บทที่ 3

การหาขอบเขตของชื่อเฉพาะ

ในบทนี้จะกล่าวถึงการทำงานส่วนแรกของระบบ คือการใช้วิธีการทางสถิติเพื่อหาขอบเขตของชื่อเฉพาะ หรือคือการหารายการกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ และเนื่องจากวิธีการทางสถิติที่ใช้ เป็นการคำนวณความสัมพันธ์ระหว่างสองหน่วยหรือสองพยางค์ ในตอนแรกของบทนี้จึงจะกล่าวถึงการปรับวิธีการทางสถิติเพื่อใช้คำนวณความสัมพันธ์ระหว่างหลายพยางค์ได้ ซึ่งจะกล่าวถึงวิธีทางสถิติที่นำมาปรับเพื่อใช้เลือกกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) และนำเสนอผลวิธีการทางสถิติที่เหมาะสมที่สุดในการดึงกลุ่มพยางค์ที่น่าจะเป็นชื่อเฉพาะออกมา และในสองส่วนสุดท้ายจะกล่าวถึงผลและการอภิปรายผลของวิธีทางสถิติที่ใช้ในงานวิจัยนี้

3.1 วิธีทางสถิติที่ประยุกต์ใช้

จากวิธีการทางสถิติที่กล่าวไปในบทที่แล้ว จะพบว่าวิธีการใช้ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) และค่า Dunning's Log Likelihood นั้น ปกติจะใช้คำนวณหาค่าความสัมพันธ์ของ 2 คำ ซึ่งชื่อเฉพาะหลายชื่อมักมีความยาวมากกว่า 2 คำ อีกทั้งชื่อสถานที่หรือชื่อองค์กรก็มักมีความยาวมากกว่า 2 คำขึ้นไป ทำให้วิธีการทางสถิติเหล่านี้มีข้อจำกัด ซึ่งอาจแก้ไขโดยใช้วิธีการขยายหน้าต่าง (Lopes และคณะ ,1999) กล่าวคือ เมื่อหาค่าความสัมพันธ์ระหว่างคำ 2 คำได้แล้ว ก็จะมอง 2 คำนั้นให้เป็นหน่วยเดียวกัน แล้วมองคำถัดไปให้เป็นอีกหน่วยหนึ่ง แล้วหาความสัมพันธ์ระหว่างหน่วยนั้นๆ กับคำถัดไป จากนั้น หน้าต่างก็จะขยายต่อไปเรื่อยๆ หลักการนี้เป็นการมองเอ็นแกรมใดๆ ในรูปของไบแกรมเทียม (pseudo-bigram) เช่น มองเอ็นแกรม (w_1, w_2, ...w_m) เป็นไบแกรมของ w_1, w_2, ... w_i กับ w_{i+1} ... w_n เป็นต้น ดังนั้น เมื่อต้องการคำนวณค่าอย่างเช่นค่า Mutual Information (MI) ซึ่งปกติจะเท่ากับ log P(x,y)/P(x)P(y) ก็จะคำนวณเป็น log P(w_1, w_2,... w_n)/P(w_1, w_2,... w_i)P(w_{i+1},... w_n) และเนื่องจากเราสามารถแบ่งครึ่งเอ็นแกรมใดๆ ที่ตำแหน่งระหว่าง w_1- w_2, w_2- w_3, ... หรือ w_{n-1}- w_n ก็ได้ (Lopes และคณะ ,1999) จึงต้องหาค่าเฉลี่ยของไบแกรมเทียมทั้ง n-1 แบบนี้

ในงานวิจัยนี้ ผู้วิจัยก็ใช้หลักการเดียวกันนี้ เพียงแต่เปลี่ยนจากการมองความสัมพันธ์ระหว่างคำเป็นความสัมพันธ์ระหว่างพยางค์แทน ซึ่งในงานวิจัยนี้ การปรับขยายหน้าต่างจะขยายถึง 8-gram เพื่อให้ครอบคลุมชื่อเฉพาะที่มีความยาวมากๆ

