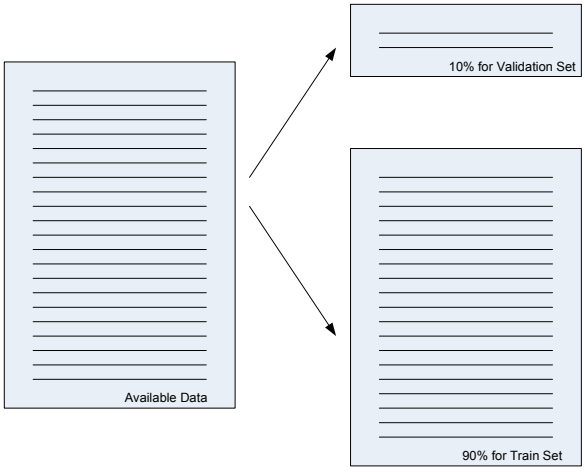
การประเมินค่าจากแต่ละมุมมองย่อมมีข้อโต้แย้งกันอีกมาก เนื่องจากการประเมินค่ามักขึ้นกับประสบการณ์และทักษะของผู้ประเมินค่าซึ่งเป็นความสามารถส่วนบุคคล ทำให้ผลการประเมินค่ามักจะมีความเอนเอียง (bias) เพื่อแก้ปัญหานี้จึงมีวิธีการประเมินค่าแบบอัตโนมัติ (Automatic Metrics for MT Evaluation) เพื่อใช้ประเมินค่าที่เป็นกลางและไม่ยึดติดกับความสามารถของบุคคล อย่างไรก็ตาม จุดด้อยของการประเมินค่าแบบอัตโนมัติคือไม่สามารถประเมินค่าความถูกต้องที่สูงกว่าระดับผิว (Surface Form) ได้ ทั้งนี้ยังไม่มีการประเมินค่าแบบใดดีที่สุด

ในงานวิจัยชิ้นนี้ได้เลือกใช้วิธีการประเมินค่าแบบอัตโนมัติของ BLEU-4 [16] ซึ่งเป็นวิธีที่ใช้กันเป็นที่แพร่หลายในการประเมินค่าแบบอัตโนมัติสำหรับระบบแปลภาษาด้วยเครื่อง

4.2 วิธีการทดสอบความถูกต้องของการเรียนรู้

4.2.1 การตรวจสอบความสมเหตุสมผลแบบไขว้ (Cross-Validation)

สำหรับวิทยานิพนธ์นี้เป็นการทดสอบเกี่ยวกับการเรียนรู้ของเครื่อง (Machine Learning) ด้วยวิธีซูเปอร์ไวส์เลินนิ่ง (Supervised Learning) ซึ่งเป็นการเรียนรู้ที่จะต้องสอนคำตอบที่ถูกต้องให้ ดังนั้นจึงต้องเตรียมข้อมูลตัวอย่างพร้อมคำตอบที่ถูกต้อง



ภาพที่ 4.1 แสดงการเตรียมชุดข้อมูลเพื่อสอนและทดสอบ

วิธีการหนึ่งที่เป็นที่นิยมใช้กันมากในการทดสอบหาค่าความถูกต้อง ซึ่งมีพื้นฐานทางด้านสถิติก็คือการทำ Cross-Validation โดยนำข้อมูลทั้งหมด (Available Data) นำมาแบ่งออกเป็นส่วนๆ ส่วนแรกเพื่อใช้ในการทดสอบ (Validation Set) และส่วนที่สองใช้ในการสอน (Train Set) วิทยานิพนธ์นี้ได้ใช้เทคนิค Cross-Validation เพื่อหาค่าความถูกต้องของแต่ละอัลกอริทึม โดยการแบ่งข้อมูลออกเป็น 10 ส่วน เรียกว่า 10-fold Cross Validation โดยมีการทำงานคือ ให้ข้อมูลแต่ละส่วนเขียนแทนด้วย Data1, Data2, ..., Data10 ในครั้งแรกให้ Data2 ถึง Data10 เป็นชุดข้อมูลที่ใช้เรียนรู้ และให้ Data1 เป็นชุดข้อมูลทดสอบ ครั้งต่อไปให้ Data1, Data3 ถึง Data10 เป็นข้อมูลสอนและให้ Data2 เป็นชุดข้อมูลทดสอบ ทำอย่างนี้จนครบ 10 ครั้งแล้วนำค่าความถูกต้องของการเรียนรู้ทั้ง 10 ครั้งมาหาค่าเฉลี่ย

