

บทที่ 3

วิธีการดำเนินงานวิจัย

3.1 องค์ประกอบที่ใช้

องค์ประกอบที่ใช้ในการดำเนินงานวิจัยนี้ประกอบด้วยขั้นตอน 4 ส่วนด้วยกัน คือ

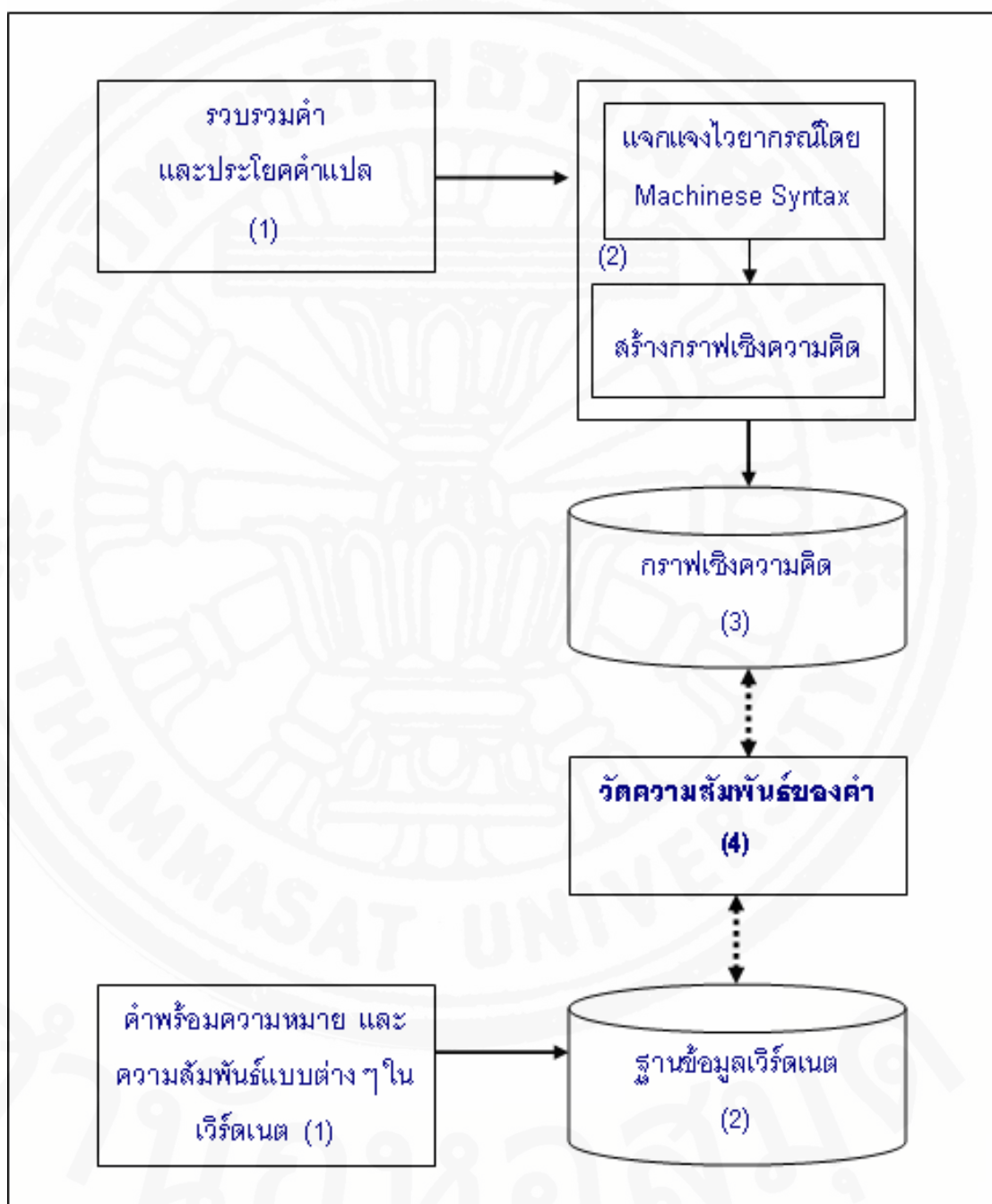
ขั้นแรกเป็นการเตรียมข้อมูล ซึ่งประกอบด้วยการรวบรวมคำพร้อมประโยคคำแปลของคำนั้น ๆ ที่จะใช้ในการทดลอง

ขั้นตอนที่สองเป็นการเตรียมเครื่องมือที่ใช้ ซึ่งในงานวิจัยนี้จะใช้เวิร์ดเน็ตเป็นฐานความรู้โดยการแปลงข้อมูลจากเวิร์ดเน็ตเข้าจัดเก็บในระบบฐานข้อมูลตามโครงสร้างที่ได้ออกแบบไว้ เครื่องมือที่ใช้อีกส่วนคือโปรแกรมเครื่องมือในการสร้างกราฟเชิงความคิด

ขั้นตอนที่สามเป็นการแทนความหมายของประโยคด้วยกราฟเชิงความคิดโดยใช้โปรแกรมเครื่องมือในขั้นตอนที่สอง และจัดเก็บกราฟที่ได้ในระบบฐานข้อมูล