ซึ่งจากการปรับสมการจะทำให้ได้สมการของการหาค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) และค่า Dunning's Log Likelihood ออกมาเป็นดังนี้

- การขยายการคำนวณแบบ pseudo bigram เพื่อหาค่า Mutual Information

จากสมการ $I(w_1, w_2) = \log \frac{p(w_1 w_2)}{p(w_1)\,p(w_2)}$

เมื่อแก้สมการ ทำการขยายหน้าต่างจะได้

$$I(w_1 \ldots w_n) = \log \frac{p(w_1 \ldots w_n)}{Avp}$$

โดยที่ $Avp = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} p(w_1 \ldots w_i) * p(w_{i+1} \ldots w_n)$

- การขยายการคำนวณแบบ pseudo bigram เพื่อหาค่า Pearson's Chi-square

จากสมการ $\chi^2 = \frac{\sum_{i,j}(O_{i,j} - E_{i,j})^2}{E_{i,j}}$

โดยที่ $O_{i,j}$ คือ ค่าความถี่จากการสังเกตของค่า 4 ค่า ดังตาราง

O_{11} = ความถี่ของการปรากฏร่วมระหว่างคำ w_1 และ w_2	O_{12} = ความถี่ของการเกิดระหว่างคำที่ไม่ใช่ w_1 ที่ปรากฏร่วมกับ w_2
O_{21} = ความถี่ของการเกิดระหว่างคำ w_1 แต่ไม่ได้ตามด้วยคำ w_2	O_{22} = ความถี่ของการปรากฏร่วมระหว่างคำที่ไม่ใช่ทั้ง w_1 และ w_2

ส่วนค่า $E_{i,j}$ คือค่าความถี่ของแต่ละเซลล์ในตารางเมื่อคำสองคำเกิดร่วมกันโดยบังเอิญ ซึ่งค่า $E_{i,j}$ ของแต่ละเซลล์ในตารางจะหาได้จาก ผลรวมของแต่ละแถวคูณกับผลรวมของแต่ละสดมภ์หารด้วยผลรวมของค่าความถี่ทั้งหมด

เมื่อทำการแทนค่า $O_{i,j}$และ $E_{i,j}$ แล้ว สมการจะได้ดังนี้

$$\chi^2 = \frac{[f(w_1, w_2) * N - (f(w_1) * f(w_2))]^2}{f(w_1) * f(w_2) * (N - f(w_1)) * (N - f(w_2))}$$

โดย $f(w_1)$, $f(w_2)$ = ค่าความถี่ของคำ (w_1) และ (w_2) ตามลำดับ

N = จำนวนคำทั้งหมดในคลังข้อมูล

เมื่อคำนวณด้วยการขยายการคำนวณแบบ pseudo bigram จะได้

$$\chi^2(w_1 \ldots w_n) = \frac{[f(w_1 \ldots w_n) * N - Avp]^2}{Avp * (N - Avx) * (N - Avy)}$$

โดยที่ $Avx = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} f(w_1 ... w_i)$

$$Avy = \frac{1}{(n-1)} * \sum_{i=2}^{i=n} f(w_i ... w_n)$$

$$Avp = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} p(w_1 ... w_i) * p(w_{i+1} ... w_n)$$

- การขยายการคำนวณแบบ pseudo bigram เพื่อหาค่า Cubic association ratio (MI^3)

จากสมการ

$$MI^3 = log_2 \frac{a^3 N}{(a+b)(a+c)}$$

จากสูตร *a* คือ ค่าความถี่ของ bigram ของคำที่1 และ 2 ($w_1 - w_2$)

b คือ ค่าความถี่ของ bigram ของคำที่ไม่ใช่คำที่ 1 และ คำที่ 2 ($\sim w_1 - w_2$)

c คือ ค่าความถี่ของ bigram ของคำที่1 และ คำที่ไม่ใช่คำที่ 2 ($w_1 - \sim w_2$)

