การวิเคราะห์ความหมายของคำ

Word Sense Analysis

โดย

นางสาวอรวรรณ โชติกิจนุสรณ์

รหัส 42067025

H001766

อาจารย์ที่ปรึกษา

ดร. โชติพัชร์ ภรณวลัย

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2543

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การวิเคราะห์ความหมายของคำ
นักศึกษา	นางสาวอรวรรณ โชติกิจนุสรณ์
อาจารย์ที่ปรึกษา	ดร.โชติพัชร์ กรณวลัย
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2543

บทคัดย่อ

เราได้นำเสนอแนวคิดของการใช้ฐานข้อมูลที่ช่วยในการประเมินทางสถิติ (Corpus-based) มาใช้ในการแก้ปัญหาเรื่องการลดกำกวมของคำ (Word Sense Disambiguation) โดยต้องการที่จะดึงข้อมูลออกมาได้อย่างอัตโนมัติจากข้อมูลดิบที่ยังไม่ได้ผ่านการวิเคราะห์ (Untagged text) เช่น กำหนดความหมาย การที่จะดึงข้อมูลนี้ออกมาได้จะใช้รูปแบบการเรียนรู้ทางสถิติแบบเรียนรู้ด้วยตนเอง (Unsupervised Learning of Probabilistic) โดยการพิจารณาจากคุณสมบัติของคำแวดล้อมที่เราทราบค่ากับคำที่มีความหมายกำกวม และนำข้อมูลที่ได้มาทำเป็นข้อมูลที่ผ่านการเรียนรู้แล้ว (Training Data) มาช่วยในการพิจารณาความหมายของคำที่ต้องการกำหนดความหมายต่อไป

Title	Word Sense Analysis
Student	Miss.Orawan Chotkijnusorn
Advisor	Dr. Chotipat Pornavalai
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2000

ABSTRACT

We present a corpus-based approach to solve word-sense disambiguation problem that only requires information that can be automatically extracted from untagged text such as defined meaning of words. We use Unsupervised Learning of Probabilistic Model Techniques and analyze features of ambiguous contexts that we known value and ambiguous word. All Process words that is Training Data are helper to decide meaning of new words.

สารบัญ

หน้า

สารบัญภาพ

หน้า

ภาพที่

สารบัญตาราง

หน้า

ตารางที่

บทที่ 1
บทนำ

ในกระบวนการประมวลผลภาษาธรรมชาติ (Natural Processing Language) ขั้นตอนหนึ่งในการทำงานคือ การแปลความหมาย (semantic interpretation) ให้ถูกต้อง แต่ปัญหาสำคัญที่พบในขั้นตอนนี้ คือ การกำหนดความหมายที่แท้จริงของคำในประโยค (word sense) เนื่องจากคำๆ หนึ่งอาจจะมีได้หลายความหมาย (Polysemous) หรือที่เรียกได้ว่ามีได้หลาย Sense การที่คำๆ นี้ไปเรียงตัวอยู่ในประโยคที่ต่างกันก็อาจที่จะมีการอ้างถึงความหมายที่แตกต่างกันได้

ดังนั้นงานหลักของการแก้ปัญหานี้ คือ การหาวิธีในการกำหนดความหมายของคำในประโยค หรือการถอดความกำกวมของคำ (Word Sense Disambiguation)

สำหรับการศึกษาในครั้งนี้จะเป็นการศึกษาถึงแนวความคิดในการกำหนดความหมายในรูปแบบต่างๆ โดยเฉพาะการใช้แนวคิดแบบ Corpus เพราะเป็นการใช้ประโยคตัวอย่างจากความเป็นจริงมาช่วยในการตัดสินความหมายของคำที่อยู่ในประโยคที่พิจารณา ทำให้มีการกำหนดความหมายใกล้เคียงกับคนมากขึ้น แต่เนื่องจากการที่จะเตรียมข้อมูลตัวอย่างให้เพียงพอ และครบถ้วนยังเป็นเรื่องที่จะต้องใช้เวลา และใช้แรงงานของคนมาก จึงสนใจไปที่การที่จะหาแนวทางในการกำหนดความหมายแบบอัตโนมัติ ถึงแม้การทำงานตามกระบวนนี้ยังคงจะเป็นเพียงแค่กำหนดกลุ่มของความหมาย (Sense Group) ท้ายที่สุดของการทำงานก็เป็นหน้าที่ของคนที่จะเข้าไปช่วยตัดสินความหมาย แต่ก็สามารถที่จะช่วยลดขั้นตอนที่คนจะเข้าไปกำหนดทุกความคำในแต่ละประโยค เป็นกลุ่มของประโยคแทน ผลสุดท้ายของการศึกษาจึงเป็นการนำแนวคิดนี้ไปช่วยในการสร้างฐานข้อมูล (Corpus) ที่จะทำงานร่วมกับการตัดสินใจของคน ให้ได้ Corpus ที่จะทำหน้าที่เป็นตัวอย่างประโยคที่มีการกำหนดความหมายในแต่ละคำแล้ว (Training Data) สำหรับประโยคใหม่ที่ต้องการพิจารณา (Testing Data)

ในบทที่ 2 จะอธิบายถึงทฤษฎีพื้นฐานในการแก้ปัญหา เพื่อที่จะใช้เป็นแนวทางในการพัฒนาประสิทธิภาพของการกำหนดความหมายของคำให้ถูกต้องมากยิ่งขึ้น

ในบทที่ 3 จะอธิบายถึงแนวทางในถอดความกำกวมของการกำหนดความหมายของคำในการศึกษาในครั้งนี้ ได้แก่ วิธีการการเรียนรู้ถึงความหมายของคำแต่ละคำด้วยวิธีการทางสถิติ (Probabilistic Model) แบบเรียนรู้ด้วยตนเอง (Unsupervised Learning)

ในบทที่ 4 จะแสดงถึงขั้นตอนในการสร้างฐานข้อมูลเพื่อที่จะใช้เป็น Training Data ว่าในแต่ละขั้นตอนมีการทำงานเป็นอย่างไร และต้องใช้ข้อมูลอะไรบ้าง นอกจากนั้นยังจะรวมถึงผลจากการที่ได้ทดสอบจริง ส่วนบทที่ 5 จะแสดงถึงตัวอย่างโปรแกรมประยุกต์ที่มีการนำฐานข้อมูล (Corpus) ไปเป็น Training data ในการตัดสินใจความหมายของประโยคที่สนใจ

ในบทที่ 6 จะแสดงถึงบทสรุป และแนวทางที่จะพัฒนาจะปรับปรุงต่อไปในอนาคต

บทที่ 2

วิธีการกำหนดความหมายของคำ

ในบทนี้จะกล่าวถึงวิธีการ (Algorithm) ในการกำหนดความหมายของคำที่ผ่านมา โดยในที่นี้จะอธิบายถึงวิธีในการแก้ปัญหาหลักๆ อยู่ 2 วิธี ได้แก่ การกำจัดความหมายที่เป็นไปไม่ได้ออกไป (Selection Restriction) ที่จะอธิบายในส่วนที่ 2.1 และแนวความคิดแบบ Corpus-based Approach ในส่วนที่ 2.2

2.1 การกำจัดความหมายที่เป็นไปไม่ได้ออกไป (Selection Restriction)

ซึ่งจะเป็นวิธีการที่ใช้กำจัดรูปแบบประโยคที่ความหมายของคำไม่สามารถเรียงต่อกันได้ โดยการกำหนดเป็นข้อจำกัด

ตัวอย่าง

"Mary drank burgundy"

คำว่า burgundy มี 2 ความหมาย ได้แก่

- สี
- เครื่องดื่ม

แต่ถ้าเรามีข้อจำกัดที่ว่า drink + LIQUID ด้วยข้อจำกัดนี้ทำให้เราสามารถกำหนดความหมายของคำว่า burgundy ว่า เครื่องดื่มได้

ข้อดี

- กำจัดรูปแบบที่ยอมรับไม่ได้ออกไป

ข้อเสีย

- ถ้าขอบเขตของสิ่งที่สนใจเพิ่มมากขึ้น การสร้างข้อจำกัดด้วยคน มีความยากลำบาก และเสียเวลา
- เกิดปัญหากับประโยคปฏิเสธ และประโยคอุปมาอุปไมย เช่น

  "My car drinks gasoline"

2.2 แนวคิดแบบ Corpus-based approach

เนื่องจากวิธี Selection Restrictions จะเป็นเพียงการแบ่งแยกอย่างหยาบๆ ของรูปแบบที่ยอมรับได้หรือยอมรับไม่ได้ ผลลัพธ์คือ ทำให้อีกหลายๆ กรณีที่เป็นความหมายที่กำกวมยังไม่สามารถที่ได้รับการแก้ไข วิธีที่ดีกว่าก็จะเป็นรูปแบบที่เป็นการประมวลผลแบบมนุษย์ คือ วิธีการที่ใช้ฐานข้อมูลขนาดใหญ่ (Corpus) ในการเก็บรวบรวมตัวอย่างของคำที่เรียงตัวอยู่ในประโยคต่างๆ โดยที่ฐานข้อมูลนี้จะอาศัยคนในการพิจารณา และมีการกำหนด (tagged) ความหมายของคำในแต่ละประโยคเอาไว้ เพื่อที่จะใช้ฐานข้อมูล (Corpus) นี้เป็นข้อมูลตัวอย่างของการเรียนรู้ (Training Data) และนำไปช่วยในการวิเคราะห์คำในประโยคที่กำลังพิจารณาทางสถิติ ได้หลายแบบดังนี้

- Unigram Statistic
- Context

2.2.1 Unigram Statistic

เป็นวิธีที่ง่ายที่สุดในการวิเคราะห์ทางสถิติ ซึ่งจะเป็นวิธีที่อาศัยการเก็บความถี่ของการเกิดคำๆ หนึ่ง ในความหมายต่างๆ ที่เป็นไปได้จาก Training Data เพื่อใช้ในการทำนายความหมายของคำในประโยคใหม่ ผลลัพธ์ที่ได้ คือ จะใช้ความหมายเดียวกับความหมายที่ถูกใช้บ่อย (มีความถี่สูงสุดที่ได้จาก Training Data) ตัวอย่างเช่น

> ถ้าพิจารณาถึงความหมายของคำว่า "bridge" และใน Corpus พบคำว่า bridge 5845 ครั้ง
> มี 5651 ครั้ง หมายถึง สะพาน
> 194 ครั้ง หมายถึง เหล็กดัดฟัน

ซึ่งถ้าเราพิจารณตามวิธีการของ Unigram Statistic เราจะทำนายความหมายของคำว่า bridge ว่าเป็น ความหมายเดียวกับความหมายที่ถูกใช้บ่อย คือ bridge จะหมายถึง สะพาน เท่านั้นแต่ถ้าเราพบคำว่า bridge ในเอกสารที่เกี่ยวกับการแต่งฟัน ความหมายของคำว่า bridge ก็มีความเป็นไปได้ที่จะหมายถึง เหล็กดัดฟันมากกว่า

2.2.2 Context

เป็นวิธีการทางสถิติอีกแบบหนึ่ง ที่จะไม่ได้สนใจที่ความถี่ของคำที่จะกำหนดความหมายของคำที่กำกวมเพียงอย่างเดียว แต่จะพิจารณาถึงคำแวดล้อม (Context) ของคำที่สนใจ วิธีนี้จะทำให้สามารถกำหนดความได้ถูกต้องยิ่งขึ้น อย่างตัวอย่างที่แล้ว ถ้าเราพบคำว่า teeth (ฟัน), dentist (ทันต

แพทย์), cavity (ช่องปาก), orthodontics (วิชาที่ว่าด้วยการแต่งฟัน) พบอยู่ในประโยคเดียวกับคำว่า bridge ทำให้อาจจะต้องเลือกความหมายว่า "เหล็กดัดฟันมากกว่า" โดยที่คำแวดล้อมเหล่านี้เราจะเรียกว่า **word collocation** ซึ่งก็คือ คำอะไรก็ตามที่มีแนวโน้มที่จะปรากฏร่วมกับคำที่พิจารณา ในการพิจารณาถึงคำแวดล้อมเราสามารถแบ่งแยกตามจำนวนคำที่ถูกพิจารณาร่วมได้ดังนี้

1. Bigram เป็นวิธีการที่จะพิจารณาคำที่เกิดร่วมกับคำที่จะกำหนดความหมายเพียง 1 คำ
2. Trigram เป็นวิธีการที่จะพิจารณาคำที่ปรากฏก่อนหน้าคำที่จะกำหนดความหมาย 2 คำ
3. N-gram เป็นวิธีการที่จะพิจารณาคำที่ปรากฏก่อนหน้า n คำ
4. พิจารณาทั้งประโยค
5. พิจารณาประโยคอื่น (Discourse)

