KKU Engineering Journal

http://www.en.kku.ac.th/enjournal/th/

การปรับปรุงประสิทธิภาพของระบบตรวจแก้คำผิดภาษาไทยโดยใช้มีมีติกอัลกอริทึม

Improvement of Thai error correction system by memetic algorithm

กริช สมกันธา* และ วิไลพร กุลตังวัฒนา

Krit Somkantha* and Wilaiporn Kultangwattana

สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุดรธานี จังหวัดอุดรธานี 41000

Department of Computer Science and Information Technology, Faculty of Science, Udon Thani Rajabhat University, Udon Thani, Thailand, 41000.

Received August 2013

Accepted December 2013

บทคัดย่อ

งานวิจัยนี้นำเสนอเทคนิควิธีการปรับปรุงประสิทธิภาพของระบบตรวจแก้คำผิดภาษาไทยโดยใช้มีมีติกอัลกอริทึม ซึ่งในงานวิจัยกระบวนการส่งผ่านโทเค็นถูกใช้ในการสร้างกราฟของกลุ่มคำและรูปแบบจำลองภาษาถูกใช้ในการตรวจสอบประโยคที่ถูกต้อง โดยกระบวนการตรวจแก้คำผิดเริ่มจากการสร้างกลุ่มคำที่เป็นไปได้จากวิธีการส่งผ่านโทเค็น จากนั้นประโยคที่ถูกต้องจะถูกค้นหาโดยใช้กระบวนการของมีมีติกอัลกอริทึมด้วยค่าความเหมาะสมที่ดีที่สุดจากรูปแบบจำลองทางภาษา จากวิธีการส่งผ่านโทเค็นในกรณีของประโยคที่มีจำนวนตัวอักษรมากขนาดของการค้นหาจะมีขนาดใหญ่มาก ซึ่งในงานวิจัยนี้สามารถแก้ปัญหาในการค้นหาโดยการใช้วิธีการของมีมีติกอัลกอริทึม โดยวิธีการของมีมีติกอัลกอริทึมจะถูกใช้ในการค้นหาประโยคที่ถูกต้องเพื่อที่จะลดเวลาในการค้นหาประโยคที่ดีที่สุด ในการทดสอบประสิทธิภาพของวิธีการที่นำเสนอจะถูกประเมินและถูกทดสอบด้วยการค้นหาทั้งหมดและเจเนติกอัลกอริทึม ซึ่งจากผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถทำงานได้อย่างมีประสิทธิภาพและมีประสิทธิภาพมากกว่าวิธีที่นำมาเปรียบเทียบ โดยวิธีการที่นำเสนอสามารถค้นหาประโยคที่ดีที่สุดได้อย่างถูกต้องและรวดเร็ว

คำสำคัญ : มีมีติกอัลกอริทึม เจเนติกอัลกอริทึม รูปแบบจำลองทางภาษา ระบบตรวจแก้คำผิด

Abstract

This paper presents an efficient technique for improving the efficiency of Thai error correction system by using memetic algorithm. In this paper, the token passing algorithm is used for constructing word graph and the language model is used checking the correct sentence. The correction process starts with word graph construction from token passing algorithm, then the correct sentence are searched by memetic algorithm with the best fitness function from language model. For a long sentence from the token passing algorithm, a search space is very huge which can be resolved by using memetic algorithm. The memetic algorithm is used for searching the correct sentence in order to reduce the analysis time. The performance of the proposed method are evaluated and compared to the full search and genetic algorithm. From the

*Corresponding author. Tel.: +66(0)812672205; fax: +66(0)42341614-5

Email address: dr_krit@hotmail.com

experimental results show that the proposed method performs very well and yields better performance more than the compared method. The proposed method can search the best sentence accurately and quickly.
Keywords : Memetic algorithm, Genetic algorithm, Language model, Error correction system

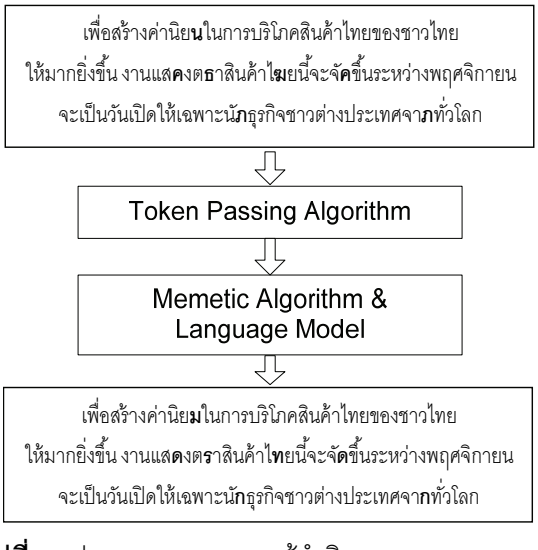
1. บทนำ

ในงานวิจัยด้านการประมวลภาษาทางธรรมชาติในปัจจุบันได้มีการวิจัยจำนวนมากในภาษาต่างๆ [1-3] แต่อย่างไรก็ตามงานวิจัยในการประมวลผลภาษาทางธรรมชาติในภาษาไทยยังกระทำได้ยาก เช่นในการตรวจแก้คำผิดภาษาไทยจะกระทำได้ซับซ้อนกว่าภาษาอังกฤษ เนื่องจากภาษาไทยไม่มีช่องว่างระหว่างคำ แต่อย่างไรก็ตามก็ได้มีนักวิจัยนำเสนอเทคนิคต่างๆ ในการตรวจแก้คำผิด ซึ่งในงานวิจัยด้านการตรวจแก้คำผิดได้มีเทคนิคหลายวิธี [4-7] ดังเช่นการใช้ระบบผู้เชี่ยวชาญโดยจะเป็นการตรวจแก้คำผิดที่ใช้การตัดสินใจจากการเลือกค่าน้ำหนักต่างๆ การใช้ Winnow golding เป็นการแก้คำผิดจากการรวบรวมคำที่มีลักษณะคล้ายกันรวมกลุ่มไว้ด้วยกันเพื่อทำการแยกแยะรายละเอียดว่ามีความสัมพันธ์กับคำใกล้เคียงอย่างไรบ้าง และการใช้ Tri-Gram และ Winnow เป็นการตรวจแก้คำผิดโดยการนำเอา Tri-Gram เข้ามาช่วยตรวจสอบแล้วเลือกเฉพาะคำที่น่าจะเป็นไปได้แล้วนำไปตรวจสอบต่อไป ซึ่งในแต่ละวิธีก็จะมีข้อดีข้อเสียแตกต่างกัน