เมื่อแก้สมการจะได้

$$MI^3 = log_2 \frac{(f(w_1, w_2))^3 * N}{[f(w_1, w_2) + (f(w_2) - f(w_1, w_2))] * [f(w_1, w_2) + (f(w_1) - f(w_1, w_2))]} \quad \text{------z}$$

จากสมการ z จะได้เป็น

$$MI^3 = log_2 \frac{(f(w_1, w_2))^3 * N}{f(w_2) * f(w_1)}$$

ดังนั้น เมื่อทำการขยายการคำนวณแบบ pseudo bigram จะได้

$$MI^3 (w_1 ... w_n) = log_2 \frac{(f(w_1 ... w_n))^3 * N}{Avy * Avx}$$

โดยที่ $Avx = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} f(w_1 ... w_i)$

$$Avy = \frac{1}{(n-1)} * \sum_{i=2}^{i=n} f(w_i ... w_n)$$

- การขยายการคำนวณแบบ pseudo bigram เพื่อหาค่า Dunning's log likelihood

จากสมการ

$$Loglike (w_1, w_2) = 2 * (log\ l(p1,k1,n1) + log\ l(p2,k2,n2) - log\ l(p,k1,n1) - log\ l(p,k2,n2))$$

เมื่อ $log\ l(P,K,M) = K * ln(P) + (M - K) * ln(1 - P)$

โดยที่ *k1* คือ ความถี่ของ bigram w_1-w_2 = $f(w_1, w_2)$

$k2 = f(w_1) - k1$

$n1 = f(w_2)$

$n2 = N - n1$

$p1 = k1/n1 = f(w_1, w_2) / f(w_2)$

$p2 = k2/n2 = (f(w_1) - f(w_1, w_2)) / (N - f(w_2))$

$p = (k1 + k2)/N = f(w_1)/N$

N คือ จำนวนคำทั้งหมดในคลังข้อมูล

เมื่อทำการขยายการคำนวณแบบ pseudo bigram จะได้

$Loglike\ (w_1...w_n) = 2 * (log\ l(pf1,kf1,nf1) + log\ l(pf2,kf2,nf2) - log\ l(pf,kf1,nf1) - log\ l(pf,kf2,nf2))$

โดยที่ $kf1 = f\ (w_1...w_n)$

$kf2 = Avx - kf1$

$pf = (kf1 + kf2)/N = Avx$

$nf1 = Avy$

$nf2 = N- nf1$

$pf1 = kf1/nf1$

$pf2 = kf2/nf2$

$$Avx = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} f(w_1...w_i)$$

$$Avy = \frac{1}{(n-1)} * \sum_{i=2}^{i=n} f(w_i...w_n)$$

3.2 การกำหนดขอบเขตของชื่อเฉพาะ

ข้อมูลใดใดที่ส่งเข้ามาจะถูกแยกพยางค์และมองเป็น n-gram ขนาดต่างๆ พร้อมทั้งคำนวณค่าทางสถิติเพื่อบอกความสัมพันธ์ภายใน n-gram นั้น เมื่อได้ค่าความสัมพันธ์ของ n-gram จากแต่ละวิธีที่กล่าวไปข้างต้นได้แก่ ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) และค่า Dunning's Log Likelihood (ที่ผ่านการขยายหน้าต่างเพื่อให้หาความสัมพันธ์ระหว่าง n-grams ได้กว้างขึ้นแล้ว) รวมทั้งค่า Mutual Expectation (ตามที่กล่าวถึงแล้วในบทที่ 2) ก็จะมาทำการเลือก n-gram หรือกลุ่มพยางค์ที่อาจจะเป็นชื่อเฉพาะ (candidate) ซึ่งโดยทั่วไปอาจใช้ระดับหรือ threshold ในการตัดสินว่าขอบเขตของตำแหน่งใดที่น่าจะเป็นจุดเริ่มต้นและจุดสิ้นสุดของชื่อเฉพาะ แล้วทำการคัดให้เป็น กลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) แต่วิธีการนี้จะมีข้อเสียตรงที่ ระดับที่ใช้คัดกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) นั้นจะได้จากคลังข้อมูลใช้ฝึก (training corpus) ซึ่งทำให้ระดับจะขึ้นอยู่กับขนาดของ