จำนวนของคำที่ร่วมในการพิจารณากับคำที่กำกวมเราจะเรียกว่า window แต่ด้วยวิธีนี้จำเป็นจะต้องใช้ Corpus ที่มีการกำหนดว่าแต่ละคำที่ปรากฏในประโยคมีความหมายว่าอย่างไร เพื่อสามารถนำไปคำนวณหาค่าความน่าจะเป็นทางสถิติ สำหรับแนวความคิดแบบง่ายๆ ในการประเมินความน่าจะเป็นของความหมายของคำที่กำกวม (กำหนดให้เป็น I) ที่สัมพันธ์กับ window ของคำ สมมติให้ window มีขนาด n คำ จะพิจารณาคำ w ให้อยู่ตรงกลางของคำแวดล้อม n - 1 คำ ดังนี้

$$w_1 w_2 .. w_n / 2 \mathbf{W} w_n / 2+1 .. w_n - 1$$

วัตถุประสงค์ คือ เราต้องการคำนวณความหมาย S ของคำ w ที่ให้ค่าสูงสุดเป็นความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) ตามนี้

$$PROB(w / s \mid w_1 w_2 .. w_n / 2 \mathbf{W} w_n / 2+1 .. w_n - 1) \qquad (1)$$

โดยที่ w / s คือ คำ w ที่มีความหมายเป็น S

จากกฎของ Bayes (Bayes's rule)

$$PROB(A \mid B) = \frac{PROB(B \mid A) * PROB(A)}{PROB(B)} \qquad (2)$$

ตามกฎของ Bayes จะได้รูปแบบของสมการที่ (1) ดังนี้

$$PROB(w / s) = \frac{PROB(w_1 w_2 .. w_{n/2} w w_{n/2+1} .. w_{n-1} / w / s) * PROB(w / s)}{PROB(w_1 w_2 .. w_{n/2} w w_{n/2+1} .. w_{n-1})} \quad (3)$$

โดยสมมติฐานที่ว่าแต่ละ w_i ปรากฏเป็นอิสระกับคำอื่นๆ ใน window ทำให้เราสามารถประมาณค่าของ $PROB(w_1 w_2 .. w_{n/2} w w_{n/2+1} .. w_{n-1} / w / s)$ ได้ดังนี้

$$PROB(w_1 w_2 .. w_{n/2} w w_{n/2+1} .. w_{n-1} / w / s) = \prod_{i=1, n-1} PROBn(w_i / w / s) \quad (4)$$

โดยที่ $PROBn(w_i / w / s)$ เป็นความน่าจะเป็นที่คำ w_i ที่เกิดขึ้นภายใน n คำของ window เพราะฉะนั้นความหมายที่ดีที่สุดจะคำนวณได้จากสูตร

$$PROB_n(w_i / w / s) = PROB(w / s) * \prod_{i=1, n-1} \quad (5)$$

โดยที่ $PROB(w / s)$ คือ ความน่าจะเป็นของคำ w จะมีความหมาย S ตลอดทั้ง Corpus และ

$$PROB_n(w_i / w / s) = \frac{\text{จ.น.ครั้งในCorpusที่เกิดกับคำพและมีความหมายS}}{\text{จ.น.ครั้งในCorpusที่คำพจะมีความหมายS}} \quad (6)$$

พิจารณาดังตัวอย่างต่อไปนี้ สมมติให้มีข้อมูลใน Corpus 10 ล้านคำ ข้อมูลที่เกี่ยวข้องกับคำว่า bridge สรุปได้ดังตารางที่ 2.1 ซึ่งจะใช้ window ที่มีขนาด 11 ตัว จากข้อมูลในตารางที่ 2.1 มีการคำนวณตามนี้

	Bridge (สะพาน)	Bridge (เหล็กดัดฟัน)	ใน window อื่น
teeth	1	10	300
suspension	200	1	2,000
the	5,500	180	500,000
dentist	2	35	900
จ.น. รวมที่เกิด	5651	194	501,500

ตารางที่ 2.1 สรุปจำนวนความหมายสำหรับคำว่า bridge ใน corpus ที่สมมติขึ้นมา

ในกรณีคำนวณ Bridge ในความหมายว่า "สะพาน" จะได้ดังนี้

$PROB(teeth \mid bridge / สะพาน) = 1/5{,}651 = 1.77*10^{-4}$

$PROB(suspension \mid bridge / สะพาน) = 200/5{,}651 = 0.035$

$PROB(the \mid bridge / สะพาน) = 5{,}500/5{,}651 = 0.97$

$PROB(dentistbridge / สะพาน) = 2/5{,}651 = 3.54*10^{-4}$

$PROB(bridge / สะพาน) = 5{,}651/501{,}500$

สรุปจากสูตรที่ (5) จะได้ความเป็นไปได้ที่ bridge จะมีความหมายว่า "สะพาน" มีค่าเท่ากับ $(0.113) * (1.77*10^{-4}) * (0.035) * (0.97) * (3.54*10^{-4}) = 2.4*10^{-10}$

ในกรณีคำนวณ Bridge ในความหมายว่า "เหล็กดัดฟัน" จะได้ดังนี้

$PROB(teeth \mid bridge / เหล็กดัดฟัน) = 10/109 = 0.052$

$PROB(suspension \mid bridge / เหล็กดัดฟัน) = 1/194 = 5.15*10^{-3}$

$PROB(the \mid bridge / เหล็กดัดฟัน) = 108/194 = 0.93$

$PROB(dentistbridge / เหล็กดัดฟัน) = 35/194 = 0.18$

$PROB(bridge / เหล็กดัดฟัน) = 194/501{,}500 = 3.87*10^{-4}$

สรุปจากสูตรที่ (5) จะได้ความเป็นไปได้ที่ bridge จะมีความหมายว่า "เหล็กดัดฟัน" มีค่าเท่ากับ $(3.87*10^{-4}) * (0.052) * (5.15*10^{-3}) * (0.93) * (0.18) = 1.7*10^{-8}$

เพราะฉะนั้นจะเห็นว่าความเป็นไปได้ของคำว่า bridge ถ้าเกิดร่วมกับคำว่า teeth, suspension, the และ dentist จึงจะมีความหมายว่า "เหล็กดัดฟัน" มากกว่า

จากวิธีการที่กล่าวข้างต้นเป็นวิธีพื้นฐานของแนวคิดแบบ Corpus-based approach อีกหลาก หลายรูปแบบ ถึงแม้จะมีการกำหนดความหมายได้ดีกว่าการกำหนดแบบ Selection Restriction แต่ก็ ยังพบปัญหาอีกหลายเรื่อง ได้แก่

- ต้องมีฐานข้อมูล (Corpus) ที่มีการกำหนดความหมายโดยคน
- ถ้ามีฐานข้อมูลแล้วฐานข้อมูลก็ต้องมีประสิทธิภาพพอเพียง เนื่องจากอาจจะเกิดปัญหา ในกรณีที่มีข้อมูลน้อยๆ หรือไม่พบข้อมูลนั้นเลยอาจจะทำให้ไม่สามารถกำหนดความ หมายได้เลย (Data sparseness)

บทที่ 3
ตัวแบบการเรียนรู้ทางสถิติด้วยตนเอง

ในการศึกษาครั้งนี้จะมุ่งเน้นไปตามแนวคิดแบบ Corpus-based ประกอบกับตัวแบบความน่าจะเป็น (Probabilistic Models) เพื่อใช้แก้ปัญหาความหมายของคำที่กำกวม โดยที่ตัวแบบนี้จะเป็นตัวบ่งชี้ว่าความหมายใดควรจะเป็นความหมายที่ถูกต้องเมื่อเรียงตัวอยู่ในคำแวดล้อมที่เกิดขึ้น

เนื่องจากตัวแบบความน่าจะเป็น (Probabilistic Models) เป็นส่วนหนึ่งของการอนุมานทางสถิติ ดังนั้นในบทนี้จะได้กล่าวถึง ความรู้ทั่วไปเกี่ยวกับการอนุมานทางสถิติ ก่อนในส่วนที่ 3.1 และจะอธิบายถึงวิธีการนำกระบวนการอนุมานมาประยุต์ใช้กับปัญหาการกำหนดความหมายของคำที่กำกวมในส่วนที่ 3.2 สำหรับในส่วนที่ 3.3 และ 3.4 จะได้อธิบายถึงการที่จะใช้ตัวแบบความน่าจะเป็นประกอบกับการเรียนรู้ด้วยตนเอง

3.1 การอนุมานทางสถิติ (Inferential Statistic)

การอนุมาน (Inference) นี้เป็นการศึกษาเกี่ยวกับการนำข้อมูลตัวอย่าง (Sample) ที่ได้มาจากการสังเกต (Observation) หรือ จากการทดลอง (Experiment) หนึ่งๆ ไปอธิบายถึงข้อมูลทั้งหมดของเรื่องใดเรื่องหนึ่งที่กำลังศึกษา (ซึ่งเรียกว่า ประชากร (Population)) นั้นว่ามีลักษณะอย่างไร สำหรับการอนุมานทางสถิติ (Inferential Statistic) จะเป็นการสร้างตัวแบบ (Model) ทางสถิติโดยอาศัยข้อมูล (Data) หรือสารสนเทศ (Information) จากกลุ่มตัวอย่างแล้วนำตัวแบบที่ได้มาพิจารณาอนุมานว่าเมื่อข้อมูลตัวอย่างที่สุ่มมามีลักษณะเป็นเช่นนี้แล้ว อาศัยทฤษฎีทางสถิติจะสรุปผลหรืออ้างอิงถึงประชากรที่กำลังศึกษาได้เช่นไร โดยคำนึงถึงคุณภาพของตัวอย่างสุ่มที่ได้ว่าเป็นตัวแทนที่ดีของประชากรหรือไม่ รวมทั้งคุณภาพของการสรุปผลด้วย ทั้งนี้เพื่อลดความผิดพลาดอันเกิดขึ้นจากการสรุปผลที่อาศัยเพียงตัวอย่างสุ่มจำนวนหนึ่งเท่านั้น

ประชากรหนึ่ง ๆ จะมีลักษณะเป็นอย่างไร เนื่องจากข้อมูลในประชากรถูกควบคุมด้วยกลไกบางอย่าง ซึ่งกลไกนี้สามารถอธิบายได้ด้วยตัวแบบความน่าจะเป็น (Probability Model) กล่าวคือ ลักษณะการเกิดขึ้นของข้อมูลนั้นมักจะมีรูปแบบที่ค่อนข้างคงที่แน่นอน และสามารถหารูปแบบดังกล่าวได้จากการอาศัยข้อมูลซึ่งได้จากการทดลองซ้ำๆ กันหลายๆ ครั้ง โดยทั่วไปแล้วตัวแบบความน่าจะเป็นนั้นมักจะมีค่าคงที่ที่ไม่ทราบค่าซึ่งใช้แสดงคุณสมบัติบางประการของประชากรรวมอยู่ด้วย เรียกค่าคงที่นี้ว่า พารามิเตอร์ (Parameter) เช่น ค่าเฉลี่ย (μ) ความแปรปรวน (σ) และสัดส่วน

(P) เป็นต้น ทั้งนี้จำนวนพารามิเตอร์ในตัวแบบความน่าจะเป็นใด ๆ อาจจะมีมากกว่าหรือเท่ากับหนึ่งตัวก็ได้ ซึ่งในที่นี้จะใช้สัญลักษณ์อักษรกรีก θ แทน พารามิเตอร์

หากตัวแบบความน่าจะเป็นสามารถอธิบายลักษณะการเกิดขึ้นของข้อมูลได้อย่างถูกต้องแล้ว การอนุมานทางสถิติจะเข้ามามีบทบาทในขั้นตอนต่อไป โดยจะพยายามค้นหาค่าพารามิเตอร์ที่ปรากฏอยู่ในตัวแบบนั้นว่ามีค่าเป็นเท่าไร ด้วยการอาศัยข้อมูลในตัวอย่างที่สังเกตได้ และวิธีการอย่างหนึ่งอย่างใดหรือมากกว่าดังต่อไปนี้

1. การประมาณค่า (Estimation) เป็นการใช้ข้อมูลจากตัวอย่างสุ่มที่มีอยู่มาคะประมาณค่าหรือสรุปผลเกี่ยวกับค่าพารามิเตอร์ว่ามีค่าเป็นเท่าไร ซึ่งตัวประมาณค่า สามารถกระทำด้วยการประมาณค่าแบบค่าเดียว และการประมาณค่าแบบช่วง
2. การทดสอบสมมติฐาน (Hypothesis testing) เป็นการใช้ข้อมูลจากตัวอย่างสุ่มที่มีอยู่และความรู้เกี่ยวกับตัวแบบความน่าจะเป็นมาทดสอบความเชื่อเกี่ยวกับค่าพารามิเตอร์หรือประชากรที่สนใจว่าเป็นเช่นนี้หรือไม่ ในบางสถานการณ์สามารถใช้การประมาณค่าแบบช่วงแทนการทดสอบสมมติฐาน
3. การตัดสินใจ (Decision) เป็นการอาศัยข้อมูลที่สังเกตได้ในตัวอย่างสุ่มมาช่วยตัดสินใจเลือกการกระทำ (action) ที่เหมาะสมเพื่อให้ความสูญเสียที่จะเกิดขึ้นจากการตัดสินใจเลือกการกระทำนั้นมีค่าน้อยที่สุด

ตัวอย่าง สมมติว่า "วิธีการรักษาโรคมะเร็งในปอดแบบใหม่ได้ถูกเสนอขึ้น โดยทำการทดลองกับคนไข้ที่เป็นโรคมะเร็งปอดจำนวน 100 คน ภายหลังจากการได้รับการรักษาวิธีนี้ ก็ได้มีการบันทึกผลการรักษาว่าจะมีชีวิตอยู่รอดต่อไปหรือไม่หลังจากได้รับการรักษาโดยใช้วิธีนี้ไปแล้ว 3 ปี และผลปรากฏว่ามีคนไข้ 80 คนที่มีชีวิตอยู่รอด"
จากข้อมูลที่ได้