ขั้นตอนที่สี่เป็นการวัดค่าความสัมพันธ์ของคำ โดยการเปรียบเทียบกราฟเชิงความคิดที่ใช้แทนความหมายของประโยคของคำนั้น ๆ

ภาพที่ 3.1
แสดงองค์ประกอบของงานวิจัย



3.2 ขั้นตอนการเตรียมข้อมูล

งานวิจัยนี้เสนอการเปรียบเทียบหาความสัมพันธ์ในเชิงความหมายของคำสองคำโดยพิจารณาความหมายในระดับประโยค ค่าความสัมพันธ์ในเชิงความหมายนั้นจะได้จากผลของการเปรียบเทียบกันของประโยคซึ่งเป็นคำแปล (glossary) ของคำทั้งสอง ดังนั้นในการเตรียมข้อมูลจึงต้องรวบรวมคำพร้อมคำแปลของคำนั้น ๆ ด้วย โดยชุดของคำพร้อมประโยคคำแปลที่ทดลองในงานวิจัยนี้ได้แบ่งออกเป็นสองส่วนดังนี้

1. กลุ่มคำที่เขียนเหมือนกันแต่ความหมายแตกต่างกันตามโดเมน และกลุ่มคำที่มีความหมายเหมือนกัน เพื่อเปรียบเทียบผลการหาความสัมพันธ์ในเชิงความหมายที่ได้ระหว่างกลุ่มคำดังกล่าว
2. ชุดคำที่เขียนต่างกันและคำแปลจำนวน 30 คู่ เพื่อเปรียบเทียบลำดับผลของค่าความสัมพันธ์ที่ได้กับงานวิจัยอื่น ซึ่งประกอบด้วยชุดคำดังนี้

ตารางที่ 3.1

แสดงคู่ลำดับคำที่ใช้ทดลอง

ลำดับที่	คำที่ 1	คำที่ 2
1	car	automobile
2	Gem	jewel
3	Journey	voyage
4	Boy	lad
5	Coast	shore
6	Asylum	Madhouse
7	Magician	wizard
8	Midday	Noon
9	Furnace	Stove
10	Food	Fruit
11	Bird	Cock
12	Bird	crane
13	Tool	Implement

ตารางที่ 3.1 (ต่อ)

ลำดับที่	คำที่ 1	คำที่ 2
15	Crane	Implement
16	Lad	Brother
17	Journey	Car
18	Monk	Oracle
19	cemetery	woodland
20	Food	Rooster
21	Coast	Hill
22	Forest	Graveyard
23	shore	woodland
24	Monk	Slave
25	Coast	Forest
26	Lad	wizard
27	Chord	Smile
28	Glass	Magician
29	Noon	String
30	Rooster	voyage

คำแปลของคำทั้งสองส่วนนี้รวบรวมจาก Wikipedia ซึ่งเป็นออนไลน์ดิคชันนารี โดย
 ประโยคคำแปลของแต่ละคำมีความยาวไม่เกิน 5 ประโยค ซึ่งรายละเอียดคำแปลของคำที่ใช้ใน
 การทดลองนี้แสดงในภาคผนวก ก ตารางคู่ลำดับและคำแปล

3.3 ขั้นตอนการเตรียมเครื่องมือ

โปรแกรมเครื่องมือที่ใช้ในการดำเนินงานวิจัยได้แก่

3.3.1 ฐานข้อมูลเวิร์ดเน็ต

การจัดเตรียมฐานข้อมูลเวิร์ดเน็ตโดยทำการแปลงมูลเวิร์ดเน็ตซึ่งถูกจัดเก็บอยู่ในลักษณะที่เป็นไฟล์ข้อความเข้าจัดเก็บในระบบฐานข้อมูล ซึ่งเครื่องมือและซอฟต์แวร์ที่ใช้ในขั้นตอนการเตรียมฐานข้อมูลเวิร์ดเน็ตมีรายละเอียดดังต่อไปนี้

ตารางที่ 3.2

แสดงรายละเอียดของเครื่องมือและซอฟต์แวร์ในการเตรียมฐานข้อมูลเวิร์ดเน็ต

เวิร์ดเน็ต	WordNet Version 2.1
ฐานข้อมูล	Oracle9i Release 9.0.2
ซอฟต์แวร์ที่ใช้ในการพัฒนาโปรแกรม	PL/SQL Oracle SQL*Loader
ซอฟต์แวร์อื่น ๆ	Microsoft Excel ในการแบ่ง segment ของข้อมูล Text file ใน WordNet

ข้อมูลของเวิร์ดเน็ตนั้นถูกจัดเก็บในรูปแบบของแฟ้มข้อมูล (Text file) ซึ่งประกอบด้วย