คลังข้อมูลใช้ฝึก (training corpus) ด้วย ดังนั้น จึงมีอีกแนวคิดหนึ่งที่ช่วยแก้ปัญหานี้ นั่นก็คือ การใช้ LocalMaxs algorithm ในการเลือก n-gram ที่มีแนวโน้มที่จะเป็นหน่วยเดียวกัน

การใช้ LocalMaxs algorithm เป็นวิธีการที่ Dias และคณะ (2000) นำมาใช้เพื่อระบุตำแหน่งเริ่มต้นและสิ้นสุดของศัพท์เฉพาะทาง (term) ซึ่งมักจะประกอบด้วยคำย่อยๆ หลายคำซึ่ง Localmax algorithm จะเลือกตัดสินว่ากลุ่มคำนั้นเป็นศัพท์เฉพาะทางหรือไม่ โดยพิจารณาจากคุณสมบัติ 2 ข้อ

1. ค่าที่ได้จากการวัดค่าความสัมพันธ์ยิ่งมีค่ามากจะบอกถึงความสัมพันธ์ระหว่างกลุ่มคำที่มีมากตามไปด้วย
2. ศัพท์เฉพาะทางมักเป็นกลุ่มคำที่ปรากฏร่วมกันและมีตำแหน่งที่แน่นอน

ดังนั้นเราจะอนุมานเอ็นแกรมของคำนี้ให้เป็นศัพท์เดียวกันก็ต่อเมื่อค่าที่ได้จากการวัดความสัมพันธ์สูงกว่าหรือเท่ากันกับค่าที่ได้จากการวัดความสัมพันธ์ของส่วนย่อย (sub-group) ทั้งหมดภายใน (n-1) gram และก็สูงกว่าค่าที่ได้จากการวัดความสัมพันธ์ของ (n+1) gram ด้วย

เมื่อกำหนดให้ *assoc* เป็นค่าที่ได้จากการวัดความสัมพันธ์

W เป็น n-gram

Ω_{n-1} เป็นเซตของ (n-1) gram ทุกตัวที่อยู่ใน *W*

Ω_{n+1} เป็นเซตของ (n+1) gram ทุกตัวที่อยู่ใน *W*

sizeof จะบอกจำนวนของคำใน n-gram

ซึ่ง Localmax algorithm จะนิยามได้เป็น ดังนี้

$\forall x \in \Omega_{n-1}, \forall y \in \Omega_{n+1}$,

W จะเป็น term เดียวกันก็ต่อเมื่อ

$$(sizeof(W) = 2 \wedge assoc(W) > assoc(y))$$

$$\vee$$

$$(sizeof(W) \neq 2 \wedge assoc(W) \geq assoc(x) \wedge assoc(W) > assoc(y))$$

ตัวอย่างการใช้ Localmax algorithm จะเลือกศัพท์โดยพิจารณาบริบทที่อยู่ข้างเคียงแล้วเลือกกลุ่มคำที่มีความสัมพันธ์ระหว่างกันสูงๆ เช่น เมื่อพิจารณา "Operating System" และ "Operating System Windows" จะพบว่า "Operating System" มีความสัมพันธ์ระหว่างกันสูงกว่าค่าความสัมพันธ์ระหว่าง "Operating System Windows" ซึ่งเป็นเพราะว่ามีคำอื่นที่มีความสัมพันธ์กับ "Operating System" มากกว่า "Windows"

ดังนั้น Localmax algorithm ก็จะเลือกให้ "Operating System" ว่ามีโอกาสที่จะเป็นศัพท์ได้มากกว่า"Operating System Windows"