ตัวแบบความน่าจะเป็น : ให้ X แทนจำนวนคนไข้ที่ได้รับการรักษาด้วยวิธีใหม่นี้ และมีโอกาสที่แต่ละคนจะมีชีวิตอยู่รอดต่อไปได้อีก 3 ปี เท่ากับ P โดยคนไข้แต่ละคนเป็นอิสระกันไม่ส่งผลกระทบต่อกัน ดังนั้น X มีตัวแบบความน่าจะเป็นแบบทวินามซึ่งมีพารามิเตอร์ (Parameter) คือ P ที่ซึ่งอยากจะรู้ว่ามีค่าเป็นเท่าไร

การประมาณค่า : จากข้อมูลที่ได้หากประมาณค่า P โดยใช้ $\hat{P}$ ซึ่งกำหนดให้มีค่าเท่ากับจำนวนคนไข้ที่มีชีวิตรอดหารด้วยจำนวนคนไข้ทั้งหมดที่ได้รับการทดลองอาศัยข้อมูลในตัวอย่างทดลองนี้จะได้

$\hat{P} = 80/100 = 0.80$

การทดสอบสมมติฐาน : หากอยากทราบว่า วิธีการรักษาแบบใหม่นี้จะให้ประสิทธิภาพในการรักษาแบบปัจจุบันที่ใช้อยู่หรือไม่ ถ้าทราบว่าความน่าจะเป็นที่วิธีการรักษาแบบปัจจุบันจะช่วยให้คนไข้มีชีวิตอยู่รอดต่อมาได้อีก 3 ปี มีค่าเท่ากับ 0.70 ซึ่งต้องทำการทดสอบด้วยการนำข้อมูลตัวอย่างสุ่มที่ได้นั้นมาใช้ทดสอบ

การตัดสินใจ : ควรจะตัดสินใจเลือกวิธีการรักษาแบบใดจึงจะมีประสิทธิภาพสูงสุดและก่อให้เกิดความสูญเสียน้อยที่สุด ถ้าทราบเพิ่มเติมว่าวิธีการรักษาแบบปัจจุบันเสียค่าใช้จ่าย 10,000 บาท/คน แต่หากรักษาโดยวิธีการรักษาแบบใหม่ จะเสียค่าใช้จ่าย 25,000 บาท/คน

3.1.1 ตัวแบบความน่าจะเป็น (Probabilistic Model)

วัตถุประสงค์อย่างหนึ่งของนักวิทยาศาสตร์ คือ การพยายามที่จะอธิบายลักษณะการเกิดขึ้นของสิ่งต่าง ๆ หรือปรากฏการณ์ต่างๆ ที่เกิดขึ้นในโลก ซึ่งวิธีการหนึ่งที่สามารถกระทำได้ คือ การสร้างตัวแบบทางคณิตศาสตร์ขึ้นมาเพื่อใช้อธิบายการเกิดขึ้นของปรากฏการณ์นั้นๆ การพยายามที่จะสร้างตัวแบบความน่าจะเป็น เพื่อนำมาใช้อธิบายการเกิดขึ้นของผลลัพธ์ที่ไม่สามารถคาดการณ์ได้ล่วงหน้าได้อย่างแน่นอนจากการทดลองสุ่ม (Random experiment) ครั้งหนึ่งๆ

ในเรื่องของตัวแบบความน่าจะเป็นจะเกี่ยวข้องกับคำว่า "ตัวแปรสุ่ม" และ "เวคเตอร์สุ่ม" โดยที่จะคำจะมีความหมาย ดังนี้

ตัวแปรสุ่ม (random variable) เป็นฟังก์ชั่นค่าจริงทางคณิตศาสตร์อย่างหนึ่งสำหรับคาดการณ์ลักษณะที่จะเกิดขึ้นจากการทดลองสุ่มอย่างหนึ่งที่มีผลลัพธ์ ซึ่งอาจทำนายหรือคาดการณ์ได้ล่วงหน้าอย่างแน่นอน หรือ หมายถึง ฟังก์ชันที่มีโดเมนเป็นกลุ่มผลการทดลองเชิงสุ่ม และมีพิสัยเป็นเซตของเลขจำนวนจริง

ถ้าใช้ X, Y และ Z เป็นสัญลักษณ์ แทน ตัวแปรสุ่ม และใช้ x,y และ z แทนค่าตัวแปรสุ่ม

ตัวแปรสุ่มแบ่งได้ 2 ชนิด คือ ตัวแปรสุ่มชนิดไม่ต่อเนื่อง (discrete random variable) และตัวแปรสุ่มชนิดไม่ต่อเนื่อง (continuous random variable) ทั้งนี้ ตัวแปรสุ่มชนิดไม่ต่อเนื่องได้แก่ ตัวแปรสุ่ม ที่มีค่าเป็นจำนวนเต็มนับได้ หรือ จำนวนเต็มแบบอนันต์แต่นับได้ เช่น

ให้ X เป็นจำนวนของสินค้าที่เสียในช่วงการผลิตกะหนึ่ง

x มีค่าที่เป็นไปได้เท่ากับ 0, 1, 2, ...

ให้ Y เป็นจำนวนครั้งของการเกิดก้อยจากการโยนเหรียญบาทหนึ่งอัน 2 ครั้ง

y มีค่าที่เป็นไปได้เท่ากับ 0, 1, 2

และจะใช้สัญลักษณ์ P(X=x) แทน ฟังก์ชันความน่าจะเป็นของตัวแปรสุ่ม X เมื่อ X เป็นตัวแปรสุ่มชนิดไม่ต่อเนื่อง

สำหรับตัวแปรสุ่มชนิดต่อเนื่อง ได้แก่ ตัวแปรสุ่มที่มีค่าหลายค่านับไม่ถ้วน เช่น

ให้ Z เป็นน้ำหนักของนักศึกษาคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีเจ้าคุณทหารลาดกระบัง มีหน่วยเป็นกิโลกรัม

$$Z = \{z \mid 0 <= z >= \infty\}$$

ซึ่งค่า z ที่ได้จะมีค่าได้มากมายอาจเป็นตัวเลขจำนวนเต็ม หรือตัวเลขที่เป็นทศนิยมก็ได้ ส่วนตัวอย่างอื่นๆ ของตัวแปรสุ่มชนิดต่อเนื่อง เช่น ส่วนสูง รายได้ คะแนะ เวลา เป็นต้น และจะใช้สัญลักษณ์ f(x) แทนฟังก์ชันความน่าจะเป็นของตัวแปรสุ่ม X เมื่อ X เป็นตัวแปรสุ่มชนิดต่อเนื่อง

เวคเตอร์สุ่ม (random vector) โดยทั่วๆ ไปแล้วในการทดลองหนึ่งๆ มักจะมีการวัดค่าผลลัพธ์ที่คาดว่าจะเป็นได้มากกว่า 1 ครั้ง หรือกระทำการทดลองซ้ำๆ กัน n ครั้ง ดังนั้นจะใช้สัญลักษณ์ X แทนเซตของการสังเกตที่ได้นี้จะเรียกว่า เวคเตอร์สุ่ม

ปกติแล้วตัวแปรสุ่มจะมีค่าได้หลายค่า ค่าที่ถือว่าเป็นตัวแทนที่ดีของตัวแปรสุ่ม คือ ค่าเฉลี่ย แต่เนื่องจากตัวแปรสุ่ม จะมีค่าเท่าใดก็ด้วยความน่าจะเป็นหนึ่งๆ ดังนั้น ค่าเฉลี่ยดังกล่าวจึงเป็นค่าเฉลี่ยที่คาดว่าจะเป็นหรือ ค่าคาดหวัง (expected value) ของตัวแปรสุ่ม

ดังนั้นแล้วตัวแบบความน่าจะเป็น (ฟังก์ชันความน่าจะเป็น) จะเป็นการอธิบายลักษณะของผลลัพธ์ที่คาดว่าจะเกิดขึ้นจากการทดลองสุ่มที่มีลักษณะต่างๆ กัน โดยตัวแบบความน่าจะเป็นน่าจะแบ่งเป็น 2 ประเภทตามชนิดของตัวแปรสุ่ม ได้แก่

- ตัวแบบความน่าจะเป็นของตัวแปรสุ่มชนิดต่อเนื่อง เช่น ตัวแบบความน่าจะเป็นแบบทวินาม (binomial probability model) ตัวแบบความน่าจะเป็นแบบพัวซอง (poisson

probability model) ตัวแบบความน่าจะเป็นแบบเรขาคณิต (geometric probability model) เป็นต้น

- ตัวแบบความน่าจะเป็นของตัวแปรสุ่มชนิดต่อเนื่อง ซึ่งได้แก่ ตัวแบบความน่าจะเป็นแบบสม่ำเสมอ (uniform probability model) ตัวแบบความน่าจะเป็นแบบปกติ (normal probability model) และตัวแบบความน่าจะเป็นแบบเอกซ์โปเนนเชียล (exponential probability model)

3.1.2 แนวทางความคิดของการอนุมานทางสถิติ

การอนุมานทางสถิติมักเกี่ยวข้องกับการสุ่มตัวอย่างเพียงบางส่วนจากประชากรที่มีขนาดใหญ่ และไม่รู้ลักษณะ แล้วจึงใช้ข้อมูลที่ได้นี้ทำการอนุมานเพื่ออธิบายลักษณะของประชากรนั้น ด้วยเหตุนี้จึงไม่สามารถที่จะเชื่อได้อย่างแน่นอนว่า ผลสรุปที่ได้นั้นจะถูกต้องหรือไม่ ความไม่แน่นอนนี้จะมีความเกี่ยวข้องกับความน่าจะเป็นซึ่งอาจจะอยู่ในรูปต่างๆ กัน เช่น ระดับนัยสำคัญ (significance level) ระดับความเชื่อมั่น (confidence level) ฟังก์ชันการแจกแจงเบื้องต้น θ (prior distribution) และฟังก์ชันการแจกแจงภายหลังของ θ (posterior probability function) เป็นต้น และเมื่อนักสถิติหลายท่านได้พิจารณาความหมายของความน่าจะเป็นจะเป็นในแง่มุมที่แตกต่างกันออกไปด้วย เหตุนี้เองจึงทำให้แนวคิดของการอนุมานทางสถิติแตกต่างกันออกไป ซึ่งพอจะจำแนกออกเป็นกลุ่มๆ ดังนี้

- Frequentist inference (บางครั้งจะเรียกว่า Classical inference)
- Bayesian inference
- Likelihood inference
- Fiducial inference
- Structural inference
- Decision inference

3.2 การประยุกต์การอนุมานทางสถิติเพื่อใช้กำหนดความหมายของคำที่กำกวม

ตัวแบบความน่าจะเป็น (Probabilistic Model) สำหรับในปัญหาการกำหนดความหมายที่ถูกต้องจะใช้เป็นรูปแบบในการแบ่งแยก โดยที่เราจะใช้ตัวแบบนี้ในการจะตัดสินใจว่าคำที่กำลังพิจารณาควรจะมีความหมายแบบใด นอกเหนือจากนั้นจะใช้พารามิเตอร์ (Parameter) ในการอธิบายคุณลักษณะของประชากร แต่ตามปกติเราไม่สามารถที่จะศึกษาถึงประชากรได้อย่างละเอียดถี่ถ้วน จึงจำเป็นที่จะต้องเลือกตัวอย่างข้อมูลจากประชากรแบบสุ่ม ทำให้คุณลักษณะของประชากรจึงต้องเป็นในรูปแบบของการประมาณแทน

เนื่องจากเหตุการณ์ที่สนใจ ณ ขณะนี้คือ การที่ประโยคมีคำที่กำกวมเกิดขึ้น แต่ละประโยคจะแสดงแทนด้วยการรวมกันของค่าของตัวแปรสุ่ม (Random variables) โดยที่ตัวแปรสุ่มถูกแทนด้วยลักษณะ (Feature) ของตัวแปรสุ่มนั้น การขึ้นต่อกันระหว่างลักษณะ (feature) จะถูกพิจารณาจาก Parametric form ของตัวแบบความน่าจะเป็น (Probabilistic model)

เมื่อคุณลักษณะ (Feature) คือ ตัวแปรสุ่ม Feature vecter เป็นตัวอย่างที่เฉพาะเจาะจงของตัวแปรสุ่มนั้น โดยที่แต่ละ Feature vector จะแสดงค่าที่เรารู้ได้ หรือตัวอย่างของเหตุการณ์ เช่น ประโยคที่มีคำที่กำกวม

สมมติว่ามีประโยคทั้งหมด N และแต่ละประโยคประโยคจะมีคำที่กำกวมปรากฏอยู่ เราจะแปลงประโยคนั้นให้ในรูปของ Feature vector ได้ดังนี้

$$(F_1, F_2, \ldots, F_n, S)$$

โดยที่

$F_1, F_2, \ldots, F_3$ แสดงถึงคุณลักษณะที่ถูกเลือกของคำแวดล้อม (Context) ในประโยคที่มีคำที่กำกวมปรากฏอยู่