4.2.2 การประเมินค่าแบบอัตโนมัติของ BLEU-4

การประเมินค่าอัตโนมัติของ BLUE-4 ตั้งอยู่บนสมมติฐานที่ว่าหากผลลัพธ์การแปลมีส่วนของข้อความที่เกิดร่วมกันกับส่วนของข้อความในผลลัพธ์การแปลอ้างอิงเป็นจำนวนที่มากกว่าก็น่าเชื่อได้ว่าผลลัพธ์การแปลนั้นจะมีความถูกต้องสูง ดังนั้นการประเมินค่าอัตโนมัติของ BLEU –*n* (เรียกโดยย่อว่า BLEU) คือการคำนวณหาค่าเฉลี่ยเลขคณิตของสัดส่วนของคำที่ซ้อนทับ

กัน (overlapping) ของชุดคำตอบจำนวน n คำติดกันที่ปรากฏในชุดคำตอบอ้างอิง โดย BLEU -4 หมายถึง n มีค่าเท่ากับ 4 (คำนวณหาค่าเฉลี่ยเลขคณิตของสัดส่วนการซ้อนทับกันของคำที่ติดกัน 4 คำในชุดคำตอบกับชุดคำตอบอ้างอิง) การคำนวณค่า BLEU ถูกคำนวณตามสมการต่อไปนี้

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4.1}$$

โดย

BP (Brevity Penalty) คือ ค่าถ่วงน้ำหนักของชุดคำตอบที่สั้นกว่าชุดคำตอบอ้างอิง โดยคำนวณจากสมการต่อไปนี้

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases} \tag{4.2}$$

โดย c คือจำนวนคำของชุดคำตอบ

r คือจำนวนคำของชุดคำตอบอ้างอิง

N จำนวนคำที่เรียงติดกัน

หากต้องการคำนวณค่า BLUE-4 จะทำให้ $N = 4$

w_n คือ ค่าถ่วงน้ำน้ำหนักของคะแนนความแม่นยำ โดย $\sum_{n=1}^{N} w_n = 1$

หากใช้การถ่วงน้ำหนักแบบ Uniform Weight จะทำให้ $w_n = \frac{1}{N}$

p_n คือ คะแนนความแม่นยำ (Precision Score) โดยคำนวณจากสมการต่อไปนี้

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}\left(n-gram\right)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count\left(n-gram\right)} \tag{4.3}$$

อย่างไรก็ดี ค่า BLEU ที่ได้มักมีค่าที่ค่อนข้างน้อยมากๆ ดังนั้นจึงต้องเปลี่ยนเป็นหน่วย log เพื่อให้เปรียบเทียบคะแนนได้ง่ายขึ้น

$$\log BLEU = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n \tag{4.4}$$

4.3 การเตรียมข้อมูลสำหรับการประเมินค่าการแปล

การทดลองนี้ใช้คลังข้อความแบบคู่ของตัวอย่างประโยคจากหนังสือ English by Examples จำนวน 100,804 ประโยค (ขอเรียกโดยย่อว่า “คลังข้อความแบบคู่”) จะถูกใช้เป็นข้อมูลตั้งต้นเพื่อนำไปใช้ทดสอบความถูกต้องของการเรียนรู้ด้วยวิธีการ 10-fold Cross-Validation