แต่ในวิทยานิพนธ์นี้ ผู้วิจัยได้นำ Localmax algorithm มาประยุกต์ใช้ในการหาขอบเขตของชื่อเฉพาะ โดยมองหากลุ่มพยางค์ที่มีค่าความสัมพันธ์ภายในสูงกว่าค่าความสัมพันธ์ภายในของกลุ่มพยางค์ที่ถูกลดจำนวนลงหรือถูกเพิ่มจำนวนขึ้น โดยคำนวณค่าความสัมพันธ์ภายในด้วยวิธีการสถิติแบบต่างๆ อันได้แก่ ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation ในที่นี้จะเรียกกลุ่มพยางค์ที่ถูกเลือกมานี้ว่า "ชื่อเฉพาะที่เลือกมา" เมื่อได้ชื่อเฉพาะที่เลือกมาจากวิธีทางสถิติเหล่านี้แล้ว จากนั้นจะมีการตัดสินว่าชื่อเฉพาะที่เลือกมาจากวิธีทางสถิติวิธีไหนที่ให้ผลดีที่สุด โดยจะตัดสินจาก อัตราการรู้จำ (recognition rate) และ เวลาที่ใช้ในการรู้จำ (recognition time) โดยอัตราการรู้จำในที่นี้ คือค่า F ซึ่งค่า F สามารถหาได้จาก ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall) ซึ่งจะกล่าวถึงในตอนต่อไป

เมื่อเปรียบเทียบค่าอัตราการรู้จำ (recognition rate) และ เวลาที่ใช้ในการรู้จำ (recognition time) แล้ว จะทำให้ได้คำตอบว่าวิธีการสถิติแบบไหนที่เลือกชื่อเฉพาะได้ดีที่สุด

3.3 เปรียบเทียบวิธีทางสถิติที่ใช้ในการหาขอบเขตชื่อเฉพาะ

การทดสอบจะทำโดยการนำรายการชื่อเฉพาะที่เลือกมาด้วยวิธีทางสถิติ 5 วิธีได้แก่ ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation มาตรวจสอบว่าเป็นชื่อเฉพาะจริงหรือไม่ ทั้งนี้เพื่อจะหาคำตอบว่าวิธีการสถิติแบบใดที่เหมาะสมที่สุดที่จะใช้ร่วมกับ Localmax algorithm ในการหาขอบเขตของชื่อเฉพาะ ซึ่งชื่อเฉพาะที่เลือกมานี้ ในงานวิจัยนี้จะจำกัดขนาดมากที่สุด 8 พยางค์ ชื่อเฉพาะที่เลือกมาจึงมีตั้งแต่ 2-8 พยางค์

การตรวจสอบจะเปรียบเทียบรายการของชื่อเฉพาะที่เลือกมากับรายการของชื่อเฉพาะที่ถูกต้อง โดยจะเปรียบเทียบทั้งตำแหน่งที่เป็นจุดเริ่มและจำนวนพยางค์ด้วย ซึ่งถ้าหากตำแหน่งที่เป็นจุดเริ่มต้นของชื่อเฉพาะที่ถูกต้องและจุดเริ่มต้นของชื่อเฉพาะที่เลือกมาไม่ตรงกัน ชื่อเฉพาะที่เลือกมานั้นจะถูกคัดทิ้ง และถ้าจำนวนพยางค์ของคำในชื่อเฉพาะที่เลือกมามีมากหรือน้อยกว่าคำจากรายการชื่อเฉพาะที่ถูกต้อง คำนั้นๆ ก็จะถูกคัดทิ้งเช่นกัน

ตัวอย่าง ถ้าชื่อเฉพาะที่เลือกมามี 2 พยางค์ เช่น "สิ – ริ" แต่คำจากรายการชื่อเฉพาะที่ถูกต้องคือ "สิ-ริ-กร" ดังนั้น ชื่อเฉพาะที่เลือกมาตัวนี้ก็จะถูกคัดทิ้งไป

ส่วนในอีกกรณีหนึ่ง ที่ชื่อเฉพาะที่เลือกมามีส่วนของคำนำหน้าชื่อติดมาด้วย เช่น "นาง-สิ-ริ-กร" ชื่อเฉพาะที่เลือกมานี้ก็จะถูกคัดทิ้งไปด้วย