ในงานวิจัยนี้ได้ใช้การตรวจสอบคำและรูปแบบจำลองทางภาษา เพื่อตรวจสอบความถูกต้องของภาษาไทย ซึ่งเป็นการแก้ไขคำผิดในรูปแบบของคำและในรูปแบบของประโยค ในส่วนรูปแบบของคำจะตรวจสอบคำที่มีอยู่ในพจนานุกรมโดยใช้วิธีการส่งผ่านค่าไปยังตัวอักษรแต่ละตัวหรือการส่งผ่านโทเค็น (Token passing algorithm) [8,16] ซึ่งจะทำให้ได้คำที่ถูกต้องและมีอยู่ในพจนานุกรม ส่วนในรูปแบบของประโยคจะใช้การจำลองรูปแบบทางภาษา [3,7,9,16] (Language model) ซึ่งเป็นฐานข้อมูลทางภาษาที่เก็บรวบรวมขึ้นมา โดยรูปแบบจำลองทางภาษาจะถูกนำมาใช้ในการตรวจคำผิดในส่วนของประโยค จากทั้งสองส่วนนี้กระทำทั้งในส่วนของการตรวจสอบของคำและการตรวจสอบของประโยค ก็จะส่งผลให้มีประสิทธิภาพที่ดีมากขึ้น ยกตัวอย่างเช่นประโยค “**นน**ของน้องหมด**ยา**ยุ” ถ้าเป็นการตรวจสอบในรูปแบบของคำเพียงอย่างเดียวก็จะถูกต้องเพราะทุกคำมีอยู่ในพจนานุกรม แต่ถ้าตรวจสอบในรูปแบบของแบบจำลองทางภาษาก็จะเป็นประโยคที่ผิด ซึ่งประโยคที่ถูกต้องคือ “นมของน้องหมดอายุ” ในงานวิจัยที่นำเสนอผลลัพธ์ของการส่งผ่านโทเค็นจะมีจำนวนผลลัพธ์ที่ได้จำนวนมากซึ่งอาจจะทำให้มีประโยคที่เป็นไปได้จำนวนล้านกว่าประโยค ดังนั้นถ้าใช้กระบวนการค้นหาทุกประโยคที่เป็นไปได้แล้วนำไปเข้าสู่รูปแบบการตรวจสอบประโยคจากรูปแบบจำลองทางภาษาก็จะส่งผลให้ใช้เวลานานหรืออาจจะไม่พบคำตอบที่ต้องการ ดังนั้นในงานวิจัยนี้จึงได้เสนอวิธีการมีมีติกอัลกอริทึม (Memetic algorithm) [10-12] เพื่อใช้ในการค้นหาประโยคที่ถูกต้องจากค่าความคลุมเคลือที่ดีที่สุด เพื่อทำให้ระบบตรวจแก้คำผิดภาษาไทยมีประสิทธิภาพมากขึ้น

2. วิธีการวิจัย

ในวิธีการดำเนินงานวิจัยเริ่มจากประโยคภาษาไทยจากนั้นนำไปเข้าสู่กระบวนการตรวจสอบคำโดยใช้ส่งผ่านค่าไปยังตัวอักษรแต่ละตัว จากนั้นเมื่อได้ผลลัพธ์ก็จะใช้มีมีติกอัลกอริทึมในการค้นหารูปประโยคที่ถูกต้องโดยจะมีรูปแบบจำลองทางภาษาเป็นฟังก์ชันความเหมาะสม (Fitness function) ดังแสดงวิธีการดำเนินการวิจัยในรูปที่ 1



รูปที่ 1 รูปแบบของการตรวจแก้คำผิด

2.1 วิธีการตรวจสอบคำ

ในการตรวจสอบคำจะใช้วิธีการส่งผ่านโทเค็น (Token passing algorithm) [8,16] ซึ่งในงานวิจัยนี้จะเพิ่มประสิทธิภาพของระบบตรวจแก้คำผิดโดยจะเพิ่มตัวอักษรที่เป็นไปได้ในแต่ละตัว ซึ่งจะทำให้สามารถแก้ไขเป็นคำที่ถูกต้องได้ โดยตัวอักษรที่มีการเพิ่มเข้าไปจะเป็นตัวอักษรที่ใกล้เคียงกันและตัวอักษรที่มีการพิมพ์ผิดบ่อย โดยจะกำหนดสูงสุดไว้ 5 ตัวอักษร ดังตัวอย่างของตัวอักษรที่เพิ่มเข้าไปดังเช่น

ตารางที่ 1 ตัวอย่างตัวอักษรที่ถูกกำหนดเพิ่มเพื่อใช้เพิ่มความถูกต้องของประโยค

ก	ข	ค	ผ
ภ	ช	ด	พ
ถ	จ	ต	ฝ
ด	บ	ศ	บ
ฏ	ซ	อ	ป

ในการตรวจแก้ความผิดพลาดยิ่งถ้ามีการเพิ่มจำนวนตัวอักษรที่เป็นไปได้มากขึ้นก็จะทำให้การตรวจแก้ความผิดพลาดยิ่งมีความถูกต้องสูงขึ้น แต่อย่างไรก็ตามในการเพิ่มจำนวนของตัวอักษรก็เป็นสิ่งที่สิ้นเปลืองเพราะจะทำให้ได้คำที่เป็นไปได้จำนวนมากเช่นมีการใช้ตัวอักษรเพิ่ม 10 ตัว ของคำว่า “การสอบ” ซึ่งจะสามารถสร้างคำที่เป็นไปได้สูงสุดถึง 10^6 ตัว ประมาณ 1,000,000 คำ (สมมุติว่าเป็นคำที่มีอยู่ในพจนานุกรมทั้งหมด) ดังนั้นในงานวิจัยนี้จึงมีการจำกัดอยู่ที่ 5 ตัวอักษร ซึ่งจากการทดสอบเพียงพอต่อความถูกต้องของระบบตรวจแก้คำผิดภาษาไทย

จากตัวอักษรทั้งหมดที่เป็นไปได้จะถูกนำไปใช้ในกระบวนการการส่งผ่านโทเค็น โดยกระบวนการนี้จะทำการตรวจสอบตัวอักษรแต่ละตัวแล้วนำไปเปรียบเทียบกับพจนานุกรม ถ้าพบก็จะเก็บไว้เป็นคำที่สามารถเป็นไปได้ ถ้าไม่พบหรือไม่มีโอกาสที่จะเกิดเป็นคำที่มีอยู่ในพจนานุกรมก็จะทำการตัดทิ้งไป แต่ถ้ามีโอกาสก็จะเก็บไว้เพื่อสร้างเป็นคำต่อไป ซึ่งผลลัพธ์ที่ได้จากวิธีการนี้ดังแสดงในตารางที่ 2

ตารางที่ 2 ตัวอย่างกลุ่มคำที่ได้มาจากวิธีการส่งผ่านโทเค็นของประโยค “ให้กับสินค้าเกษตร”