การทดลองจะนำแต่ละส่วนที่แบ่งไว้ มาแบ่งเป็นชุดสอนและชุดทดสอบ โดยมีอัตราส่วนชุดสอนต่อชุดทดสอบเป็น 9 ต่อ 1 จากนั้นนำข้อมูลชุดสอนไปเรียนรู้และนำข้อมูลชุดทดสอบมาแบ่งออกเป็นเฉพาะภาษาต้นทางและปลายทาง โดยประโยคต้นทางที่ได้จากชุดทดสอบเรียกว่าชุดต้นทาง (Source Set) ประโยคปลายทางจากชุดทดสอบจะถูกใช้เป็นผลการแปลอ้างอิงหรือชุดคำตอบอ้างอิง (Reference Result Set) จากนั้นนำชุดต้นทางไปแปลให้เป็นภาษาปลายทางด้วยระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม (jEBMT) ผลลัพธ์ที่แปลออกมาได้เรียกว่าชุดคำตอบ (Result Set) และทำเช่นเดียวกันนี้กับระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่อง ”สุภาษิต” (Parsit)

ขั้นตอนสุดท้ายจะนำ ชุดต้นทาง ชุดคำตอบอ้างอิง และชุดคำตอบจากระบบทั้งสอง ไปใช้ทำการทดลองตามกรณีต่างๆ ต่อไป

4.4 ผลการทดลอง

การทดลองสามารถทำได้โดยประเมินค่าความถูกต้องของการแปล โดยจะแบ่งการประเมินออกเป็น 2 กลุ่ม โดยกลุ่มแรกจะประเมินค่าความถูกต้องของการแปลแบบอัตโนมัติ ซึ่งจะถูกแบ่งเป็น การประเมินค่าแบบอิงจำนวนการเข้าคู่แบบแม่นตรง (Exact Matching) และการประเมินค่าอัตโนมัติของ BLUE-4 กลุ่มที่สองจะประเมินค่าความถูกต้องของการแปลโดยมนุษย์ประกอบด้วย ความถูกต้องของการเรียงลำดับคำ ความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท และความถูกต้องของการแปลวลี

4.4.1 ประเมินค่าความถูกต้องของการแปลแบบอัตโนมัติ

4.4.1.1 ผลการทดลองของการเข้าคู่แบบแม่นตรง (Exact Matching)

การทดลองนี้ใช้เทคนิค 10-fold Cross-Validation เพื่อเทียบผลการแปลชุดคำตอบแต่ละชุดที่ได้จากแต่ละระบบกับชุดคำตอบอ้างอิง จากนั้นจึงนับจำนวนที่สามารถเข้าคู่แบบแม่นตรง (เหมือนกันทั้งหมด) ผลลัพธ์ที่ได้แสดงไว้ในตารางที่ 4.1

ตารางที่ 4.1 ตารางสรุปการเข้าคู่แบบแม่นตรง

ครั้งที่ของการการทำ 10-fold Cross-Validation	จำนวนที่สามารถเข้าคู่แบบแม่นตรง				
	Parsit		jEBMT		จำนวนชุดคำตอบที่นำมาเทียบ
	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	
1	279	2.77%	308	3.06%	10,081
2	239	2.37%	294	2.92%	10,081
3	223	2.21%	289	2.87%	10,081
4	261	2.59%	254	2.52%	10,081
5	237	2.35%	282	2.80%	10,080
6	233	2.31%	296	2.94%	10,080
7	276	2.74%	287	2.85%	10,080
8	247	2.45%	254	2.52%	10,080
9	239	2.37%	283	2.81%	10,080
10	238	2.36%	299	2.97%	10,080
ค่าเฉลี่ย	***2,472***	***2.45%***	***2,846***	***2.82%***	***100,804***