จากการตรวจสอบรายการของชื่อเฉพาะที่เลือกมาจากวิธีทางสถิติทุกแบบแล้ว จะได้รายการของชื่อเฉพาะที่เลือกมาและตรงกับรายการชื่อเฉพาะที่ถูกต้องออกมา ซึ่งจำนวนของรายการของชื่อ

เฉพาะที่เลือกมาจากวิธีทางสถิติแต่ละวิธีนั้น จะถูกนำมาตัดสินจาก อัตราการรู้จำ (recognition rate) และ เวลาที่ใช้ในการรู้จำ (recognition time) โดยอัตราการรู้จำในที่นี้ คือค่า F ซึ่งค่า F สามารถหาได้จาก ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall)

ซึ่งค่าเหล่านี้จะหาได้จากสูตร

ค่าความแม่นยำ(P) = (จำนวนชื่อเฉพาะที่เลือกมาแล้วถูกต้อง *100)/ จำนวนชื่อเฉพาะที่เลือกออกมาทั้งหมด

ค่าความครบถ้วน (R) = (จำนวนชื่อเฉพาะที่เลือกมาแล้วถูกต้อง *100) / จำนวนชื่อเฉพาะจริงทั้งหมดในข้อมูล

อัตราการรู้จำ หรือ ค่า F คือค่าที่เฉลี่ยให้ความสำคัญกับค่าความแม่นยำและความครบถ้วนเท่าๆกัน = (2 *P *R) / (P+R)

เวลาที่ใช้ในการรู้จำ (recognition time) สามารถคำนวณได้จากจำนวนชื่อเฉพาะที่เลือกออกมาได้ต่อ 1 นาที

ซึ่งผลการทดสอบรายการของชื่อเฉพาะที่เลือกมาจากวิธีทางสถิติอันได้แก่การหาค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation ร่วมกับการใช้ Localmax algorithm นั้น เมื่อคำนวณค่าความแม่นยำ (Precision) , ค่าความครบถ้วน (Recall) และค่า F (อัตราการรู้จำ) แล้วจะได้ผลดังแสดงอยู่ในตารางที่ 2 โดย จำนวน candidate ทั้งหมด จะเป็น จำนวนรายการของชื่อเฉพาะที่เลือกมาด้วยวิธีทางสถิติแต่ละแบบ

ตารางที่ 2 ผลการทดสอบค่าความแม่นยำ ค่าความครบถ้วนและอัตราการรู้จำที่ได้จากวิธีทางสถิติ 5 วิธี

วิธีทางสถิติ	ชื่อเฉพาะที่เลือกมาแล้วถูกต้อง	ชื่อเฉพาะที่เลือกมาทั้งหมด	ชื่อเฉพาะที่ถูกต้องทั้งหมด	ค่าความแม่นยำ (%)	ค่าความครบถ้วน (%)	ค่า F
Mutual Information	8746	534999	11683	1.635	74.861	3.200
Pearson's Chi-square	5800	542241	11683	1.070	49.645	2.095
Cubic Association Ratio	7406	735421	11683	1.007	63.391	1.983
Dunning's Log Likelihood	7396	738974	11683	1.001	63.306	1.971
Mutual Expectation	698	23270	988	3.000	70.648	5.756

จากตารางที่ 2 จะสังเกตได้ว่า วิธี Mutual Expectation ให้ผลออกมาต่างจากวิธีอื่นๆ โดยจำนวนชื่อเฉพาะทั้งหมดมีแค่เพียง 988 ชื่อเท่านั้น ทั้งนี้เป็นเพราะในการประมวลผลของระบบที่ใช้วิธีทางสถิติวิธีนี้ ใช้เวลานานมากกว่าจะประมวลผลสำเร็จ ดังนั้นจึงมีการตัดทอนข้อมูลลงเพื่อให้การประมวลผลเป็นไปได้ ซึ่งคลังข้อมูลใช้ฝึก (training corpus) ที่ใช้ประมวลผลกับระบบนี้นั้นจึงมีประมาณ 10% ของคลังข้อมูลใช้ฝึกทั้งหมดที่ใช้กับวิธีทางสถิติวิธีอื่น