จุดเริ่ม	จุดสิ้นสุด	ตัวอักษร
0	3	ไท้
0	3	ให้
0	3	ไห้
0	3	โท้
0	3	ไห่
0	3	โห่
3	6	กับ
3	6	ถับ
6	9	ลืม
6	9	สิน
6	12	สินค้า
9	12	ค้า
9	12	ค่า
9	12	ด่า
9	12	คัว
12	17	เกษตร

2.2 วิธีการตรวจสอบรูปแบบประโยค (Language model)

การตรวจสอบรูปแบบประโยคในงานวิจัยได้ใช้วิธีการจากรูปแบบจำลองของภาษาซึ่งเป็นกรรมวิธีทางสถิติที่นำมาใช้สร้างแบบจำลอง [7,9,16] ซึ่งผู้วิจัยได้มีการเก็บข้อมูลของประโยคที่ถูกต้องไว้เพื่อทำการใช้ในรูปแบบของแบบจำลอง

แบบจำลองทางสถิติ N-Gram สามารถคำนวณหาค่าความน่าจะเป็นของคำ $P(W)$ เปรียบเสมือนว่า W มีความถี่เท่าไหร่ในการเกิดคำนั้นในฐานข้อมูลซึ่งในงานวิจัยนี้ได้ใช้ Tri-Gram ในการตรวจแก้คำผิดภาษาไทย เช่น P(การ ดำ นาน) จะเกิดความถี่น้อยกว่า P(การ ทำ งาน) ดังนั้นกลุ่มของคำที่มีการเกิดความถี่ยิ่งมากแสดงว่าน่าจะเป็นกลุ่มของคำที่ถูกต้องสูง เปรียบเสมือนการคำนวณค่าความน่าจะเป็นของคำโดยดูจากความถี่ของคำที่เกิดขึ้นในฐานข้อมูล ดังนั้นเราสามารถคำนวณความน่าจะเป็นของคำได้ดังนี้

$$\begin{aligned} P(W) &= P(w_1, w_2, ..., w_n) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\cdots P(w_n \mid w_1, w_2, ..., w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i \mid w_1, w_2, ..., w_{i-1}) \end{aligned} \quad (1)$$

ซึ่ง $P(w_i \mid w_1,w_2,\ldots,w_{i-1})$ คือความน่าจะเป็นของ w_i ที่สืบเนื่องมาจากลำดับของคำที่เกิดขึ้นก่อนหน้า w_i

จากสมการข้างบนนี้ w_i จะขึ้นอยู่กับคำทั้งหมดที่อยู่ก่อนหน้า w_i สำหรับคำศัพท์ขนาด v จะมี v^{i-1} ของข้อมูลเก่าที่เก็บไว้ ดังนั้น $P(w_i \mid w_1,w_2,\ldots,w_{i-1})$ จึงมีขนาด v^i ค่าที่นำมาใช้ในการคำนวณ ในทางปฏิบัติเราแบ่งแยกกลุ่มคำออกเป็นช่วงละเท่าๆ กัน โดยกลุ่มคำสามารถแบ่งได้โดยใช้คำก่อนหน้า $w_1,w_2,\ldots,w_{i-1}$ เป็นจำนวนช่วงเท่าๆ กัน ซึ่งจะได้ว่า $P(w_i|w_1,w_2,\ldots,w_{i-1})$ หาจากกลุ่มคำโดยแบ่งเป็นส่วนย่อยๆ จำนวน i คำ เช่นถ้าเราใช้การแบ่งเป็นกลุ่มละสามคำ (Tri - gram) เราจะหาความน่าจะเป็นของคำจาก $P(w_i \mid w_{i-2},w_{i-1})$ ดังนั้นเราสามารถหาค่าความน่าจะเป็นได้โดยการนับจำนวนเหตุการณ์ ที่เราสนใจในคลังฐานข้อมูล (Corpus) โดยให้ $C(w_{i-2},w_{i-1},w_i)$ เป็นการนับจำนวน w_{i-2},w_{i-1},w_i ที่เกิดขึ้นในคลังฐานข้อมูลที่นำมาฝึก ดังนั้นจึงสามารถหาค่าความน่าจะเป็นได้จาก

$$P(w_i \mid w_{i-2}\ w_{i-1}) \approx \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})} \qquad (2)$$

จากนั้นใช้เทคนิคทำให้ราบเรียบ [7] เพื่อป้องกันให้ค่าความน่าจะเป็นมีค่าเป็นศูนย์ ซึ่งจะทำให้เกิดการกระจายค่าของความน่าจะเป็นให้มีความราบเรียบมากขึ้น อีกทั้งยังทำให้รูปแบบมีความถูกต้องเพิ่มมากขึ้น

ผลลัพธ์ที่ได้จากรูปแบบจำลองทางภาษาซึ่งจะให้ค่าความคลุมเครือของภาษาต่ำในกรณีที่ประโยคนั้นถูกต้อง และจะให้ค่าความคลุมเครือของภาษาสูงในกรณีที่ประโยคนั้นไม่ถูกต้องมาก [3,7] ดังตัวอย่างที่แสดงในตารางที่ 3 ซึ่งข้อมูลที่นำมาสร้างเป็นรูปแบบจำลองทางภาษาได้มาจากการเก็บข้อมูลจากฐานข้อมูลของหนังสือพิมพ์ในทุกหัวข้อข่าวเป็นเวลา 3 เดือน

ตารางที่ 3 ผลลัพธ์ความถูกต้องของประโยคที่ได้จากแบบจำลองทางภาษา

ประโยค	ความคลุมเครือ (ppx)
หด ลอน	21015
หด ลอน	46548
ทด สอบ	11169
หด สอบ	20690

2.3 วิธีการค้นหาประโยคที่ถูกต้องโดยใช้มีมีติกอัลกอริทึม (Memetic algorithm)

จากผลลัพธ์ที่ได้จากวิธีการส่งผ่านโทเค็นดังแสดงในตารางที่ 2 แสดงให้เห็นว่ายิ่งถ้ามีจำนวนตัวอักษรยิ่งมากเท่าไหร่ก็จะมีโอกาสที่จะเกิดคำจำนวนมาก ซึ่งจะทำให้สามารถค้นหาประโยคที่ถูกต้องได้กระทำได้ยากดังนั้นในงานวิจัยนี้จึงได้มีการประยุกต์วิธีการของมีมีติกอัลกอริทึมเพื่อใช้ในการค้นหาประโยคที่ถูกต้อง โดยเทคนิคมีมีติกอัลกอริทึมเป็นเทคนิคการค้นหาที่จำลองมาจากกระบวนการทางธรรมชาติที่สามารถแก้ปัญหาที่ซับซ้อนได้ ซึ่งวิธีการที่สำคัญของเทคนิควิธีมีมีติกอัลกอริทึมคือการค้นหาคำตอบแบบเฉพาะที่ (Local search) เพื่อช่วยปรับปรุงการหาคำตอบให้ดียิ่งขึ้น [10-12] ซึ่งในการตรวจแก้คำผิดภาษาไทยจะเริ่มจากตัวอักษรทั้งหมดทีละประโยคหรือทีละ 1 บรรทัด ในกรณีที่ไม่มีการเว้นวรรค โดยจะดำเนินการตรวจสอบทีละประโยคไปจนครบทั้งหมด โดยสามารถอธิบายเป็นขั้นตอนดังนี้