ผลการทดลองแสดงให้เห็นว่า jEBMT ให้ผลการเข้าคู่แบบแม่นตรงโดยเฉลี่ยดีกว่า Parsit ร้อยละ 15 เนื่องจากชุดคำตอบอ้างอิงถูกแปลโดยมนุษย์ทำให้การแปลแบบอิงตัวอย่างให้ผลลัพธ์ที่ใกล้เคียงกับการแปลของมนุษย์มากกว่าการแปลโดยใช้กฎ

4.4.1.2 ผลการทดลองการประเมินค่าอัตโนมัติของ BLUE-4

ในการทดลองนี้ได้ใช้ชุดเครื่องมือ NIST MT Evaluation kit [17] สำหรับประเมินค่าแบบอัตโนมัติของ BLEU-4 ค่าความถูกต้องของการแปลที่ได้จะแสดงไว้ในตารางที่ 4.2

ตารางที่ 4.2 ตารางสรุปผลการประเมินค่าแบบอัตโนมัติของ BLEU-4

ครั้งที่ของการการทำ	ค่าความถูกต้องของการประเมินค่าแบบอัตโนมัติของ BLEU-4	
10-fold Cross-Validation	Parsit	jEBMT
1	0.0276	**0.0301**
2	0.0234	**0.0268**
3	0.0245	**0.0281**
4	0.0265	**0.0303**
5	0.0255	**0.0310**
6	**0.0267**	0.0266
7	**0.0273**	0.0255
8	0.0264	**0.0303**
9	0.0262	**0.0313**
10	0.0262	**0.0292**
ค่าเฉลี่ย	***0.0260***	***0.0289***

จากการทดลองพบว่า jEBMT สามารถแปลโดยเฉลี่ยได้ดีกว่า Parsit ร้อยละ 11 โดยอิงจากผลการทดลอง ปัญหาหลักที่ทำให้ผลการแปลจาก jEBMT ไม่ดีส่วนใหญ่เกิดจากการเรียงประโยคปลายทางที่ผิดพลาด ปัญหาเหล่านี้สามารถแก้ไขได้โดยการรวบรวมคลังข้อความแบบเดี่ยวสำหรับภาษาปลายทางที่มีเนื้อหาเกี่ยวข้องกับขอบเขตที่แปล

4.4.2 ประเมินค่าความถูกต้องของการแปลโดยมนุษย์

ในงานวิจัยชิ้นนี้ได้ทดสอบปัญหาพื้นฐานของระบบแปลภาษา 3 ชนิด อันได้แก่ ความถูกต้องของการเรียงลำดับคำ ความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท ความถูกต้องของการแปลวลี การประเมินความถูกต้องทั้งหมดนี้กระทำโดยนักภาษาศาสตร์

การทดสอบทั้ง 3 ชนิดกระทำโดยสุ่มเลือกประโยคมา 200 ประโยค จำนวน 5 ชุด โดยแต่ละชุดมีข้อมูลไม่ซ้ำกัน จากนั้นจึงให้นักภาษาศาสตร์คนเดียวกันทำการประเมินโดยวิธีนับจำนวนประโยคที่ยอมรับได้ตามเงื่อนไขของแต่ละกรณี

4.4.2.1 ความถูกต้องของการเรียงลำดับคำ

เกณฑ์ในการทดสอบความถูกต้องของการเรียงลำดับคำ จะพิจารณาเฉพาะตำแหน่งที่เหมาะสมของคำแปล แต่อาจไม่ได้เลือกคำแปลที่มีความหมายเหมาะสมกับบริบท ตัวอย่างเช่น “experience in administration” อาจจะถูกแปลเป็น “ลิ้มลอง/ใน/การบริหาร/” จะเห็นได้ว่า “experience” ถูกแปลเป็น “ลิ้มลอง” ซึ่งไม่เหมาะสมกับบริบท แต่ตำแหน่งคำแปลอยู่ในตำแหน่งที่ถูกต้องเพียงแต่ไม่สามารถเลือกคำแปลที่เหมาะสมกับบริบทได้ ผลการทดลองถูกแสดงไว้ในตารางที่ 4.3