ตารางที่ 3 ผลการทดสอบเวลาที่ใช้ในการรู้จำที่ได้จากวิธีทางสถิติ 5 วิธี

วิธีทางสถิติ	เวลาที่ใช้ (นาที)	ชื่อเฉพาะที่เลือกมาทั้งหมด	เวลาที่ใช้ในการรู้จำ (จำนวนชื่อ/นาที)
Mutual Information	1	534999	53499.9
Pearson's Chi-square	1	542241	54224.1
Cubic Association Ratio	1	735421	73542.1
Dunning's Log Likelihood	1	738974	73897.4
Mutual Expectation	1	23270	0.86

และจากตารางที่ 2 ผลการวัดค่าความแม่นยำ ค่าความครบถ้วน และ คำนวณหาค่าอัตราการรู้จำ (ค่า F) ดังกล่าว จะพบว่า วิธี Mutual Expectation ให้ค่าอัตราการรู้จำ (ค่า F) ออกมามากที่สุด ซึ่งสามารถสรุปได้ว่าจากวิธีทางสถิติทั้ง 5 วิธี อันได้แก่วิธี Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation นั้น วิธีที่ใช้ค่า Mutual Expectation สามารถดึงเอาชื่อเฉพาะออกมาแล้วมีความถูกต้องมากที่สุด แต่จากตารางที่ 3 พบว่าวิธีที่ใช้ค่า Mutual Expectation นี้ใช้เวลาในการประมวลผลนานมากกว่าการใช้วิธีการสถิติแบบอื่นๆ ทำให้เวลาที่ใช้ในการรู้จำหรือ recognition time นั้นมีค่าน้อยที่สุด โดยที่วิธีนี้ใช้เวลาในการดึงกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะด้วยความเร็ว 0.86 ชื่อ/ 1 นาที ในขณะที่เมื่อใช้วิธี Mutual Information จะดึงกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะด้วยความเร็ว 53499.9 ชื่อ/ 1 นาที ดังนั้น เมื่อเปรียบเทียบอัตราในการรู้จำแล้ว ผู้วิจัยจึงเลือกใช้ค่าทางสถิติที่ได้มาจากวิธี Mutual Information แทน ซึ่งวิธีนี้เป็นวิธีที่ให้ผลการทดสอบออกมาดีเป็นที่สุดเมื่อเทียบกับอีก 3 วิธีที่เหลือ ดังนั้น วิธี Mutual Information จึงเป็นวิธีทางสถิติที่เหมาะสมที่สุดที่จะนำมาใช้ในการคำนวณค่าความสัมพันธ์ภายในของกลุ่มพยางค์ โดยจะนำมาใช้ร่วมกับ Localmax algorithm ในการหาขอบเขตของชื่อเฉพาะในงานวิจัยนี้