S แสดงถึงความหมายของคำที่กำกวม

จุดมุ่งหมายคือ การแบ่งประโยค N ประโยคที่มีคำที่กำกวมให้อยู่ตามกลุ่มของความหมาย (Sense Group) ที่กำหนด โดยที่กลุ่มเหล่านี้ก็ต้องมีการจับคู่กับความหมายจริงหลังจากการแบ่งกลุ่มได้ทำเสร็จสิ้นแล้ว

เพราะฉะนั้นแล้วการใช้ตัวแบบความน่าจะเป็น (Probabilistic Model) เพื่อช่วยในการแก้ปัญหาเรื่องของความกำกวมนี้ก็ต้องประกอบด้วย 2 ส่วนได้แก่

- **Parametric Form** ที่จะเป็นรูปแบบที่จะใช้ในการอธิบายการขึ้นต่อกันระหว่างคุณลักษณะ (Feature) ทำให้เราสามารถที่จะกำหนดกลุ่มของความหมายให้กับคำที่กำกวมได้
- **Parameter Estimation** เป็นตัวที่จะบอกว่าแต่ละเหตุการณ์ที่เกิดขึ้นมีความเป็นไปได้อย่างไร

3.2.1 Parametric Form

เราจะใช้ Decomposable Models เพื่อใช้ในการอธิบายการขึ้นต่อกันของแต่ละ Feature ว่าจะเป็นลักษณะแบบไหน ซึ่ง Decomposable Model จะเป็น subset ของ Graphical Model ใน Graphical Model การขึ้นต่อกันของ features อาจจะเป็นแบบ dependent หรือ conditionally independent

Model นี้จะใช้เป็นตัวที่แยกแยะ ในการระบุถึงความหมายที่เป็นไปได้มากที่สุดของคำที่กำกวม ที่ปรากฏใน Context นั้นๆ เช่น ถ้าประโยคที่ประกอบไปด้วยคำที่กำกวม แสดงในรูปของ Feature vector ดังนี้

$$(C = c, V = v, R = r, T = t, S = ?)$$

ตัวแปร S แสดงถึงความหมายของคำที่กำกวม

ตัวแปร C, V, R, T เป็น feature ที่แสดงถึง context ที่มีคำที่กำกวมเกิดขึ้น

ค่าของ Feature อื่นเรารู้ แต่ S เราไม่รู้ถ้าให้ x เป็นค่าที่เป็นไปได้ของ S แต่ละความหมายที่เป็นไปได้จะแสดงแทนด้วย S_x คือ มีเหตุการณ์ที่เป็นไปได้ x เหตุการณ์ที่เกี่ยวข้องกับ Feature vector ที่ไม่สมบูรณ์

เพราะฉะนั้นการลดความกำกวมจะถูกกระทำผ่าน Maximization function โดยให้ค่าสำหรับ contextual feature ที่ทราบ probabilistic classifier จะตัดสินใจที่ค่าของ S

$$S = \overset{\text{argmax}}{s} \; x \; p(s_x \mid c,v,r,t) = \overset{\text{argmax}}{s} \; x \; \frac{p(c,v,r,t,s_x)}{p(c,v,r,t)}$$

ตัวหารจากสมการด้านบนนี้เป็นค่าคงที่เพราะไม่มี S รวมอยู่ด้วย เราสามารถที่จะเอาออก และทำให้การหาค่าง่ายขึ้น สรุปออกมาเป็นการหาค่า S ที่ Maximizes ที่เมื่อพิจารณาร่วมกัน Feature C, V, R, T และ S เป็นดังนี้

$$S = \overset{\arg\max}{s} \; x \; \frac{p(c,v,r,t,s_x)}{p(c,v,r,t)}$$

3.2.2 Parameter Estimation

มีอยู่ 2 วิธีการในการประมาณค่าของ Parameter ได้แก่

- **Maximum Likelihood Estimation (MLE)** จะเป็นวิธีการที่ประกอบไปด้วยการคำนวณความเป็นไปได้ของข้อมูลตามรูปแบบ และหาค่าของ Parameter ที่ทำให้มีค่ามากที่สุด
- **Bayesian Estimation** เป็นวิธีการประมาณค่าของ Parameter อีกแบบหนึ่งที่แตกต่างจากการประเมินแบบ Maximum Likelihood Estimation

3.3 การเรียนรู้จากข้อความ

เนื่องจากในการศึกษาครั้งนี้จะมุ่งเน้นไปตามแนวคิดแบบ Corpus-based เพื่อที่จะเรียนรู้ตัวแบบความน่าจะเป็น (Probabilistic Models) ที่จะใช้แก้ปัญหาความหมายของคำที่กำกวม เพราะตัวแบบเหล่านี้จะบ่งชี้ว่าความหมายของคำที่กำกวมส่วนใหญ่แล้วน่าจะขึ้นอยู่กับคำแวดล้อมที่ปรากฏอยู่ด้วยกัน ทำให้ในงานของการลดความกำกวมจึงประกอบไปด้วยการแยกคำที่กำกวมให้เป็นเพียง 1 หมายความจากหลายๆ ความหมายที่เป็นไปได้

ตัวแบบความน่าจะเป็น (Probabilistic Models) ที่มีการเรียนรู้นี้จะเรียนรู้ผ่านการเรียนรู้อยู่ 2 แนวทางได้แก่

- การเรียนรู้แบบมีผู้สอน (Supervised Learning)
- การเรียนรู้ด้วยตนเอง (Unsupervised Learning)

ถ้าเรามีตัวอย่างของการลดความกำกวมที่พิจารณาโดยคนเพื่อให้เป็นข้อมูลในการฝึก (Training Data) แล้วการเรียนรู้แบบมีผู้สอน (Supervised Learning) จะเป็นวิธีที่มีประสิทธิภาพดี เพราะตัวอย่างจะอยู่ในรูปแบบที่มีการกำหนดความหมายของคำที่กำกวมในแต่ละประโยคที่รวบรวมเป็นจำนวนมากไว้แล้ว (Sense tagged text) ซึ่งการกำหนดความหมายนี้จะกระทำโดยคน และหลังจากนั้นการเรียนรู้แบบนี้จะสร้างตัวแบบที่เป็นลักษณะทั่วๆ ไปจากชุดของตัวอย่าง และใช้ตัวแบบนี้ในการลดความกำกวมกับคำที่ปรากฏในประโยคใหม่ที่ต้องการพิจารณา (Test data)

แต่ถ้าไม่มีตัวอย่าง เราก็จะใช้การเรียนรู้แบบเรียนด้วยตนเอง นอกเหนือจากนั้นการเรียนรู้แบบนี้จะอาศัยเพียงข้อมูลดิบ ที่ยังไม่มีการกำหนดความหมาย (raw of untagged text) กระบวนการแบบเรียนรู้ด้วยตนเองนี้จะแบ่งการใช้คำที่กำกวมเป็นกลุ่มโดยพิจารณาจากคำแวดล้อม

ในการเรียนรู้ทั้ง 2 เทคนิคจะต้องประกอบไปด้วย Parametric form และ Parameter estimates โดยที่ Parametric form จะเป็นตัวที่แสดงว่าคุณลักษณะแวดล้อม (contextual feature) มีผลกับค่าของคุณลักษณะแวดล้อมอื่นได้ดีเหมือนกับคุณลักษณะแวดล้อมมีผลกับความหมายของคำที่กำกวม และ Parameter estimates จะเป็นตัวบอกว่าการรวมกันของค่าสำหรับคุณลักษณะแวดล้อม (contextual features) ที่เกิดร่วมกันมีผลกับความหมายกับคำที่กำกวมอย่างไร

3.4 การเรียนรู้ด้วยตนเอง (Unsupervised Learning)

ด้วยข้อจำกัดของการเรียนรู้ตามแบบมีผู้สอน (Supervised Learning) ที่จะต้องมี Sense-tagged text เพื่อที่จะใช้เป็น Training data จึงเป็นกิจกรรมที่ต้องใช้เวลา และแรงงานคนในการที่จะมากำหนดความหมายมาก และเพื่อที่จะกำจัดความยุ่งยากนี้จึงเกิดการเรียนรู้อีกแบบที่จะเป็นการเรียนรู้ตามแนวคิดแบบความรู้ที่ไม่สมบูรณ์ (Knowledge-lean) จุดประสงค์หลักของการเรียนรู้แบบนี้เป็น

การกำหนดกลุ่มของความหมายของคำที่กำกวม (sense group)โดยที่ไม่ต้องอาศัย Training Data แต่จะเป็นการเรียนรู้จากข้อมูลดิบ (Raw untagged text) โดยที่ข้อมูลดิบในที่นี้จะประกอบไปด้วยคำและเครื่องหมายวรรคตอนที่ปรากฏในประโยคนั้นๆ เท่านั้น

เนื่องจากสร้างกลุ่มของความหมาย (sense group) จะเป็นเพียงการกำหนดกลุ่มของความหมายที่ประโยคนั้นควรจะเป็น โดยพิจารณาจากคุณลักษณะที่เราสามารถรู้ได้จากข้อความ การทำงานจึงจะเป็นแค่การแบ่งแยกกลุ่มของความหมาย แต่ยังไม่ได้กำหนดความหมายกับกลุ่มนั้นๆ ถ้าเกิดจะมีการกำหนด จึงต้องกระทำหลังจากที่มีการสร้างกลุ่มของความหมายไว้แล้ว

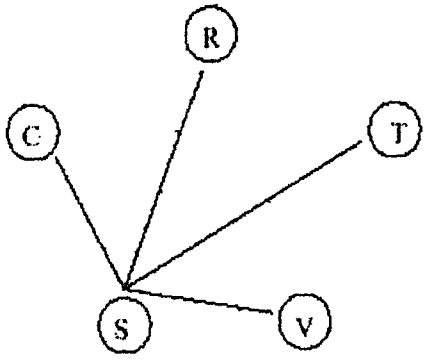
เพราะว่าการเรียนรู้แบบ Unsupervised เป็นแนวคิดแบบ Probabilistic Model ในการทำงานจึงต้องประกอบได้ด้วย 2 องค์ประกอบ ได้แก่ Parametric Form และ Parameter Estimation ซึ่งในที่นี้ Parametric Form จะใช้แบบ Naive Bayes และ Parameter Estimation จะใช้ Maximum Likelihood Estimation with EM Algorithm

3.4.1 Naive Bayes Model

Naive Bayes เป็น Decomposable Model แบบหนึ่งที่กำหนดว่าทุกๆ คุณลักษณะ (Feature) มีความสัมพันธ์แบบอิสระจากกันอย่างมีเงื่อนไข (conditionally independent) เมื่อมาประยุกต์กับการแก้ปัญหาเรื่องความกำกวม จะเป็นแบบที่ทุกๆ contextual features จะให้ความหมายกับคำที่กำกวมอย่างเป็นอิสระจากกัน ทำให้ความน่าจะเป็นของการรวมกันของ contextual features กับความหมายที่ต้องการหาจะเป็นดังนี้

$$p(F_1, F_2, \ldots, F_n, S) = p(S) \prod_{i=1}^{n} p(F_i \mid S)$$

Parameter ของ Model นี้ได้แก่ $p(S)$ และ $P(F_i|S)$ คือ ผลรวมของข้อมูลที่จำเป็นในการประเมินค่าของ parameter โดยที่ทั้ง 2 ค่านี้เป็นความถี่ของเหตุการณ์ที่ได้มาจาก (F,S) แต่เนื่องจากเราไม่ทราบความหมายของคำที่กำกวมจึงได้ใช้ EM Algorithm มาช่วยในการประเมินค่าของ parameter แทน



รูปที่ 3.1 Naive Bayes Model

ตัวอย่างเช่น ต้องการหาความหมายที่แท้จริงของคำว่า ***bill*** โดยการเลือกจาก 2 ความหมาย ดังนี้

- พระราชบัญญัติ
- ใบเสร็จรับเงิน

สมมติว่ากำหนดให้แต่ละประโยคที่มีคำว่า bill มี binary feature อยู่ 5 feature ได้แก่

S เป็นความหมายของคำว่า bill ซึ่งจะเป็นค่าที่เราต้องการหา

ส่วนอีก 4 feature เป็น Contexture feature ที่บ่งบอกว่ามีคำนั้นในประโยคร่วมกับคำที่กำกวมหรือไม่ ได้แก่ คำว่า Congress(C), Veto(V), Restaurant(R), Tip(T) โดยที่จะแสดงเป็นตัวแปร ตามนี้ C, V, R, T และแต่ละค่าของตัวแปรจะเป็น binary ใช่ ถ้ามีคำนั้นอยู่ หรือไม่ใช่ ถ้าไม่มีคำนั้นอยู่

Parametric Form จะแสดงดังนี้ *(CS)(RS)(TS)(VS)* แสดงได้ดังรูปที่ 1

โดยที่การคำนวณตามแบบของ Naive Bayes จะคำนวณตามนี้

$$\hat{p}(c,v,r,t)=\hat{p}(s)\times\hat{p}(c|s)\times\hat{p}(v|s)\times\hat{p}(r|s)\times\hat{p}(t|s)$$

3.4.2 Maximum Likelihood Estimation With EM-Algorithm

จะใช้การประมาณค่า Parameter แบบ Maximum Likelihood Estimation ร่วมกับ EM-Algorithm เนื่องจากค่าของ Parameter ที่จะหาเราไม่สามารถทราบค่าได้ คือ ความหมายของคำที่กำกวม ที่เราจะเรียกว่าเป็น Missing Data

EM Algorithm มี 2 Step

- Expectation (E-Step)
- Maximization (M-step)

E-step จะเป็นคำนวณค่าคาดหวังของทางสถิติของ Model โดยการใช้การประมาณค่าของ Parameter ปัจจุบัน สำหรับในกรณีของ Decomposable model จะเป็นจำนวนความถี่ของเหตุการณ์ที่ Feature เกิดร่วมกับความหมายของคำที่กำกวมในความหมายต่างๆ ที่เป็นไปได้

ส่วน M-step จะทำ Maximum Likelihood Estimates ของ Model parameter โดยใช้ค่าทางสถิติจาก E-Step

เราจะกระทำทั้ง 2 ขั้นตอนนี้จะทำจนกระทั่งการประมาณค่าของ Parameter ในรอบที่ k-1 และรอบที่ k จะแตกต่างน้อยกว่า ε (คือ ค่าๆ หนึ่งที่มีการกำหนดไว้)

ขั้นตอนของการทำ EM Algorithm ใน Naive Bayes Model สามารถทำได้ดังนี้

1. หาค่าเริ่มตัวแบบสุ่ม $p(F_i|S)$ set k = 1
2. E-step : $count(F_i,S) = p(S|F_i) \times count(F_i)$
3. M-Step : ประเมินค่าใหม่

   $p(Fi|S) = count(Fi,S)/count(S)$
4. k = k + 1
5. กลับไป step ที่ 2 ถ้า parameter estimates จาก k และ k-1 ต่างกันมากว่า ε

ตัวอย่าง เป็นการแสดงขั้นตอนของ EM Algorithm ในแบบของ Naive Bayes แบบลำดับขั้น ดังนี้

Feature 1 (F_1)	Feature 2 (F_2)	Sense (S)
1	2	?
1	2	?
2	2	?
2	2	?
1	2	?
1	1	?
1	1	?
1	1	?
1	2	?
2	2	?