ตารางที่ 4.3 ตารางสรุปผลการทดลองความถูกต้องของการเรียงลำดับคำ

ครั้งที่ทดลอง	ค่าความถูกต้องของการเรียงลำดับคำ				
	Parsit		jEBMT		จำนวนที่ใช้ในการทดลอง
	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	
1	100	50.00%	56	28.00%	200
2	91	45.50%	50	25.00%	200
3	113	56.50%	49	24.50%	200
4	99	49.50%	61	30.50%	200
5	98	49.00%	55	27.50%	200
ค่าเฉลี่ย	***100.2***	***50.10%***	***54.2***	***27.10%***	***200***

ผลการทดลองเห็นได้ชัดว่า Parsit มีความถูกต้องในการแก้ไขปัญหาการเรียงลำดับคำโดยเฉลี่ยสูงกว่า jEBMT อย่างเห็นได้ชัดถึงร้อยละ 84 เหตุที่เป็นเช่นนี้สืบเนื่องจากค่าสถิติของตัวแบบเอ็นแกรม มีความเบาบาง (sparse) เกินกว่าจะนำมาตัดสินความเป็นส่วนประชิด (constituent) ของผลการแปลในภาษาปลายทางได้

4.4.2.2 ความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท

เกณฑ์ในการทดสอบความถูกต้องของการเลือกคำให้เหมาะสมตามบริบทจะพิจารณาเฉพาะคำแปลที่เหมาะสมกับประโยคต้นฉบับเท่านั้น ไม่จำเป็นต้องมีการเรียงลำดับคำที่ถูกต้อง ตัวอย่างเช่น “close an account at the bank in her name” อาจจะถูกแปลเป็น “ปิดบัญชี/ณ/ใน/ชื่อ/ธนาคาร/เธอ/” จะเห็นได้ว่าทุกคำมีการเลือกคำที่มีความหมายถูกต้องแต่การเรียงลำดับของคำไม่ถูกต้อง ผลการทดลองแสดงไว้ในตารางที่ 4.4

ตารางที่ 4.4 ตารางสรุปผลการทดลองความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท

ครั้งที่ทดลอง	ค่าความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท				
	Parsit		jEBMT		จำนวนที่ใช้ในการทดลอง
	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	
1	88	44.00%	124	62.00%	200
2	100	50.00%	112	56.00%	200
3	92	46.00%	121	60.50%	200
4	95	47.50%	106	53.00%	200
5	87	43.50%	111	55.50%	200
ค่าเฉลี่ย	***92.4***	***46.20%***	***114.8***	***57.40%***	***200***

ผลการทดลองแสดงให้เห็นว่า jEBMT มีความถูกต้องในการเลือกใช้คำให้เหมาะสมตามบริบทระดับประโยคมากกว่า Parsit พอสมควรในอัตราร้อยละ 24 โดยเฉลี่ย ทั้งนี้เป็นเพราะค่าสถิติของตัวแบบเอ็นแกรมสามารถนำมาช่วยแก้ไขความกำกวมในการเลือกคำแปลตามบริบทได้แม้ว่าข้อมูลจะมีปริมาณเบาบางก็ตาม

4.4.2.3 ความถูกต้องของการแปลวลี

เกณฑ์ในการทดสอบความถูกต้องของการแปลวลี จะพิจารณาเฉพาะคำแปลของวลีที่ปรากฏในประโยคต้นฉบับเท่านั้น ตัวอย่างเช่น “I have an account with the bank” อาจจะถูกแปลเป็น “ฉัน/มี/บัญชีเงินฝาก ที่ ธนาคาร/” จะเห็นได้ว่า “an account with the bank” ถูกแปลได้อย่างถูกต้องเป็น “บัญชีเงินฝาก ที่ ธนาคาร” ผลการทดลองแสดงไว้ในตารางที่ 4.5