3.4 อภิปรายผล

จากตารางที่ 2 การที่ค่าความแม่นยำของผลที่ได้มาจากวิธีทางสถิติต่างๆ เหล่านี้มีค่าต่ำมาก อาจจะเป็นผลมาจากการที่คลังข้อมูลที่ใช้ในงานนี้เป็นคลังข้อมูลที่ผ่านการแยกพยางค์จึงทำให้จำนวนชื่อเฉพาะที่เลือกออกมาได้มีปริมาณมาก และส่งผลให้ค่าความแม่นยำมีค่าต่ำมากด้วย ซึ่งทั้งนี้อาจเป็นเพราะจากการที่สาย n-grams ซึ่งเป็นพยางค์ต่อๆ กันนี้เข้าสู่โปรแกรมที่ใช้วิธีทางสถิติต่างๆ อันได้แก่ วิธี Mutual Information , วิธี Dunning's Log Likelihood , วิธี Pearson's Chi-square , วิธี Cubic association ratio (MI^3) และ วิธี Mutual Expectation ซึ่งวิธีทางสถิติเหล่านี้มีหลักการร่วมกันคือการคำนวณหาค่าความสัมพันธ์ระหว่างหน่วยย่อยโดยหน่วยย่อยซึ่งในที่นี้คือพยางค์ที่มีความสัมพันธ์ระหว่างกันสูงหรือปรากฏร่วมกันบ่อยๆ ในคลังข้อมูลจะถูกดึงออกมา ซึ่งกลุ่มพยางค์ที่มีค่าความสัมพันธ์ระหว่างกันมากๆ อาจจะไม่ได้เป็นชื่อเฉพาะก็ได้ เช่นคำว่า "ปรึกษา" จะถูกแบ่งออกเป็นสองพยางค์ จากนั้นเมื่อผ่านเข้าสู่กระบวนการประมวลผลทางสถิติ จะพบว่า "ปรึกษา" มีความสัมพันธ์ระหว่างกันค่อนข้างสูง เพราะพบว่าคำว่า "ปรึก" กับ "ษา" มักอยู่ติดกัน โดยคำว่า "ปรึก" จะไม่นำหน้าคำอื่นๆ ได้อีกเลย แม้ว่าคำว่า "ษา" อาจจะตามหลังคำอื่นอย่างคำว่า "รัก" ได้ก็ตาม จึงทำให้คำนี้ถูกเลือกออกมาด้วยวิธีการสถิติด้วย

แต่จากผลการคำนวณค่าความครบถ้วนจะพบว่าวิธีทางสถิติทั้ง 4 วิธีให้ผลไม่ต่ำกว่า 60% ซึ่งนับว่าค่อนข้างดี ยกเว้นวิธีไคกำลังสอง (Pearson's Chi-square) ซึ่งให้ผล 49.645% แสดงให้เห็นว่าชื่อเฉพาะก็เป็นกลุ่มคำที่มีความสัมพันธ์กันค่อนข้างสูง ชื่อเฉพาะส่วนหนึ่งที่ไม่สามารถเลือกออกมาได้ด้วยวิธีการสถิติ เพราะวิธีทางสถิติเหล่านี้ก็ยังมีข้อจำกัดประการหนึ่งคือ กรณีที่ชื่อเฉพาะเป็นคำที่มีเพียงพยางค์เดียว เช่น นาย<person>ชวน</person> หรือ ประเทศ<location>ไทย</location> จะไม่สามารถผ่านเงื่อนไขของวิธีทางสถิติที่พยายามหาค่าความสัมพันธ์ระหว่างพยางค์ เพราะชื่อเฉพาะที่จะผ่านเงื่อนไขไปได้ต้องมีไม่ต่ำกว่า 2 พยางค์ขึ้นไป โดยชื่อเฉพาะที่มีเพียงพยางค์เดียวจะมีจำนวน 1028 จากชื่อเฉพาะจำนวน 11683 ชื่อ โดยคิดเป็น 8.799% ของชื่อเฉพาะทั้งหมด

และจากผลการคำนวณอัตราการรู้จำหรือค่า F แสดงให้เห็นว่าวิธีที่ใช้ค่า Mutual Expectation นั้นให้ผลดีที่สุด ทั้งนี้อาจเป็นเพราะแนวคิดของวิธีนี้เป็นการหาความสัมพันธ์ระหว่างคำโดยการคำนวณความเป็นไปได้ที่คำต่างๆ ในสาย n-grams จะหายไป ซึ่งทำให้ได้ค่าความแม่นยำสูงที่สุดในวิธีทางสถิติทั้งหมด แต่วิธี Mutual Expectation นี้ก็ใช้เวลาในการประมวลผลมาก จนทำให้ผู้วิจัยเลือกใช้วิธีทางสถิติที่ให้ผลดีรองลงมา นั่นคือวิธี Mutual Information