ตารางที่ 3.1 ตัวอย่างข้อมูลเพื่อใช้ในกระบวนการการเรียนรู้แบบเรียนรู้ด้วยตนเอง

สมมติว่ามีข้อมูลตัวอย่างที่เหตุการณ์ถูกอธิบายด้วยตัวแปรสุ่มอยู่ 3 ตัวแปรสุ่ม โดยที่ตัวแปร F_1 และ F_2 เป็นค่าที่เราสามารถทราบได้ และมีค่าที่เป็นไปเพียง 2 ค่า (1 กับ 2) ตัวแปร S ใช้แทนกลุ่มของเหตุการณ์ที่เราไม่สามารถทราบค่าได้ เพราะฉะนั้นจากที่ได้กล่าวแล้ว ทำให้เราได้ ตัวแบบของ Parameter เป็น $p(S)$, $P(F_1|S)$ และ $p(F_2|S)$ และข้อมูลตัวอย่างที่ใช้จะมีทั้งหมด 10 เหตุการณ์ดังแสดงในตารางที่ 3.1

1. เนื่องจากยังไม่มีค่าของ S ทำให้ในรอบที่ 1 จะเป็นการกำหนดค่าของ S แบบสุ่ม สมมติถ้าเรากำหนดแบบสุ่ม ดังตารางที่ 3.2
2. E-Step รอบที่ 1 ค่าที่คาดหวัง (Expeted values) ของรูปแบบ Naive Bayes ถูกตัดสินโดยการนับจำนวนเหตุการณ์ในขอบเขตที่กำหนดโดย Naive Bayes นั่นคือ freq(F_1,S) และ freq (F_2,S) ดังที่จะแสดงเป็นตัวอย่างดังนี้

$$freq(F_1=1, S=1) = 3$$
$$freq(F_1=1, S=2) = 2$$
$$freq(F_1=1, S=3) = 2$$
$$freq(F_1=2, S=1) = 1$$
$$freq(F_1=2, S=2) = 2$$
$$freq(F_1=1, S=2) = 0$$

ส่วนค่าของ $freq(F_2=1, S=?)$ และ $freq(F_2=2, S=?)$ ก็จะเหมือนกับ $freq(F_1=1, S=?)$ และ $freq(F_1=2, S=?)$

Feature 1 (F1)	Feature 2 (F2)	Sense (S)
1	2	1
1	2	3
2	2	2
2	2	2
1	2	1
1	1	3
1	1	1
1	1	2
1	2	2
2	2	1

ตารางที่ 3.2 กำหนดค่าของ S แบบสุ่ม

3. M-Step รอบที่ 1 คือการทำ Maximum likelihood estimates สำหรับ parameter ของ Naive Bayes ได้แก่การหาค่าของ $p(S)$, $p(F_1,S)$ และ $p(F_2,S)$ เช่น

   $p(S=1) = 0.4$ ...

   $p(F_1=1|S=1) = 0.75$ ...

   $p(F_2=1|S=1) = 0.25$ ...

   เป็นต้น

4. E-Step รอบที่ 2 หลังจากมีการทำ EM Algorithm รอบแรกเรียบร้อยแล้ว จะต้องมีการหาค่า expected values ของแต่ละเหตุการณ์ (แต่ละรายการ) โดยการกำหนดค่าให้กับ S ใหม่ โดยจะแทนด้วยค่าที่ทำให้ $p(S|F_1, F_2)$ มีค่าสูงสุด

$$S = \underset{s}{\operatorname{argmax}}\ p(S \mid F_1, F_2) = \frac{p(S)xp(F_1 \mid S)xp(F_2 \mid S)}{p(F_1, F_2)}$$

ทำให้แต่ละเหตุการณ์มีการกำหนด S ใหม่ดังตารางที่ 3.3

Feature 1 (F1)	Feature 2 (F2)	Sense (S)
1	2	1
1	2	1
2	2	2
2	2	2
1	2	1
1	1	3
1	1	3
1	1	3
1	2	1
2	2	2

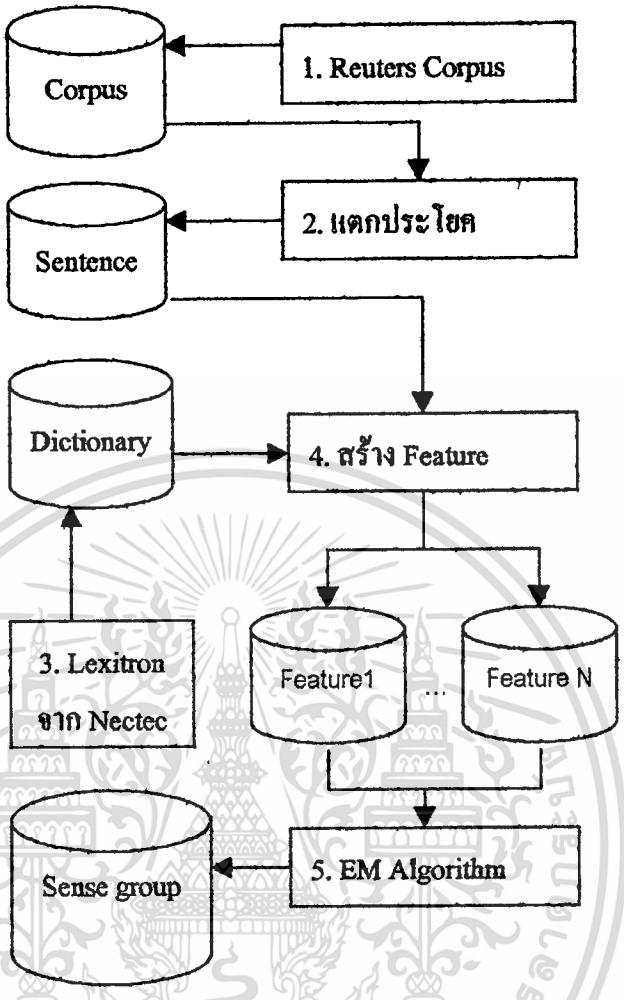
ตารางที่ 3.3 E-Step ในรอบที่ 2

5. ทำงานเหมือนขั้นตอนที่ 2-3 กับ S ใหม่ ดังที่แสดงในตารางที่ 3.3 และเมื่อหาค่าของ parameter estimates ได้แก่ $p(S)$, $p(F_1|S)$ และ $p(F_2|S)$ แล้วให้เปรียบเทียบค่า Parameter ของรอบปัจจุบันกับรอบก่อน (ค่าที่ได้จากข้อที่ 3) ถ้าค่าของ Parameter ในรอบปัจจุบัน และรอบที่ผ่านมามีค่ามากกว่าค่า Threshold ที่กำหนดให้ค่าหนึ่ง ให้วนกลับไปทำงานเหมือนกับขั้นตอนที่ 2-5 จนกว่าค่าของ Parameter ของรอบปัจจุบัน และรอบก่อนหน้าจะมีความแตกต่างน้อยกว่าค่า Threshold ทำให้การทำงานด้วย EM Algorithm ต้องทำงานมากกว่า 1 รอบ

บทที่ 4
ขั้นตอนการสร้างฐานข้อมูล (Corpus)

จากแนวคิดของการเรียนรู้แบบเรียนรู้ด้วยตนเองตามแนวทางทางสถิติ (Unsupervised Learning) เราจะนำมาประยุกต์ในการสร้างฐานข้อมูลเพื่อเป็น Training Data (corpus) โดยวัตถุประสงค์หลักคือ การที่จะสร้างฐานข้อมูล (Corpus) ที่มีการกำหนดความหมายให้กับคำที่กำกวม (sense tagged text) เหมือนกับการที่ถูกสร้างโดยใช้มนุษย์ แต่เพราะแนวคิดของการเรียนรู้แบบนี้เป็นเพียงการแยกกลุ่มของความหมาย (sense group) จึงยังไม่ได้กำหนดความหมายให้แต่ละกลุ่ม หลังจากการเรียนรู้แบบนี้จึงจะต้องใช้มนุษย์เข้าไปกำหนดความหมายให้แต่ละกลุ่มเพื่อให้ได้ตัวอย่างประโยคที่มีการกำหนดความหมายแล้ว และนำไปใช้ประกอบในการพิจารณาความหมายของประโยคใหม่ต่อไป โดยที่ขั้นตอนในการสร้างฐานข้อมูล จะแบ่งเป็นขั้นตอนย่อยดังแสดงในรูปที่ 4.1 ดังนี้

1. รวบรวมข้อมูลที่จะเป็นเอกสารภาษาอังกฤษ สำหรับในการทำงานครั้งนี้ได้ข้อมูลมาจาก Reuters Corpus (Reuters-21578 text categorization Version 1.0) เพื่อใช้เป็นข้อมูลดิบ (raw text)
2. ข้อมูลที่รวบรวมจากข้อที่ 1 จะแตกเอกสารเป็นประโยค เพื่อใช้ในการเรียนรู้แบบเรียนรู้ด้วยตนเอง
3. สร้างพจนานุกรมคำศัพท์ภาษาอังกฤษ-ไทย ซึ่งได้ข้อมูลจากพจนานุกรมคำศัพท์ (Lexitron)
4. จากประโยคที่ได้จากข้อ 2 จะทำการกำหนดคุณลักษณะต่างๆ ของคำแต่ละคำที่มารวมกันอยู่กันประโยคทั้งคำที่มีหลายความหมาย และคำแวดล้อม (contextual feature) ได้แก่
    - หน้าที่ของคำ (Part-of-speech)
    - Morphology
    - Co-Occurrences
    - Unrestricted Collocations
    - Content Collocations
5. พิจารณากลุ่มของความหมายตามวิธีการแบบ EM-Algorithm



รูปที่ 4.1 ขั้นตอนการสร้างฐานข้อมูล

สำหรับในรายละเอียดของการเตรียมข้อมูลในแต่ละขั้นตอนเป็น ดังนี้

4.1 ข้อมูลจาก Reuters Corpus

เนื่องจากการศึกษาในครั้งนี้เป็นความพยายามที่จะกำหนดความหมายที่แท้จริงของคำศัพท์ภาษาอังกฤษ เป็นภาษาไทย ซึ่งอาจจะนำไปใช้เป็นประโยชน์ในเรื่องของการแปลภาษาต่อไปในอนาคต จึงต้องการข้อมูลตัวอย่างเป็นเอกสารภาษาอังกฤษ ที่มีความหลากหลายของรูปแบบการเรียงตัวของคำ กล่าวคือ ต้องเป็นเอกสารที่กล่าวถึงเรื่องราวหลากหลายเรื่อง และที่สำคัญเอกสารที่ได้มาจะเป็นลักษณะของ Raw Text ก็คือ เอกสารนี้จะเป็นเอกสารที่มีแต่การเรียงตัวของคำศัพท์ และเครื่องหมายวรรคตอนโดยที่ยังไม่ได้กำหนดคุณลักษณะใดๆ กับคำในประโยคเหล่านั้น เช่น ความหมาย, หน้าที่ของคำ ฯลฯ