ตารางที่ 4.5 ตารางสรุปผลการทดลองความถูกต้องของการแปลวลี

ครั้งที่ทดลอง	ค่าความถูกต้องของการแปลวลี				
	Parsit		jEBMT		จำนวนที่ใช้ในการทดลอง
	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	
1	56	28.00%	108	54.00%	200
2	60	30.00%	112	56.00%	200
3	59	29.50%	101	50.50%	200
4	49	24.50%	99	49.50%	200
5	54	27.00%	97	48.50%	200
ค่าเฉลี่ย	***55.6***	***27.80%***	***103.4***	***51.70%***	***200***

ผลการทดลองแสดงให้เห็นอย่างชัดเจนว่า jEBMT มีความถูกต้องในการแปลวลีมากกว่า Parsit เป็นอย่างมากถึงร้อยละ 85 โดยเฉลี่ย ทั้งนี้เป็นเพราะ jEBMT สามารถรู้จำ (recognize) และแปลวลีที่มีส่วนประชิดต่อเนื่องกัน (continuous constituent) จากคลังข้อความแบบคู่โดยใช้ค่าสถิติของตัวแบบเอ็นแกรมได้เป็นอย่างดีแม้ว่าจะมีปริมาณข้อมูลเบาบางก็ตาม

4.5 สรุปผลการประเมินค่าการแปล

การใช้ตัวแบบเอ็นแกรมสามารถแก้ปัญหาการเลือกใช้คำให้เหมาะสมตามบริบทระดับประโยคและการแปลวลีได้เป็นอย่างดี ในขณะที่การประยุกต์ใช้ตัวแบบเอ็นแกรมกับการแก้ไขปัญหาการเรียงลำดับคำยังคงต้องการปริมาณข้อมูลเชิงสถิติเพิ่มอีกเป็นจำนวนมาก ทั้งนี้เป็นเพราะต้องตรวจสอบความเป็นส่วนประชิดซึ่งมีความหลากหลายในการสร้าง (diversity of formation) เป็นอย่างมาก

ผลการประเมินค่าโดยรวมสามารถสรุปได้ว่าระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างมีความสามารถแปลได้ในระดับหนึ่ง โดยผลการแปลส่วนใหญ่จะสามารถจับใจความได้ แต่ยังไม่ถูกต้องตามหลักไวยากรณ์ รูปแบบผลการแปลสามารถแบ่งได้เป็นประเภทต่างๆ ดังนี้

1. ไม่สามารถจับใจความได้ (แทนด้วยสัญลักษณ์ ☹)
2. สามารถอ่านจับใจความสำคัญได้ แต่ไวยากรณ์ไม่ถูกต้อง (แทนด้วยสัญลักษณ์ ☺)
3. สามารถอ่านจับใจความสำคัญได้ ไวยากรณ์ถูกต้อง แต่ไม่ตรงกับคำตอบอ้างอิง (แทนด้วยสัญลักษณ์ ☺☺)
4. สามารถอ่านจับใจความสำคัญได้ ไวยากรณ์ถูกต้อง และตรงกับคำตอบอ้างอิง (แทนด้วยสัญลักษณ์ ☺☺☺)

<u>ตัวอย่างการแปลพร้อมบทวิเคราะห์</u>

ข้อความรับเข้า:	for mutual advantage
คำตอบอ้างอิง:	เพื่อ ประโยชน์ของ ทั้งสองฝ่าย
คำตอบจาก Parsit:	สำหรับ ผล ประโยชน์ สอง ฝ่าย
คำตอบจาก jEBMT:	สำหรับ ซึ่ง กันและกัน ความ ได้ เปรียบ
สรุป:	Parsit ☺ jEBMT ☹
วิเคราะห์:	คำตอบจาก jEBMT มีการเรียงลำดับคำของคำแปลที่ผิดพลาด ทำให้ไม่สามารถจะเข้าใจใจความสำคัญของประโยคต้นทางได้ คำตอบถูกต้องควรเป็น สำหรับ ความ ได้ เปรียบ ซึ่ง กันและกัน ที่เป็นเช่นนี้เกิดจากค่าสถิติของการประกอบส่วนประชิดมีความเบาบางและไม่มากพอที่จะทำให้การเรียงลำดับคำถูกต้อง