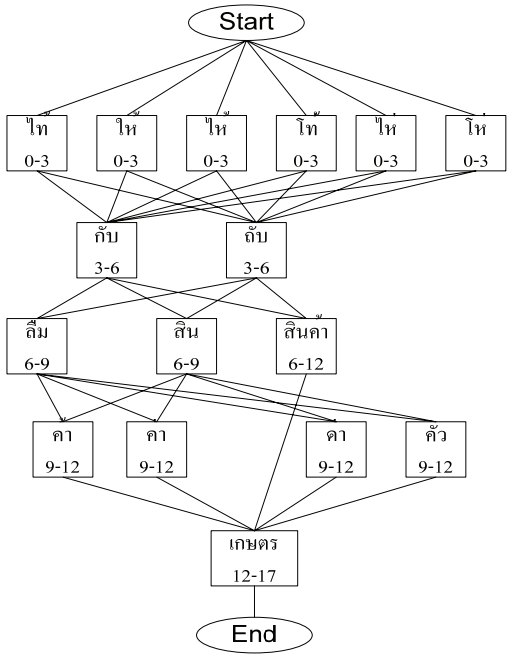
```
Function Memetic Algorithm
{
    Gen = 0;
    Population_Initialization P(Gen);
    Evaluate P(Gen);
    Local Search P(Gen);
    while not terminated(do) {
    Keep_best P(Gen);
    P_Select(Gen) = Select_Parent P(Gen);
    P'(Gen) = Crossover P_Select(Gen);
    P''(Gen) = Mutation P'(Gen);
    Evaluate P''(Gen);
    P''(Gen) = Local_Search P''(Gen);
    P(Gen+1) = Select(P''(Gen) U  Keep_best));
    Gen = Gen + 1;
    } Best_Sentence = Select_Best(Gen);
}
```

รูปที่ 2 กระบวนการของระบบตรวจแก้คำผิดภาษาไทยโดยใช้มีมีติกอัลกอริทึม

ซึ่งในฟังก์ชัน P(Gen) คือประชากรในแต่ละรุ่น Population_Initialization P(Gen) คือการสร้างประชากรเริ่มต้น Evaluate P(Gen) คือการประเมินค่าความเหมาะสม Local Search P(Gen) คือการค้นหาคำตอบเฉพาะที่

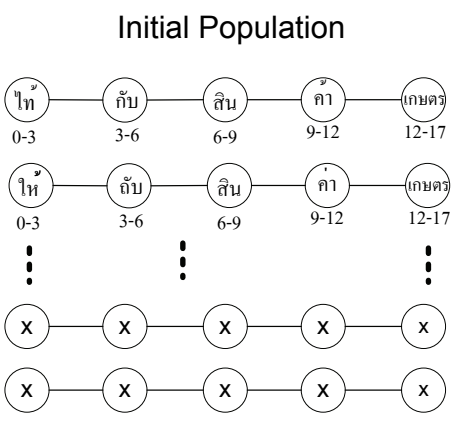
Keep_best P(Gen) คือการรักษาคำตอบที่ดีไว้ไม่ให้สูญเสีย Select_Parent P(Gen) คือการเลือกประชากรพ่อแม่ที่จะดำเนินการทางพันธุศาสตร์ Crossover P_Select(Gen) และ Mutation P'(Gen) คือการดำเนินการทางพันธุศาสตร์เพื่อหาคำตอบที่ดี Select(P''(Gen) U Keep_best)) คือการเลือกประชากรรุ่นถัดไปที่ดี Select_Best(Gen) คือการเลือกคำตอบที่ดีที่สุด

(1) Chromosome representation การกำหนดรูปแบบโครโมโซม เริ่มต้นของกระบวนการมีมีติกอัลกอริทึมจะต้องมีการปรับรูปแบบของปัญหาให้อยู่ในรูปแบบที่เหมาะสมเพื่อที่จะนำไปใช้ในการกระบวนการของมีมีติกอัลกอริทึมได้อย่างมีประสิทธิภาพ ซึ่งในงานวิจัยนี้ข้อมูลจะถูกเก็บในรูปแบบของเส้นทางของกลุ่มคำ (Word Graph) ที่ได้มาจากวิธีการส่งผ่านโทเค็น ดังแสดงในรูปที่ 3



รูปที่ 3 เส้นทางของกลุ่มคำที่เป็นไปได้ทั้งหมด

(2) Population_Initialization การสร้างประชากรเริ่มต้น ในขั้นตอนนี้เป็นการสร้างประชากรต้นแบบโดยการสุ่มค่าจำนวนประชากรตามที่ได้ออกแบบ โดยจะสุ่มมาจากเส้นทางของกลุ่มคำ ซึ่งในงานวิจัยจำนวนประชากรเริ่มต้น = 30 ซึ่งได้มากจากการทดสอบที่ดีที่สุด

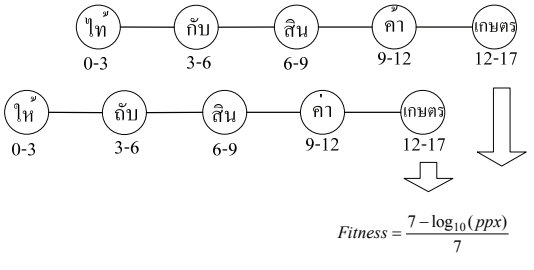


รูปที่ 4 ประชากรเริ่มต้น

(3) Evaluate การประเมินค่าความเหมาะสม (Fitness function) ในขั้นตอนนี้เป็นการดูค่าความเหมาะสมของประโยคที่ถูกต้องซึ่งในงานวิจัยเราสามารถประเมินหาค่าความเหมาะสมโดยการปรับค่าความคลุมเครือ (ppx) ให้อยู่ในช่วง 0-1 โดยถ้าเข้าใกล้ 0 แสดงว่าประโยคนั้นมีความน่าจะเป็นที่ถูกต้องแย่ที่สุด และถ้าเข้าใกล้ 1 แสดงว่าประโยคนั้นมีความน่าจะเป็นที่ถูกต้องมากที่สุด ซึ่งค่าความน่าจะเป็นของประโยคหรือค่าความเหมาะสมสามารถคำนวณได้จากสมการ

$$Fitness = \frac{7 - \log_{10}(ppx)}{7} \quad (3)$$

ซึ่ง *ppx* คือค่าความคลุมเครือของประโยค (Perplexity)