ในที่นี้ได้นำข้อมูลมาจาก Reuters Corpus (Reuters-21578 text categorization Version 1.0) ของ David D. Lewis (http://www.research.att.com/~lewis) ซึ่งเป็น Corpus ที่มีการรวมเอกสารที่เหมาะกับการศึกษาในเรื่องของ Information Retrieval, Machine Learning และการศึกษาในเรื่องอื่นๆ ที่เกี่ยวข้องกับแนวคิดแบบ Corpus-based โดยเอกสารที่รวบรวมมาเป็นข่าวของสำนักพิมพ์ Reuters ในปี 1987 ที่นำมาจัดเป็นหมวดหมู่ รวมทั้งหมด 21,578 เอกสาร และนำเสนออยู่ในรูปแบบ SGML File (มีการกำหนด TAG เป็นแบบ SGML) แต่การกำหนดเป็น TAG นั้นไม่ได้กำหนดความหมายเพียงแต่ TAG เพื่อจัดหมวดหมู่ของเรื่องราวในเอกสาร สำหรับในการศึกษาครั้งนี้เราจึงจะสนใจเฉพาะเนื้อความที่กล่าวถึงข่าวเท่านั้น ดังแสดงตัวอย่างในรูปที่ 3.2 (ในส่วนที่ขีดเส้นใต้)

<REUTERS TOPICS="NO" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="5546" NEWID="3">
<DATE>26-FEB-1987 15:03:27.51</DATE>
<TOPICS></TOPICS>
<PLACES><D>usa</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;F A
&#22;&#22;&#1;f0714&#31;reute
d f BC-TEXAS-COMMERCE-BANCSH 02-26 0064</UNKNOWN>
<TEXT>&#2;
<TITLE>TEXAS COMMERCE BANCSHARES <TCB> FILES PLAN</TITLE>
<DATELINE> HOUSTON, Feb 26 - </DATELINE><BODY>Texas Commerce Bancshares
Inc's Texas
Commerce Bank-Houston said it filed an application with the Comptroller of the Currency in
an effort to create the largest banking network in Harris County. The bank said the network
would link 31 banks having 13.5 billion dlrs in assets and 7.5 billion dlrs in deposits.
Reuter
&#3;</BODY></TEXT>
</REUTERS>

รูปที่ 4.2 ตัวอย่างเอกสารที่ได้จาก Reuters Corpus

4.2 การแตกข้อมูลจาก Reuters Corpus เป็นประโยค

จากเอกสารที่ได้จากข้างต้น จะตัดออกเป็นประโยค แล้วเก็บไว้ในฐานข้อมูล เพื่อรอประมวลผลต่อไป โดยที่ฐานข้อมูลที่ใช้เก็บประโยคจะมีโครงสร้างเป็นแบบ Relational ดังนี้

Table Name : Sentence_hdr

Table Description : ข้อความในแต่ละประโยค

Field Name	Description	Data type
sentence_id	ลำดับของประโยค	Number
sentence	ข้อความในแต่ละประโยค	String

ตารางที่ 4.1 แสดงโครงสร้างของ Table Sentence_hdr

Table Name : Sentence_dtl

Table Description : รายละเอียดของคำในแต่ละประโยค

Field Name	Description	Data type
sentence_id	ลำดับของประโยค	Number
pos_of_sentence	ตำแหน่งของคำในประโยค	Number
word_original	คำตามรูปแบบที่ปรากฏในประโยค	String
word_stem	รากศัพท์ของ word_original	String
word_pos	หน้าที่ของคำในประโยค	String

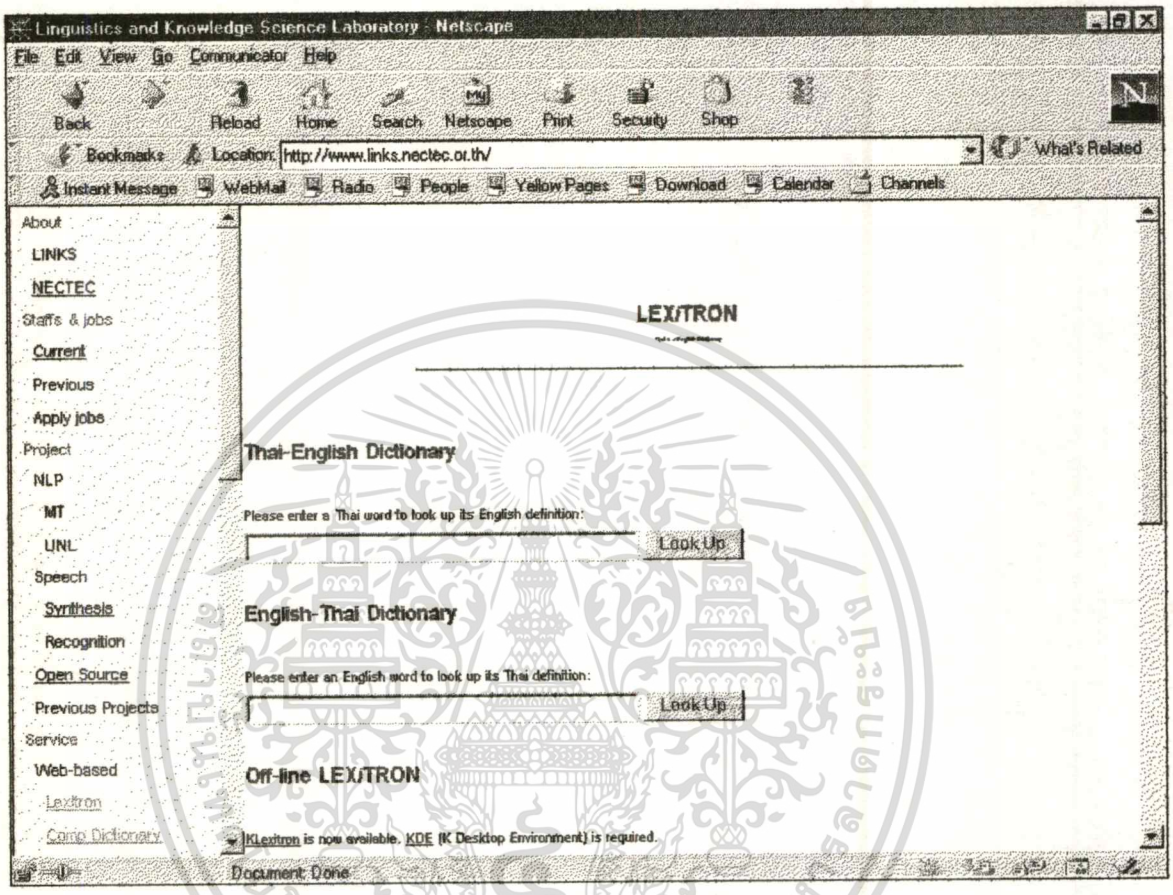
ตารางที่ 4.2 แสดงโครงสร้างของ Table Sentence_dtl

ซึ่งนอกเหนือจากการแตกข้อความออกเป็นประโยคแล้ว สิ่งที่ต้องคำนึงอีกส่วนก็คือ การกำหนดหน้าที่ของคำให้แต่ละประโยค ในที่นี้จะใช้โปรแกรม TreeTagger ที่เป็นโปรแกรมสำหรับการกำหนดความหมายของมหาวิทยาลัยสตุ๊ดการ์ด ประเทศเยอรมัน สามารถ Download ได้จาก http://www.ims.uni-stuttgart.de/projekte/corplex/DecisionTreeTagger.html

4.3 สร้างพจนานุกรมคำศัพท์ภาษาอังกฤษ-ไทย

ในการศึกษาครั้งนี้ได้ข้อมูลจากความหมายของคำศัพท์จาก Lexitron ที่พัฒนาโดยกลุ่มวิจัยภาษาและวิทยาการความรู้ (Links) ของศูนย์เทคโนโลยีอิเลกโทรนิกส์ และคอมพิวเตอร์แห่งชาติ (NECTEC) เพื่อใช้ในการกำหนดความหมาย ได้ว่าคำที่มีความหมายกำกวมควรจะแบ่งเป็นกี่กลุ่ม และก็จะใช้ในการกำหนดความหมายให้แต่ละกลุ่มด้วย ใน Lexitron จะมีการกำหนดความหมายของคำตามหน้าที่ของคำ (Part-of-speech) ต่างๆ และแต่ละความหมายก็จะมีตัวอย่างการใช้ ซึ่งเรา

สามารถที่จะเข้าไปใน web site http://www.links.nectec.or.th ได้ โดยที่คำศัพท์ที่ถูกจัดเก็บใน Lexitron ที่เป็นศัพท์จากภาษาอังกฤษเป็นภาษาไทย มีอยู่ทั้งหมด 8,973 คำ



รูปที่ 4.3 แสดง Web Site ของพจนานุกรมคำศัพท์ LexiTron

4.4 การกำหนดคุณลักษณะของคำในประโยค

การกำหนดว่าจะใช้คุณลักษณะใด (Feature) ใดบ้างที่จะช่วยในการกำหนดความหมายของคำที่กำกวมตาม (Bruce and Wiebe) ซึ่งในที่นี้จะใช้คุณลักษณะทั้งหมด 5 แบบด้วยกัน ได้แก่

- **Morphology** (Feature M) เป็นคุณลักษณะที่ใช้แสดงถึงหน้าที่ของคำที่กำกวมในประโยคนั้น เช่น
  - คำนาม (Noun) ซึ่งจะกำหนดเป็นลักษณะของ Binary feature ได้แก่ เป็นคำนามที่เป็น Single หรือ คำนามที่เป็น Plural
  - คำกิริยา (Verb) เป็นการบ่งบอกถึง Tense ของคำกิริยา ในที่นี้จะกำหนด 7 Tense ได้แก่
  - นอกเหนือจากคำนาม และกริยา ถือว่าเป็นอื่นๆ

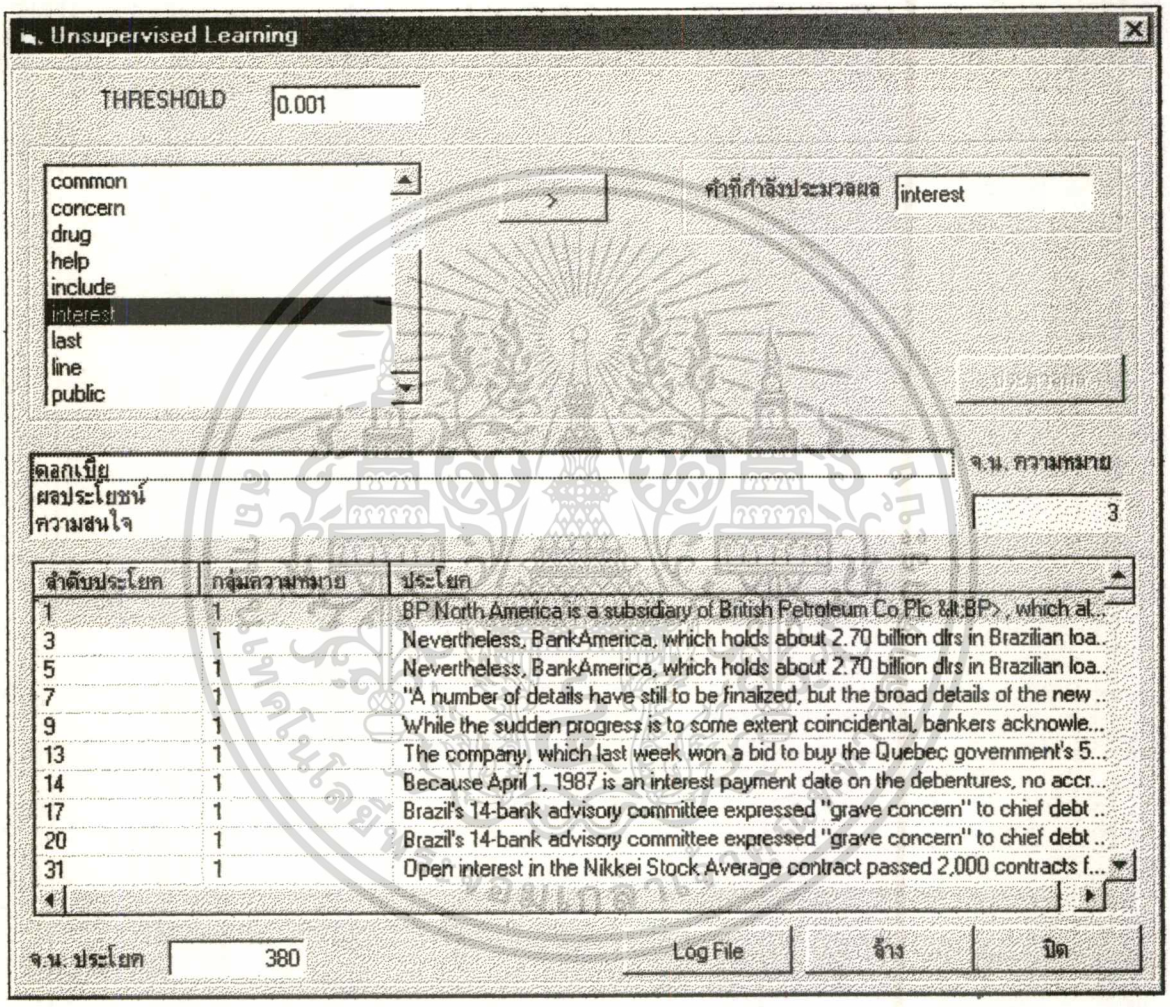
- **Part of Speech** (Feature P) เป็นคุณลักษณะที่ใช้แสดงถึงหน้าที่ของคำของคำที่ตำแหน่งที่ i ทางด้านซ้าย และขวาของคำที่กำกวม สำหรับการพิจารณาในที่นี้จะพิจารณาเฉพาะตำแหน่งที่ 1 และ 2 เท่านั้น นอกจากนั้นหน้าที่ของคำที่สนใจ เราจะสนใจหน้าที่ของคำเพียง 5 หน้าที่ ได้แก่ คำนาม (Noun), คำกิริยา (Verb), คำคุณศัพท์ (Adjective), คำวิเศษณ์ (adverb) และคำที่นอกเหนือจากที่กล่าวไป จะถือว่าเป็นคำอื่นๆ เพราะฉะนั้นคุณลักษณะนี้จะแสดงเป็น P
- **Co-Occurrences** (Feature CF) เป็นคุณลักษณะที่ใช้แสดงถึงคำที่เป็น Content word ซึ่งหมายถึงคำที่มีหน้าที่ของคำอยู่ในกลุ่มของ คำนาม (Noun), คำกิริยา (Verb), คำคุณศัพท์ (Adjective), คำวิเศษณ์ (Adverb) และคำสรรพนาม (Pronoun) ที่มีความถี่สูงสุด 3 ลำดับ ที่มีอยู่ร่วมกับคำที่กำกวมที่ตำแหน่งใดก็ได้ในประโยค โดยที่คุณลักษณะนี้จะแสดงในรูปแบบ Binary Feature (มี/ไม่มี) และใช้สัญลักษณ์แทนด้วย CF1 , CF2 , CF3 เพื่อแทน Content word ที่พบมากที่สุดในประโยคเดียวกับคำที่กำกวมลำดับที่ 1, 2 และ 3 ตามลำดับ
- **Unrestricted collocation** (Feature UC) เป็นคุณลักษณะที่ใช้แสดงถึงคำที่ภายในตำแหน่งที่ +2 และ –2 จากคำที่กำกวม โดยที่ Feature นี้จะกำหนดเป็นค่าที่เป็นไปได้ 21 เพราะฉะนั้นจะทำให้มี
  - 19 ค่าที่เป็นค่าในตำแหน่งที่กำหนดให้เป็นคำที่เจอสูงสุด
  - ค่า null ที่จะบอกว่าตำแหน่งที่ i จากด้านซ้าย และด้านขวาเป็นขอบของประโยค
  - ค่า none จะเป็นคำที่นอกเหนือจากคำ 19 คำที่กำหนดไว้