ข้อความรับเข้า:	a superior in age
คำตอบอ้างอิง:	ผู้ มี วัยวุฒิ สูง
คำตอบจาก Parsit:	กว่า ผู้ใหญ่ ใน อายุ
คำตอบจาก jEBMT:	ผู้ที่ มีอายุ สูง กว่า
สรุป:	Parsit ☹ jEBMT ☺
วิเคราะห์:	คำตอบจาก Parsit มีการสังเคราะห์ประโยคที่ผิดพลาด คำตอบจาก jEBMT มีความถูกต้องเนื่องจากคำเหล่านี้ คือ “มีอายุ” และ

คำว่า “สูงกว่า” เคยมีการปรากฏแบบประชิดร่วมกันในคลังข้อความ ทำให้ค่าสถิติบ่งชี้ให้เกิดการเรียงลำดับที่ถูกต้อง

ข้อความรับเข้า: reach an agreement
คำตอบอ้างอิง: ได้ ข้อตกลง
คำตอบจาก Parsit: เข้าถึง ข้อ ตกลง
คำตอบจาก jEBMT: บรรลุ ข้อตกลง
สรุป: Parsit ☹ jEBMT ☺☺
วิเคราะห์: คำตอบที่ได้จาก Parsit ใช้คำว่า เข้าถึง ซึ่งมีใจความสำคัญไม่ตรงกับประโยคต้นทาง ในขณะที่คำตอบจาก jEBMT สามารถแปลได้ดีกว่าคำตอบอ้างอิง แต่ใช้คำไม่ตรงกับคำตอบอ้างอิง ซึ่งแสดงให้เห็นว่าตัวอย่างภายในคลังข้อความแบบคู่ให้ผลลัพธ์การแปลที่เหมาะกว่าการแปลโดยการใช้กฎของ Parsit

ข้อความรับเข้า: have a high aim in life
คำตอบอ้างอิง: มี เป้าหมาย ที่ สูง ใน ชีวิต
คำตอบจาก Parsit: มี จุด มุ่งหมาย ที่ สูง ใน ชีวิต
คำตอบจาก jEBMT: มี สูง ใน ชีวิต เป้า
สรุป: Parsit ☺☺ jEBMT ☹
วิเคราะห์: คำตอบจาก Parsit สามารถแปลได้ดีกว่าคำตอบอ้างอิง แต่ใช้คำไม่ตรงกับคำตอบอ้างอิงแต่ยังคงความหมายถูกต้อง
คำตอบจาก jEBMT มีการเรียงลำดับคำของคำแปลที่ผิดพลาด ทำให้ไม่สามารถจะเข้าใจความหมายที่แท้จริงของประโยคต้นทางได้ ประโยคที่ถูกควรเป็น มี เป้าสูง ใน ชีวิต ที่เป็นเช่นนี้เกิดจากค่าสถิติของการประกอบส่วนประชิดมีความเบาบางและไม่มากพอที่จะทำให้การเรียงลำดับคำถูกต้อง