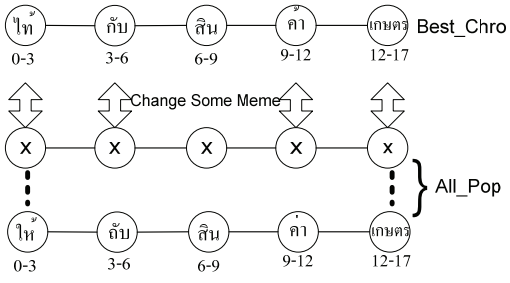
รูปที่ 5 การหาค่าความเหมาะสม

(4) Local search การค้นหาเฉพาะที่ (Local search) ซึ่งในกระบวนการของมีมีติกอัลกอริทึมการค้นหาเฉพาะที่จะเป็นส่วนที่สำคัญที่จะทำให้สามารถหาคำตอบได้อย่างรวดเร็วและมีประสิทธิภาพมากขึ้น ซึ่งการค้นหาเฉพาะที่จะเปลี่ยนแปลงคำตอบที่ดีให้ดียิ่งขึ้นไป เนื่องจากคำตอบที่ได้ในแต่ละตัวมักจะมีตัวที่ดีรวมอยู่ด้วย โดยสามารถอธิบายเป็นขั้นตอนดังนี้

```
Function local search
{
    Best_Chro = Select_Best P(Gen);
    All_Pop = P(Gen) – Best_Chro;
    Num_Meme = Length of Chromosome;
    Count = 1;
     while (Count < Num_Meme)
    {
        if (Best_Chro(Count) < All_Pop(Count))
        {
        Swap(Best_Chro(Count),All_Pop(Count));
        }
        Count = Count + 1;
    }
}
```

รูปที่ 6 กระบวนการค้นหาเฉพาะที่

Best_Chro จะทำการเลือกโครโมโซมที่ดีที่สุดในประชากรในแต่ละรุ่น ซึ่งในฟังก์ชัน P(Gen) คือประชากรในแต่ละรุ่น เมื่อทำการเลือกโครโมโซมที่ดีที่สุดก็จะเหลือโครโมโซมอื่นๆ นั่นคือ All_Pop จากนั้นจะกระทำการเปรียบเทียบในแต่ละมีมีหรือในแต่ละคำของตัวอักษร ถ้าเปรียบเทียบแล้วโครโมโซมที่ดีสุดมีคำที่ดีสุดก็จะไม่เปลี่ยนแปลง แต่ถ้าในแต่ละคำของโครโมโซมมีค่าต่ำกว่าคำอื่นของประชากรอื่นในรุ่นนั้นก็จะทำการแลกเปลี่ยนคำที่ดีที่สุดให้แก่โครโมโซมที่ดีที่สุด ซึ่งวิธีการนี้จะทำให้ได้โครโมโซมหรือคำตอบที่ดียิ่งขึ้นในการหาคำตอบเฉพาะที่



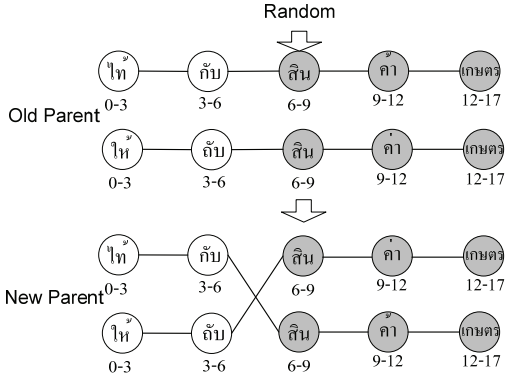
รูปที่ 7 การหาคำตอบที่ดีเฉพาะที่

(5) Parent selection method การเลือกประชากรในการดำเนินการทางพันธุศาสตร์หรือการเลือกคู่เพื่อที่จะแลกเปลี่ยนข้อมูลต่างๆ เพื่อหาคำตอบให้ดียิ่งขึ้น ซึ่งในงานวิจัยนี้จะใช้วิธีการวงล้อรูเล็ท (Roulette wheel) [13] โดยดำเนินการหาอัตราส่วนของประชากรต้นแบบทั้งหมด ถ้าประชากรต้นแบบใดมีค่าความเหมาะสมสูงในส่วนของวงล้อรูเล็ทก็จะมีพื้นที่กว้างทำให้โอกาสที่จะเลือกเป็นต้นแบบก็จะมีสูง แต่ถ้ามีค่าความเหมาะสมต่ำในส่วนของวงล้อรูเล็ทก็จะมีพื้นที่น้อยทำให้โอกาสที่จะเลือกเป็นต้นแบบก็จะมีน้อย เมื่อได้วงล้อรูเล็ทของประชากรต้นแบบทั้งหมดก็จะทำการสุ่มจับคู่โครโมโซมต้นแบบหรือโครโมโซมพ่อแม่เพื่อดำเนินการทางพันธุศาสตร์ต่อไป ซึ่งวิธีการนี้เปรียบเสมือนว่าคำตอบที่แข็งแรงมีโอกาสที่จะถูกคัดเลือกเพื่อเข้าสู่การดำเนินการทางพันธุกรรมมากกว่าตัวที่เป็นคำตอบไม่แข็งแรง

(6) Crossover method การครอสโอเวอร์หรือการแลกเปลี่ยนบางส่วนของโครโมโซม ซึ่งจุดประสงค์ของวิธีการนี้คือการสร้างประชากรรุ่นใหม่จากประชากรรุ่นเก่า ซึ่งก็จะทำให้ได้คำตอบรุ่นใหม่แทนคำตอบรุ่นเก่า ซึ่งในงานวิจัยนี้ได้ใช้การครอสโอเวอร์แบบ 1 จุด (Single Point Crossover) โดยจุดที่จะทำการครอสโอเวอร์จะถูกกำหนดโดยการสุ่ม ดังแสดงตัวอย่างในรูปที่ 8 และใช้จำนวนของการครอสโอเวอร์ดังสมการ

$$Number\ Crossover = \frac{P_c \times Population\ \ Size}{2} \qquad (4)$$

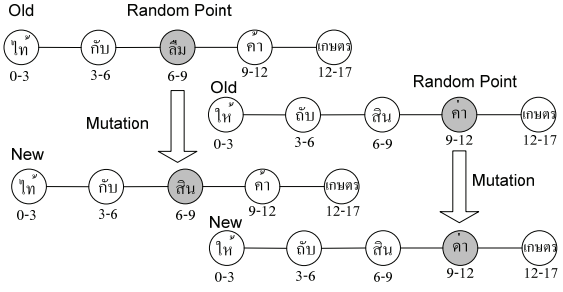
ซึ่ง Pc คือความน่าจะเป็นของการครอสโอเวอร์ (Crossover Probability)