  โดยที่คุณลักษณะนี้จึงแทนด้วยสัญลักษณ์ UC_{-2}, UC_{-1}, UC_{+1}, UC_{+2}
- **Content Collocation** (Feature CC) เป็นคุณลักษณะที่ใช้แสดงถึง Content word ซึ่งหมายถึงคำที่มีหน้าที่ของคำอยู่ในกลุ่มของ คำนาม (Noun), คำกิริยา (Verb), คำคุณศัพท์ (Adjective), คำวิเศษณ์ (Adverb) และคำสรรพนาม (Pronoun) ที่เกิดขึ้นในตำแหน่งที่ +1 หรือ –1 จากคำที่กำกวม จะมีลักษณะคล้ายๆ กับ Unrestricted Collocation คือ จะมีค่าที่เป็นไปได้ 21 ค่า ทำให้แบ่งแยกได้ดังนี้
  - 19 ค่าที่เป็นค่าในตำแหน่งที่กำหนด และเป็น Content word 19 ค่าสูงสุด
  - ค่า null ที่จะบอกว่าตำแหน่งที่ i จากด้านซ้าย และด้านขวาเป็นขอบของประโยค
  - ค่า none จะเป็นคำที่นอกเหนือจากคำ 19 คำที่กำหนดไว้

  โดยที่คุณลักษณะนี้จะแทนด้วยสัญลักษณ์ CC_{-1}, CC_{+1}

4.5 พิจารณากลุ่มของความหมายด้วยวิธีการแบบ EM-Algorithm

ในการศึกษาครั้งนี้จะทดสอบด้วยโปรแกรม Visual Basic Version 6.0 ดังแสดงในรูปที่ 4.4 ที่จะเป็นการนำ EM-Algorithm ที่ได้นำเสนอวิธีการมาแล้วในบทที่ 3 ใช้กับตัวอย่างประโยคที่ได้มาจาก Reuters Corpus



รูปที่ 4.4 แสดงโปรแกรมการเรียนรู้ด้วยตนเอง

4.6 ผลของการศึกษา

ก่อนที่จะมีการประมวลผล จะต้องมีทราบถึงข้อจำกัดของการเรียนรู้ด้วยตนเอง ก่อน ได้แก่

1. ผลที่ได้คือ ความสามารถในการแบ่งกลุ่มของความหมายเท่านั้น
2. ต้องมีการกำจัดความกำกวมของหน้าที่ของคำแล้ว
3. ต้องสามารถดึงคุณลักษณะ (Feature) ได้อย่างอัตโนมัติจากข้อความ (raw text)

ความสามารถที่จะแบ่งประโยคออกเป็นกลุ่มความหมาย (Sense group) ตามจำนวนความหมายที่เป็นไปได้ กล่าวคือ จะกำหนดว่าแต่ละประโยคที่มีคำที่กำกวมอยู่ควรจะอยู่ในกลุ่มความหมายใด แต่สุดท้ายของการทำงานจะต้องใช้มนุษย์เข้าไปตัดสินความหมายของกลุ่มความหมายแต่ละกลุ่มที่ได้มาอีกทีหนึ่ง

ตัวอย่าง เช่น การเรียนรู้ถึงคำว่า "interest" ในกรณีมีหน้าที่ของคำในประโยค คือ เป็นคำนาม (Noun) ถ้าเราตรวจสอบในพจนานุกรมคำศัพท์ใน Lexitron พบว่ามีความหมายของคำว่า interest ที่ทำหน้าที่เป็นคำนาม อยู่ 3 ความหมาย ได้แก่

- ดอกเบี้ย
- ผลประโยชน์
- ความสนใจ

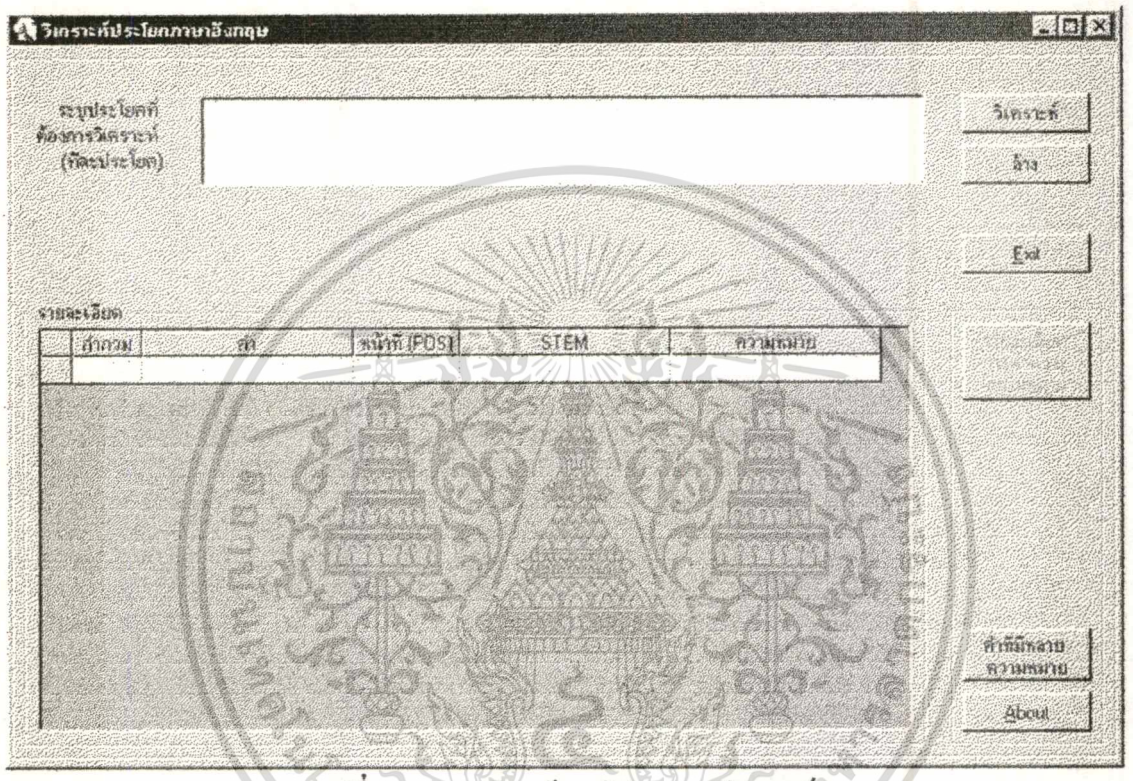
ในการประมวลผลจะใช้ทดสอบที่ค่า THRESHOLD = 0.001 ซึ่งค่า THRESHOLD นี้เป็นค่าที่จะกำหนดรอบของขั้นตอนการประมวลผลตามวิธีการของ EM Algorithm คือถ้าค่าของ Parameter ในรอบปัจจุบัน กับรอบก่อนหน้าถ้ามีความแตกต่างน้อยกว่าค่า THRESHOLD ก็จะหยุดการทำงาน และสรุปผลตามความหมายสุดท้ายที่ได้

จากจำนวนตัวอย่างประโยคที่พบคำว่า interest ที่ทำหน้าที่เป็นคำนามเกิดร่วมอยู่ด้วย 380 ประโยค ด้วยวิธีการตามทฤษฎีที่ได้กล่าวไว้แล้วในบทที่ 3 สามารถที่จะแบ่งแยกประโยค 380 ประโยคนี้ออกเป็น 3 กลุ่มความหมาย คือ กลุ่ม 1, กลุ่ม 2 และกลุ่ม 3

เมื่อเสร็จสิ้นกระบวนการวิเคราะห์ของ EM Algorithm แล้วก็จะทำการกำหนดความหมายให้กับกลุ่มโดยคน แล้วจัดเก็บตัวอย่างประโยคพร้อมกับความหมายที่มีการกำหนดลงในฐานข้อมูล ในที่นี้จะประมวลผลกับคำศัพท์ทั้งสิ้น 8,973 คำ ตามที่ได้คำศัพท์มาจากพจนานุกรมคำศัพท์ Lexitron

บทที่ 5

โปรแกรมวิเคราะห์คำภาษาอังกฤษ



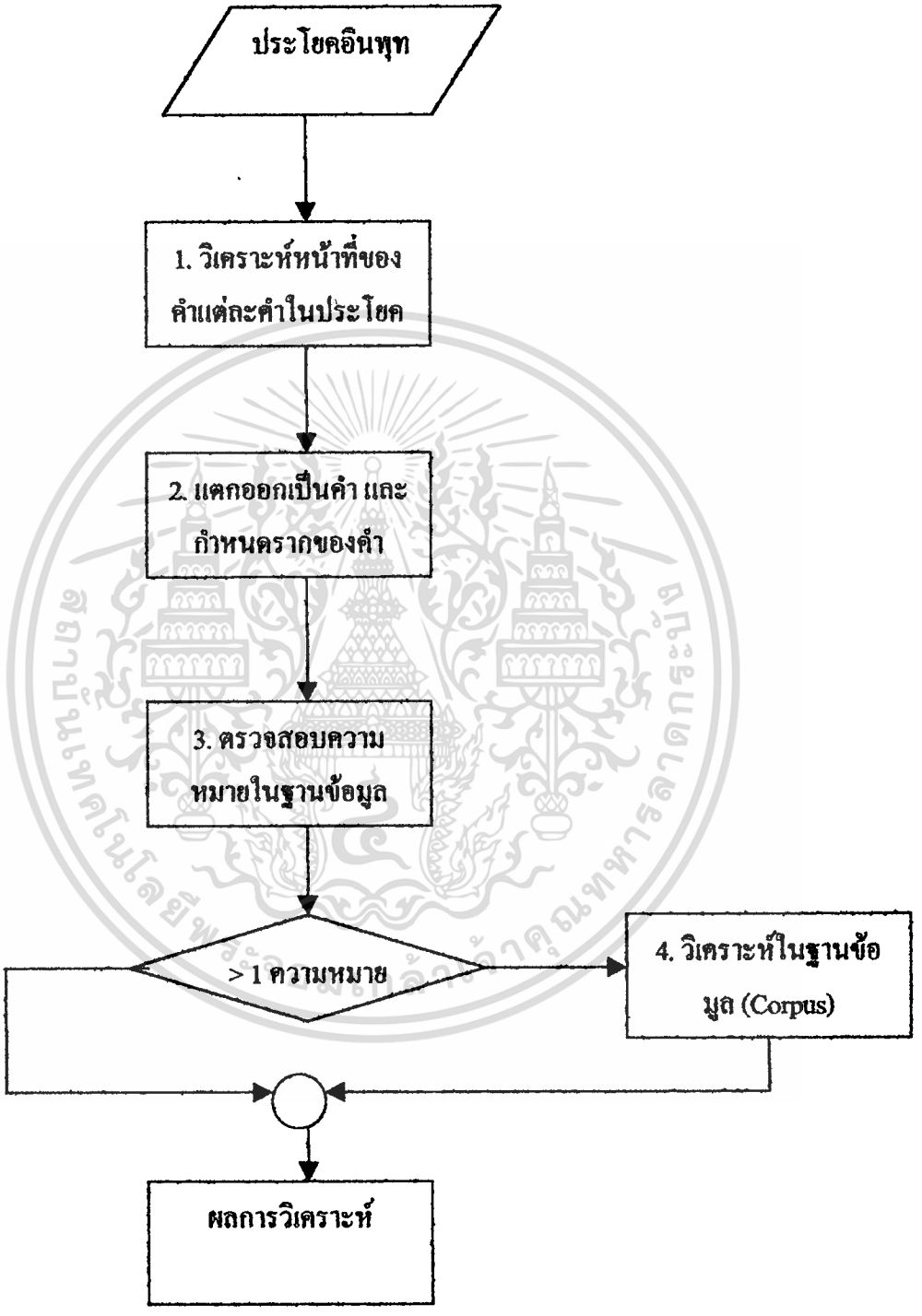
รูปที่ 5.1 แสดงตัวหน้าจอโปรแกรมประยุกต์

จากฐานข้อมูลที่ได้จากบทที่ 4 ที่ได้กล่าวมาแล้ว จะได้นำมาเป็น Training Data สำหรับวิเคราะห์ Testing Data หรือข้อมูลใหม่ที่สนใจ โดยสร้างเป็นโปรแกรมวิเคราะห์คำภาษาอังกฤษ เพื่อใช้ในการวิเคราะห์ความหมาย และหน้าที่ของคำอย่างง่ายๆ