ข้อความรับเข้า: the correct answer

คำตอบอ้างอิง: คำตอบ ที่ ถูกต้อง

คำตอบจาก Parsit: คำ ตอบ ที่ ถูก ต้อง

คำตอบจาก jEBMT: ดัด นิสัย ตอบ

สรุป: Parsit ☺☺☺ jEBMT ☹

วิเคราะห์: คำตอบจาก Parsit ถูกต้องและเหมือนคำตอบอ้างอิง

คำตอบจาก jEBMT เลือกคำผิดพลาดจากส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรม โดยแปลคำว่า the correct เป็น ดัด นิสัย เนื่องจากภายในระบบมีการตัดคำบ่งชี้เฉพาะ (article เช่น a, an, the) ทำให้ the correct ถูกตัดทอนเหลือ correct ซึ่งในคลังข้อความมีค่าสถิติของ correct ที่อยู่ต้นประโยคเป็นคำกริยา ทำให้ correct ของข้อความรับเข้าถูกเลือกความหมายเป็น ดัด นิสัย

ข้อความรับเข้า: retire from the army

คำตอบอ้างอิง: ปลดเกษียณ จาก กองทัพ

คำตอบจาก Parsit: ลา ออก กองทัพ

คำตอบจาก jEBMT: ปลดประจำการ จาก ทหารบก

สรุป: Parsit ☹ jEBMT ☹

วิเคราะห์: คำตอบจาก Parsit และ jEBMT ต่างมีใจความสำคัญที่ไม่ถูกต้องซึ่งเป็นปัญหาที่เรียกว่าปัญหาในการเลือกคำที่เหมาะสมกับบริบท เนื่องจาก retire ควรมีความหมายเป็น “เกษียณอายุ” ซึ่งมีความที่แตกต่างจาก ลา ออก ซึ่งเป็นคำตอบจาก Parsit และ ปลดประจำการ ซึ่งเป็นคำตอบจาก jEBMT

จากการวิเคราะห์ผลการแปลพบว่าระบบ jEBMT สามารถแปลวลีและสำนวนได้อย่างถูกต้องแม่นยำมากกว่า Parsit ในขณะที่ Parsit สามารถรองรับการแปลตามกฎไวยากรณ์ได้ดีกว่า jEBMT สาเหตุของผลการแปลดังกล่าวเกิดจากปัจจัยสำคัญ 2 ประการ ได้แก่ สภาพการเกาะกลุ่มของวลีและสำนวน (Clustering of Phrase and Idioms) และความเบาบางของข้อมูลที่ใช้สอนระบบ (Data Sparseness)

สภาพการเกาะกลุ่มของวลีและสำนวนเป็นปัจจัยสำคัญที่ทำให้ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรมสามารถช่วยปรับปรุงผลการแปลได้ จากการทดลองพบว่า วลีและสำนวนมักอยู่ติดกันภายใน 3 คำตามตำแหน่งของคำ มีเพียงวลีและสำนวนส่วนน้อยที่อยู่แยกจากกันเกิน 3 คำ ทำให้ผลการแปลวลีและสำนวนของ jEBMT จึงดีกว่าของ Parsit

ความเบาบางของข้อมูลเป็นปัจจัยที่ทำให้ผลความถูกต้องของการแปลของ jEBMT ลดต่ำลง การอนุมานกฎไวยากรณ์จากคลังข้อความจำเป็นต้องใช้ข้อมูลจำนวนมาก Keh-Yih Su, Tung-Hui Chiang and Jing-Shin Chang [17] กล่าวว่า จำนวนประโยคตัวอย่างสำหรับสร้างตัวแบบเอ็นแกรมสำหรับภาษาใดๆ ที่มีจำนวนคำในภาษา w คำ จะต้องมีอย่างน้อย $10w^3$ ตัวอย่าง เนื่องจากคำในภาษาอังกฤษและภาษาไทยมีจำนวนมากในระดับมากกว่าแสนคำ คลังข้อความที่ใช้จึงไม่เพียงพอต่อการสร้างตัวแบบเอ็นแกรมของภาษา ขณะที่ Parsit ซึ่งใช้อิงกฎไวยากรณ์ในการแปลที่สร้างจากมนุษย์โดยตรง จึงทำงานได้ดีกว่า jEBMT ในการแปลประโยคโดยใช้กฎ