รูปที่ 8 การครอสโอเวอร์

จากรูปเริ่มแรกจะเป็นการสุ่มตำแหน่งตัวอักษรจากรูปจะมีจำนวนตัวอักษรทั้งหมด 17 ตัว ดังนั้นจะทำการสุ่มตัวอักษร เช่นถ้าสุ่มได้เลข 8 แสดงว่าตำแหน่งที่จะทำการแลกเปลี่ยนคือตำแหน่งที่ 3 (ในกรณีที่ได้ตำแหน่งที่ 1 ให้ทำการสุ่มใหม่) ซึ่งจะทำการแลกเปลี่ยนบางส่วนของข้อมูลตั้งแต่ตำแหน่งที่สุ่มได้ไปจนถึงตำแหน่งสุดท้าย

(7) Mutation method การกลายพันธุ์ในงานวิจัยนี้คือการเปลี่ยนแปลงคำของประโยคในแต่ละตำแหน่งที่อยู่ในส่วนของประชากรต้นแบบทั้งหมด ซึ่งจะเป็นการเพิ่มประสิทธิภาพในการค้นหาคำตอบที่ดียิ่งขึ้น เป็นการเพิ่มความหลากหลายให้มากขึ้นในกลุ่มประชากรทำให้มีผลคำตอบครอบคลุมพื้นที่กว้างมากขึ้น ซึ่งในงานวิจัยนี้จำนวนการกลายพันธ์จะขึ้นอยู่กับความยาวของโครโมโซมสูงสุดของแต่ละรุ่น เช่นในกลุ่มของประชากรมีจำนวนความยาวของคำหรือมีมี สูงสุดคือ 9 ตัว ก็จะดำเนินการกลายพันธ์จำนวน 9 ตัว โดยจะทำการสุ่มทั้งในส่วนของโครโมโซมและในส่วนของของมีมี ทั้งหมด โดยข้อกำหนดในการกลายพันธ์ในแต่ละคำจะต้องอยู่ในตำแหน่งเดียวกัน ดังตัวอย่างแสดงในรูปที่ 9



รูปที่ 9 การกลายพันธ์

(8) Survive method ในการหาผู้รอดชีวิตหรือการหาคำตอบที่ดีเพื่อที่จะนำมาเป็นประชากรรุ่นใหม่ที่จะนำไปสู่การหาคำตอบที่ดีขึ้น ในงานวิจัยนี้จะใช้การเก็บรักษาประชากรที่ดีที่สุดในแต่ละรุ่นเพื่อรักษาคำตอบที่ดีที่สุดไว้ไม่ให้สูญหายเนื่องจากในบางครั้งเมื่อดำเนินการทางพันธุศาสตร์อาจจะไปในทิศทางที่ดีหรือไม่ดีก็ได้ แต่ถ้าใช้วิธีการนี้จะเป็นการหาคำตอบไปในทิศทางที่ดีขึ้นเรื่อยๆ เนื่องจากมีการเก็บประชากรที่ดีที่สุดไว้ โดยจะนำเอาคำตอบที่ดีที่สุดในประชากรรุ่นก่อนหน้ามาแทนคำตอบที่แย่ที่สุดในประชากรรุ่นใหม่

3. ผลการวิจัยและอภิปราย

ในการทดลองได้มีการทดสอบวิธีการที่นำเสนอในการหาผลลัพธ์ที่ดีที่สุดและได้มีการเปรียบเทียบกับวิธีการที่ใช้ในการค้นหาทั้งหมดเพื่อแสดงให้เห็นถึงประสิทธิภาพของวิธีการที่นำเสนอ อีกทั้งยังได้มีการทดสอบกับวิธีการเจเนติกอัลกอริทึม (Genetic algorithm) [13-15]

การทดลองอันดับแรกจะเป็นการทดลองในการตรวจแก้คำผิดภาษาไทยโดยจะมีการทดสอบตัวอักษรที่ผิดตั้งแต่ 10 ตัว ถึง 40 ตัว ในแต่ละข้อความทั้งหมดดังแสดงในตัวอย่างของผลการทดลองในตารางที่ 4 และตารางที่ 5

ตารางที่ 4 ตัวอย่างผลการทดสอบในการตรวจแก้คำผิดจำนวน 10 ตัวอักษร

ตัวอักษรผิด 10 ตัว	ผลลัพธ์จากการแก้ไข
ถาร**ห**ดสอบเป็นการ**แล**ดง	**ก**าร**ท**ดสอบเป็นการ**แส**ดง
ความถูกต้อง**บ**องวิธีการที่นำ	ความถูกต้อง**ข**องวิธีการที่
เส**ม**อกับวิธี**ภ**ารที่ถูกนำ**น**า	นำเส**น**อกับวิธี**ก**ารที่ถูกนำ**ม**า
เปรียบเทีย**น** ซึ่งแสดงให้เห็**ม**	เปรียบเทีย**บ** ซึ่งแสดงให้เห็**น**
ว่ามีคว**าน**ถูกต้องสูง	ว่ามีคว**าม**ถูกต้องสูง

ตารางที่ 5 ผลการทดสอบในการตรวจแก้คำผิด

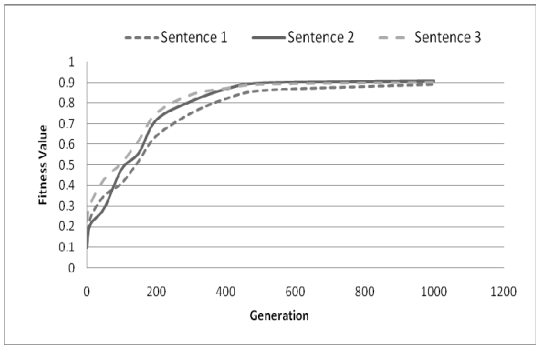
จำนวนตัวอักษรผิด	ความถูกต้อง %
10	98.0%
15	96.5%
20	94.2%
25	94.4%
30	93.6%
35	92.4%
40	92.5%

*ความถูกต้อง % ได้มาจากการเฉลี่ยของกลุ่มประโยคจำนวน 10 กลุ่ม ในแต่ละจำนวนตัวอักษรที่ผิด

จากการทดสอบในการตรวจแก้คำผิดแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถตรวจแก้คำผิดได้ถูกต้องมากกว่า 92% แต่อย่างไรก็ตามก็อาจจะมีส่วนที่ผิดพลาดเนื่องจากในส่วนของการเลือกตัวอักษรที่ใกล้เคียง แต่อย่างไรก็ตามก็นับว่าสามารถแก้คำผิดภาษาไทยได้มีประสิทธิภาพสูง

ในการทดลองถัดไปจะเป็นการทดลองในการหาคำตอบที่ดีที่สุดนั่นคือการหาคำตอบที่ดีที่สุดในการหาประโยคที่ถูกต้องโดยมีมีติกอัลกอริทึม ซึ่งการทดลองจะเป็นการทดสอบหาคำตอบที่ดีที่สุดจากจำนวนตัวอักษร 40 ตัวอักษร ซึ่งสามารถทำให้เกิดประโยคที่เป็นไปได้มากกว่าหมื่นประโยค โดยจะใช้วิธีการมีมีติกอัลกอริทึมหาประโยคที่ถูกต้องที่สุด ซึ่งจากผลการทดลอง 3 ประโยค ในการ