5.1 องค์ประกอบของโปรแกรมวิเคราะห์ความหมาย

- ฐานข้อมูลคำศัพท์ เช่นเดียวกันกับตอนสร้างฐานข้อมูล เราจะใช้ Lexitron ของ NECTEC
- ฐานข้อมูล (Corpus) ที่มีข้อมูล Sense Tagged Text ได้แก่ คำที่มีการกำหนดความหมายแล้ว
- โปรแกรมที่ใช้ในการกำหนดหน้าที่ของคำ (Part-of-Speech)

5.2 ขั้นตอนการทำงานของโปรแกรมวิเคราะห์ความหมาย



รูปที่ 5.2 แสดงผังการทำงานของโปรแกรม

จากผังการทำงานในรูปที่ 5.2 จะเห็นได้ว่า หลังจากที่รับประโยคที่ต้องการวิเคราะห์เข้ามาแล้ว จะต้องทำงานตามขั้นตอนต่างๆ ดังนี้

1. วิเคราะห์หน้าที่ของคำแต่ละคำในประโยค

สำหรับในการวิเคราะห์หน้าที่ของคำ (Part-of-Speech) ได้ใช้โปรแกรม Rule-based Tagger V1.14 ของ Eric Bill แห่งสถาบัน M.I.T. (Massachusetts Institute of Technology) ที่เป็นโปรแกรมสำหรับวิเคราะห์หน้าที่ของคำ และทำงานบน Platform UNIX เนื่องจากโปรแกรมที่สร้างขึ้นมาในครั้งนี้เป็นโปรแกรมที่ทำงานอยู่บน Platform Windows จึงได้นำเอามาดัดแปลงเพื่อให้สามารถประมวลผลผ่าน Platform Window ได้

วิธีการในการวิเคราะห์ของตัว Tagger นี้จะมีการทำงานอยู่ 2 ขั้นตอนหลัก ได้แก่

- กำหนดหน้าที่ที่น่าจะเป็นไปได้ จาก Training Corpus (ที่มีการกำหนดหน้าที่ของคำ)
- พิจารณาอีกรอบเพื่อเปลี่ยนหน้าที่จาก Contextual cues

2. แตกออกเป็นคำ และกำหนดรากของคำ

หลังจากที่มีการกำหนดหน้าที่ของคำ เราจะทำการแตกประโยคออกเป็นคำๆ แล้วเปลี่ยนรูปของคำจากรูปที่ไม่ได้เป็นรูปปกติ เป็นรากของคำแต่ละ (word stemming) เนื่องจากข้อมูลของความหมายจะจัดเก็บเป็นรากของคำ ไม่มีการจัดเก็บเป็นรูปอื่นของคำ

3. ตรวจสอบคำแต่ละคำใน Lexitron

เนื่องจากพจนานุกรมคำศัพท์ (Lexitron) จะเก็บอยู่ในรูปฐานข้อมูล (Relational DBMS) จึงจะเป็นการค้นหาคำศัพท์ที่ต้องการในฐานข้อมูล เพื่อตรวจสอบว่ามีคำใดบ้างที่มีความหมายมากกว่า 1 ความหมาย ซึ่งในการพิจารณาต้องพิจารณาประกอบกับหน้าที่ของคำด้วย เพราะจะเป็นการกรองความหมายได้อย่างหนึ่งสำหรับโครงสร้างของฐานข้อมูลที่ใช้เก็บข้อมูลพจนานุกรม จะมีโครงสร้างดังรูปที่ 5.3

4. วิเคราะห์ข้อมูลในฐานข้อมูล (Corpus)

เพื่อเป็นการสร้างการวิเคราะห์อย่างง่ายๆ เราจะใช้แนวความคิดตามแบบของ Corpus-based approach คือ การสร้างวิเคราะห์จากคำที่อยู่แวดล้อมจากประโยคอินพุท กับประโยคที่อยู่ใน Training Corpus จากบทที่ 4 ว่ามีความน่าจะเป็นความหมายใดมากกว่ากัน

Field Name	Description	Data type
word_seq	ลำดับที่ของคำ	Number
word	คำศัพท์ (ภาษาอังกฤษ)	String
word_pos	หน้าที่ของคำ	String
word_meaning	ความหมายของคำเมื่อมีหน้าที่ตามที่กำหนด	String
word_use_exam	ตัวอย่างประโยคที่ทำให้มีความหมายนี้	String

ตารางที่ 5.1 แสดงโครงสร้างของ Table ที่ใช้เก็บข้อมูล Lexitron

ซึ่งผลของการทำงาน คือ จะเป็นการวิเคราะห์แต่ละแต่ละคำในประโยคว่าจะมีหน้าที่ของคำ รากของคำ และความหมายของคำเป็นอย่างไร รวมทั้งแสดงว่าคำๆ ไหนที่เป็นคำที่ก่อให้เกิดความกำกวม

5.3 ข้อจำกัดของประสิทธิภาพของโปรแกรม

- Reuters Corpus เป็นฐานข้อมูลขนาดเล็ก ทำให้อาจจะไม่มีคำศัพท์ไม่ครบ รวมถึงถ้ามีคำศัพท์ แต่ก็อาจจะไม่มีตัวอย่างประโยคที่ให้ความหมายที่เราต้องการก็ได้
- คำศัพท์ได้มาจากพจนานุกรม Lexitron ซึ่งกำลังอยู่ในช่วยพัฒนาทำให้ทั้งคำศัพท์ และความหมายอาจจะไม่ครบถ้วน
- เกิดมาจากตัวคนที่เข้าไปตัดสินความหมายของคำเอง ถ้าตัดสินผิดก็จะส่งผลถึงการที่จะทำให้ Training Data มีความผิดพลาดได้เหมือนกัน

บทที่ 6
บทสรุป และแนวทางการพัฒนา

การพัฒนา และปรับปรุงเทคนิคการเรียนรู้แบบเรียนรู้ด้วยตนเอง (Unsupervised Learning) เป็นประเด็นที่สำคัญในกระบวนการประมวลผลภาษาธรรมชาติ เพราะวิธีนี้ได้เข้ามากำจัดการเรียนรู้แบบมีคนสอน (Supervised Learning) ในแง่ของการที่จะได้มาซึ่งความรู้ ที่ได้มาจากคน โดยที่คนจะมีภาระในการกำหนดความหมายของคำที่กำกวม เพื่อรองรับการรับการเป็น Training Data ที่มีประสิทธิภาพ แต่วิธีแบบเรียนรู้ด้วยตนเอง จะเป็นแนวคิดในการแบ่งแยกความหมายโดยซึ่งอยู่กับคุณลักษณะต่างๆ ของคำที่กำกวมเอง รวมถึงคำแวดล้อมคำที่กำกวมด้วย

เนื่องจากตัวแบบความน่าจะเป็นของการลดความกำกวมของความหมายถูกเรียนรู้จาก Raw text โดยทำเหมือนกับว่าคำที่กำกวม (Ambiguous word) เป็น Missing Data การเรียนรู้จึงถูกจำกัดกับการประมาณค่าของ Parameter ในการศึกษา รวมถึงการทดลองจึงสมมติว่า Parametric form เป็น Naive Bayes และใช้การประมาณค่าของ Parameter เป็น EM Algorithm ซึ่งในความเป็นจริงแล้วนอกเหนือจากการใช้ EM Algorithm แล้วเรายังสามารถที่จะใช้วิธีการอื่นที่ทำงานได้คล้ายกับ EM คือ Gibbs Sampling

สำหรับในการศึกษาครั้งนี้จะเป็นการทดสอบประสิทธิภาพของการเรียนรู้ด้วยตนเอง (Unsupervised Learning) ร่วมกับการตัดสินโดยคน เพื่อที่จะใช้สร้างเป็นฐานข้อมูล (Corpus) ที่มีการกำหนดความหมายแบบอัตโนมัติ

ส่วนแนวทางที่จะต้องพัฒนาต่อไปในอนาคต จะแจกแจงเป็น 2 เรื่อง ด้วยกัน คือ

1. การสร้าง Corpus แบบ Unsupervised

- เพิ่มฐานข้อมูลดิบให้มีความหลากหลายขึ้น เพื่อให้ประโยคตัวอย่างครอบคลุมทุกความหมาย
- การกำหนดความหมายให้กับกลุ่มของความหมาย (Sense group) เนื่องจากวิธีการเรียนรู้ด้วยตนเอง จะไม่มีการเชื่อมโยงระหว่างคำ และความหมายของคำ ฉะนั้นกลุ่มของความหมายที่สร้างจะไม่มีความหมายกำกับอัตโนมัติ
- การเลือกคุณลักษณะ (Feature) ในการพิจารณาความหมาย เราจะรู้ได้อย่างไรว่าคุณลักษณะใดมีความสำคัญมีความสำคัญมากกว่ากัน เมื่อมีการใช้เป็น Raw Text ค่าที่ถูก

เลือกจะขึ้นอยู่กับความถี่ที่เกิดขึ้น ผลก็คือบางทีคุณลักษณะที่ไม่ค่อยเกี่ยวข้องจะกลายเป็นคุณลักษณะที่สำคัญไป

- ความกำกวมของหน้าที่ของคำ เนื่องจากในการประมวลผลตามวิธีแบบมีผู้สอน (Supervised Learning) มีข้อจำกัดอยู่อย่างหนึ่ง คือ ก่อนหน้าที่จะมาตัดสินเรื่องของความหมายจะต้องแก้ไขเรื่องความกำกวมของหน้าที่ของคำมาก่อนหน้านี้แล้ว แต่เนื่องจากการเรียนรู้ด้วยตนเอง (Unsupervised Learning) ไม่ได้เรียนรู้จากตัวอย่างที่คนพิจารณาไปแล้ว ทำให้อาจจะยังเกิดปัญหาที่ยังคงมีความกำกวมอยู่ ซึ่งแนวทางที่สามารถนำมาใช้ได้ คือ การใช้ Rule based part-of-speech tagger แต่อย่างไรก็ตามประสิทธิภาพของการทำงานก็ยังคงเป็นสิ่งที่คลุมเครืออยู่

2. โปรแกรมประยุกต์ที่ใช้สร้างเพื่อทดสอบประสิทธิภาพของ Corpus

- สร้างกฎตามหลักไวยากรณ์ภาษาอังกฤษ เพื่อให้สามารถแปลความหมายของประโยคได้

บรรณานุกรม

อโนทัย ตรีวานิช. 2539. ทฤษฎีการอนุมานทางสถิติ. ขอนแก่น : นานาวิทยา.

Bruce, Rebecca and Wiebe, Janyce. n.d. "Word-Sense Disambiguation Using Decomposable Models". n.p.

Bruce, Rebecca and Pedersen, Ted. 1998. "Knowledge lean Word-Sense Disambiguation". In preceedings of Natural Conference Artificial Intelligence'1998.

Hogg, Robert V. and Tanis, Elliot A. 1993. 4th ed. Probability and Statistical Inference. Macmillan.

Mitchell , Tom M. 1997. **Machine Learning.** Singapore: McGraw-Hill.

Pedersen, Ted . 1998. "Learning Probabilistic models of word sense disambiguation".

ประวัติผู้เขียน

ชื่อ นางสาว อรวรรณ โชติกิจนุสรณ์

วันเกิด 25 พฤศจิกายน พ.ศ. 2518

ประวัติการศึกษา

ระดับประถมศึกษา	โรงเรียนศุภวรรณ กรุงเทพมหานคร
ระดับมัธยมศึกษาตอนต้น	โรงเรียนสตรีวัดระฆัง กรุงเทพมหานคร
ระดับมัธยมศึกษาตอนปลาย	โรงเรียนสตรีวัดระฆัง กรุงเทพมหานคร
ระดับบัณฑิตศึกษา (ปริญญาตรี)	มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพมหานคร

ประวัติการทำงาน

-