ทดสอบได้ใช้ P_c = 0.5 และจำนวนประชากรเริ่มต้น = 30 ดังแสดงผลลัพธ์ในรูปกราฟที่ 10



รูปที่ 10 การหาประโยคที่ถูกต้องโดยวิธีการที่นำเสนอ

จากการทดลองจะแสดงให้เห็นว่าคำตอบที่ได้จะดีขึ้นเรื่อยๆ โดยเฉพาะในรุ่นที่ 400 ขึ้นไป จะพบคำตอบที่ใกล้เคียงกับคำตอบที่ดีที่สุดจากวิธีการค้นหาทั้งหมด ดังนั้นในการทดสอบถัดไปจะกำหนดจำนวนรอบไว้ที่ 400 รอบ เพื่อทำการหาคำตอบที่ดีที่สุดในประโยคที่มีความยาวแตกต่างกันไป จากการทดสอบจำนวน 10 ครั้ง ในแต่ละจำนวนตัวอักษร ซึ่งจะแสดงผลลัพธ์ที่ดีที่สุดและแย่ที่สุดจากวิธีที่นำเสนอ ดังแสดงผลลัพธ์ในตารางที่ 6

ตารางที่ 6 ผลลัพธ์จากการเปลี่ยนแปลงจำนวนตัวอักษร

จำนวนตัวอักษร	ตัวอักษรที่เป็นไปได้	ค่าที่ดีที่สุด	วิธีการที่นำเสนอ	
			ดีสุด	แย่สุด
10	48	0.92	0.92	0.92
20	532	0.90	0.90	0.90
30	2,054	0.87	0.87	0.87
40	10,432	0.89	0.89	0.81
50	86,456	0.84	0.84	0.82
60	293,564	0.85	0.83	0.80
70	> 1 ล้าน	0.81	0.81	0.79
80	> 1 ล้าน	0.82	0.80	0.78

จากการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถหาคำตอบที่ดีที่สุดได้ใกล้เคียงกับคำตอบที่ถูกต้องที่สุดจากวิธีการค้นหาทั้งหมด (Full search) ในจำนวนตัวอักษรที่น้อยจะกระทำได้ดีทั้งในกรณีที่แย่สุดและดีสุด แต่ในกรณีที่จำนวนตัวอักษรเพิ่มมากขึ้น จำนวนของประโยคที่เป็นไปได้ก็จะมาก แต่อย่างไรก็ตามวิธีการที่นำเสนอก็สามารถหาคำตอบที่ใกล้เคียงกับคำตอบที่ถูกต้องได้ ซึ่งถ้าหากมีการกำหนดจำนวนรอบเพิ่มมากขึ้นก็จะมีความถูกต้องที่เพิ่มขึ้น

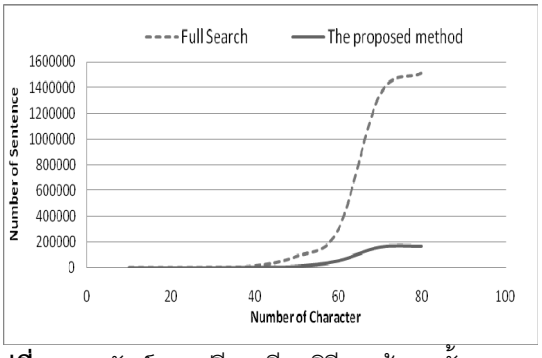
ในการทดสอบถัดไปเป็นการทดสอบการค้นหาคำตอบในแต่ละจำนวนตัวอักษร โดยจะหาคำตอบที่ดีที่สุดจากวิธีการที่นำเสนอโดยการเพิ่มจำนวนรอบไปจนกว่าจะเจอคำตอบที่ดีที่สุด ซึ่งในแต่ละจำนวนตัวอักษรจะทำการทดสอบจำนวน 10 ครั้ง จากนั้นจะได้จำนวนรอบที่ต่ำที่สุดและสูงสุดที่หาได้ ซึ่งจะนำไปคำนวณหาค่าเฉลี่ย ดังแสดงผลลัพธ์ในตารางที่ 7

ตารางที่ 7 ผลลัพธ์ของจำนวนรอบในการหาคำตอบที่ดีที่สุด

จำนวนตัวอักษร	จำนวนรอบที่เจอคำตอบที่ดีที่สุด		ค่าเฉลี่ยจำนวนรอบ	อัตราการค้นหา
	ต่ำสุด	สูงสุด		
10	1	7	4	100%
20	1	13	8	45%
30	10	24	18	26%
40	23	412	176	53%
50	45	1034	343	11%
60	37	4327	1,764	18%
70	84	7832	5,367	11%
80	89	8632	5,612	11%

อัตราการค้นหาคือจำนวนของค่าเฉลี่ยคูณกับจำนวนประชากรที่กำหนดและหารด้วยจำนวนประโยคทั้งหมดที่เป็นไปได้ แล้วทำให้เป็นเปอร์เซ็นต์ ซึ่งแสดงให้เห็นว่าวิธีการที่นำเสนอไม่จำเป็นที่จะต้องทำการค้นหาทั้งหมดซึ่งทำให้เสียเวลาในการค้นหา ยิ่งในกรณีที่มีเส้นทางที่เป็นไปได้เป็นล้านเส้นทาง วิธีการที่นำเสนอสามารถค้นหาเพียงช่วงเวลาหนึ่งก็สามารถหาคำตอบที่ถูกต้องได้ ซึ่งจากการทดลองแสดงให้เห็นว่าคำตอบที่ถูกต้องขึ้นอยู่กับจำนวนรอบ ยิ่งจำนวนรอบเพิ่มมากขึ้นคำตอบที่ได้ก็จะถูกต้องมากยิ่งขึ้น ถึงจะมีจำนวนรอบมาก แต่อย่างไรก็ตามการค้นหาที่ได้จะทำการค้นหาเฉพาะบางส่วนของประโยคที่เป็นไปได้ทั้งหมด ดังเช่นอาจจะหาเพียง 10% ของกลุ่มข้อมูลทั้งหมด ซึ่งจากวิธีการที่นำเสนอก็จะได้ผลลัพธ์เช่นเดียวกันกับการหา 100% ของวิธีการค้นหาทั้งหมด ซึ่งจากการทดลองแสดงให้เห็นว่ายิ่งมีจำนวนตัวอักษรมาก ก็จะมากทำให้การค้นหาลำบากซึ่ง

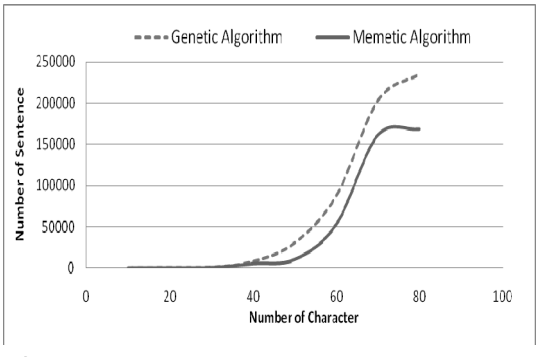
อาจจะต้องหาเป็นล้านประโยคแต่วิธีการที่นำเสนอจะสามารถหาคำตอบที่ถูกต้องได้เพียงไม่เกิน 15% ของทั้งหมด ดังแสดงการเปรียบเทียบในการหาประโยคทั้งหมดและวิธีการที่นำเสนอ ในรูปกราฟที่ 11



รูปที่ 11 ผลลัพธ์การเปรียบเทียบวิธีการค้นหาทั้งหมดและวิธีการที่นำเสนอ

จากกราฟแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถหาประโยคที่ถูกต้องโดยจะหาเพียงช่วงหนึ่งของข้อมูลทั้งหมด ยิ่งในกรณีที่มีจำนวนตัวอักษรมาก ก็จะเห็นผลประสิทธิภาพที่แตกต่างกันชัดเจน

ในการทดลองถัดไปจะเป็นการทดลองโดยเปรียบเทียบกับวิธีการของเจเนติกอัลกอริทึม ซึ่งวิธีการนี้จะเป็นวิธีการค้นหาคำตอบที่มีประสิทธิภาพอีกวิธีการหนึ่ง โดยใช้รูปแบบเช่นเดียวกับวิธีการที่นำเสนอเพียงแต่ไม่นำเอาวิธีการค้นหาเฉพาะที่มาใช้ในกระบวนการ ซึ่งจากการทดลองสรุปได้ดังรูปที่ 12 ซึ่งเป็นการเปรียบเทียบในการค้นหาประโยคที่ถูกต้องในด้านจำนวนของประโยคที่ทำการค้นหา โดยจะดูว่าวิธีการใดใช้จำนวนประโยคที่ทำการค้นหาน้อยกว่ากันแล้วเจอคำตอบที่ดีที่สุด



รูปที่ 12 การเปรียบเทียบระหว่างมีมีติกอัลกอริทึมและเจเนติกอัลกอริทึมในการค้นหาประโยคที่ถูกต้อง

จากการทดลองในการเปรียบเทียบประสิทธิภาพระหว่างวิธีการของมีมีติกอัลกอริทึมและเจเนติกอัลกอริทึม แสดงให้เห็นว่าวิธีการที่นำเสนอสามารถหาคำตอบที่ถูกต้องได้เร็วมากกว่า ซึ่งแสดงให้เห็นจากจำนวนการค้นหาที่แสดงถึงปริมาณของการค้นหาประโยคที่ถูกต้องน้อยกว่า

4. สรุป

งานวิจัยได้มีการปรับปรุงประสิทธิภาพในการตรวจแก้คำผิดภาษาไทยโดยใช้มีมีติกอริทึม ซึ่งจากผลลัพธ์จากการใช้วิธีการส่งผ่านโทเค็นจะส่งผลให้ได้คำที่ถูกต้องที่เป็นไปได้จำนวนมาก ดังนั้นในงานวิจัยนี้จึงนำกระบวนการมีมีติกอัลกอริทึมมาช่วยในการค้นหาประโยคที่ถูกต้องจากคำที่เป็นไปได้ทั้งหมดโดยเปรียบเทียบความถูกต้องจากรูปแบบจำลองทางภาษา จากผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถตรวจแก้ไขคำผิดได้อย่างถูกต้องในกรณีที่มีจำนวนตัวอักษรจำนวนมากแสดงให้เห็นถึงประสิทธิภาพในการค้นหาว่ามีประสิทธิภาพสูง อีกทั้งในงานวิจัยได้มีการเปรียบเทียบกับวิธีการของเจเนติกอัลกอริทึมและวิธีการค้นหาทั้งหมด ซึ่งแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถค้นหาประโยคที่ดีที่สุดอย่างมีประสิทธิภาพมากกว่าวิธีที่นำมาเปรียบเทียบ ดังนั้นในงานวิจัยนี้จะเป็นวิธีการตรวจแก้คำผิดโดยอัตโนมัติวิธีการหนึ่งที่สามารถตรวจแก้คำผิดได้อย่างถูกต้อง ทำให้สามารถนำไปสร้างเครื่องมือช่วยเหลือมนุษย์ได้อย่างมีประสิทธิภาพ

5. เอกสารอ้างอิง

[1] Kukich K. Technique for automatically correcting words in text. ACM Computing Surveys;1992;24(4): p.517-522.

[2] Na-Rae H, Martin C, Claudia L. Detecting error in English article usage by non-native speakers. Natural Language Engineering; 2006; 12(2):115-129.

[3] Clarson P, Rosenfeld R. Statistical language model using the CMU CAMBRIDGE toolkit. Available from: URL:http://www.eng.cam.ac.uk/~prc14/toolkit.html.

[4] Jianhua LI, Xiaolong W. Combining trigram and automatic weight distribution in Chinese spelling error correction. Journal of Computer Science and Technology; 2002; 17(6): p.915-923.

[5] Golding AR. A winnow base approach to context sensitive spelling correction. Boston. 1999. p.107-130.

[6] Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech and Signal Processing; 1987; 35(3): p.400-410.

[7] Rosenfeld R. Adaptive statistical language modeling: A maximum entropy approach. School of Computer Science. Carnegie Mellon University. 1994.

[8] Young SJ. Token passing: a simple conceptual model for connected speech recognitions system. Cambridge University Engineering Department. 1989.

[9] Feller W. An introduction to probability theory and its application. New York. 1968.

[10] Fogel DB. Evolutionary computation: toward a new philosophy of machine intelligence. IEEE Press. Piscataway. 1995.

[11] Alkan A, Ozcan E. Memetic algorithms for timetabling. Yeditepe University. 2003.

[12] Duan H, Yu X. Hybrid ant colony optimization using memetic algorithm for traveling salesman problem. Proceedings of the IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learnng. 2007. p. 92-95.

[13] Holland JH. Genetic Algorithm and the optimal allocation of trials. SIAM Journal of Computing; 1973; 2(2): p.88-105.

[14] Goldberg DE. Genetic algorithm in search, optimization, and machine learning. Addison Wesley. 1989.

[15] Phuk-in A. Method production scheduling using a comparison of genetic algorithm and other general methods. KKU Engineering Journal. 2012;39(1): 35-46. (In Thai).

[16] Somkantha K. Thai error correction system by spell checking and language model. The National Conference on Computing and Information Technology; 2005 May 24-25; Bangkok, Thailand. 2005. p.36-41. (In Thai).