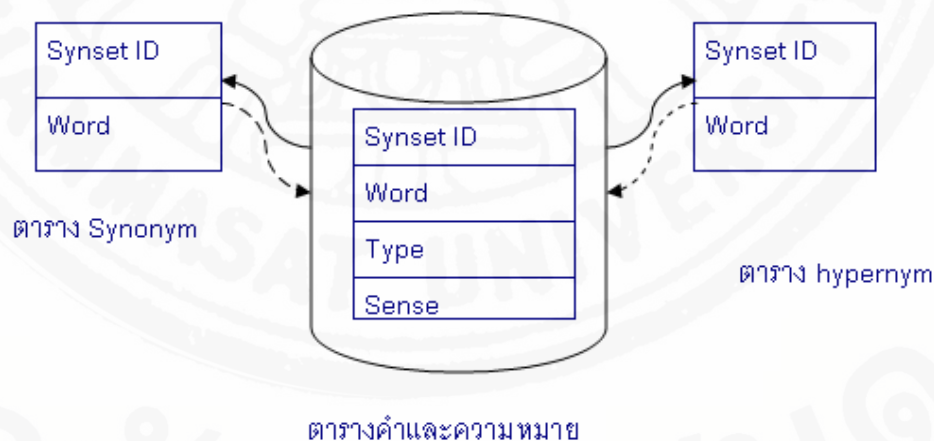
- หมายเลขของ Synset
- ชนิดของ Synset ว่าเป็น Noun Verb Adjective หรือ Adverb
- ลำดับที่ของคำใน Synset นั้น ๆ
- คำที่อยู่ใน Synset
- ตัวชี้ (Pointer) และสัญลักษณ์ของตัวชี้ ซึ่งจะเป็นบอกถึงความสัมพันธ์ Synset หมายเลขใด และมีความสัมพันธ์แบบไหน
- ความหมายของ Synset และตัวอย่างประโยคหรือวลี

ภาพที่ 3.2
แสดงข้อมูลในเวิร์ดเน็ต

00004492 29 v 01 hold 4 001 @ 02399099 v 0000 01 + 08 00 | keep from exhaling or expelling; "hold your breath"
00006735 29 v 01 wheeze 0 002 @ 00001740 v 0000 + 00824340 n 0101 01 + 02 00 | breathe with difficulty
00008233 29 v 02 bat 0 flutter 0 001 @ 00007777 v 0000 01 + 08 00 | wink briefly; "bat one's eyelids"
00008337 29 v 01 wink 2 003 @ 00982279 v 0000 + 06786526 n 0101 + 10622400 n 0101 01 + 22 00 | signal by winking;

ในการทดลองของงานวิจัยนี้ได้จัดเก็บข้อมูลเวิร์ดเน็ตในระบบฐานข้อมูลโดยจัดเก็บเป็นตาราง ซึ่งประกอบด้วย ตารางหลักสำหรับเก็บหมายเลข Synset คำใน Synset และความหมาย ส่วนตารางที่เหลือจะเป็นตารางคำที่มีความสัมพันธ์กันในแบบต่าง ๆ ได้แก่ ได้แก่ Synonym, Hypernym, Hyponym, Antonym, Holonym และ Meronym ซึ่งตารางเหล่านี้จะเก็บหมายเลข Synset ในการที่จะเชื่อมโยง และคำที่มีความสัมพันธ์กับตารางหลัก

ภาพที่ 3.3
แสดงโครงสร้างและความสัมพันธ์ตารางข้อมูลเวิร์ดเน็ต



3.3.2 เครื่องมือในการสร้างกราฟเชิงความคิด

- 1) Machine Syntax จากเว็บไซต์ www.connexor.com (Connexor Oy Inc) ซึ่งเป็นเครื่องมือในการแจกแจงไวยากรณ์ของประโยค โดย Machine Syntax พัฒนาตามการแจก

แจกแจงไวยากรณ์แบบ Functional Dependency Grammar โดยจะบันทึกผลที่ได้จากการแจกแจงไวยากรณ์ของ Machine Syntax ให้อยู่ในรูปของ text file

เช่นประโยค “He evaluates the animals based on official standards for each type of rat or mouse.” จะได้ผลจากการแจกแจงไวยากรณ์ ดังนี้

ภาพที่ 3.4

แสดงผลที่ได้จากการแจกแจงไวยากรณ์ของ Machine Syntax

Analysis of Machine Syntax for English:

#	Text	Baseform	Syntactic relation	Syntax and morphology
1	He	he	subj:>2	@SUBJ %NH PRON PERS NOM SG3
2	evaluates	evaluate	main:>0	@+FMAINV %VA V PRES SG3
3	the	the	det:>4	@DN> %>N DET
4	animals	animal	obj:>2	@OBJ %NH N NOM PL
5	based	base	mod:>4	@-FMAINV %VP EN
6	on	on	phr:>5	@ADVL %EH PREP
7	official	official	attr:>8	@A> %>N A ABS
8	standards	standard	pcomp:>6	@<P %NH N NOM PL
9	for	for	mod:>8	@<NOM %N< PREP
10	each	each	det:>11	@DN> %>N DET SG
11	type	type	pcomp:>9	@<P %NH N NOM
12	of	of	mod:>11	@<NOM-OF %N< PREP
13	rat	rat	pcomp:>12	@<P %NH N NOM SG
14	or	or	cc:>13	@CC %CC CC
15	mouse	mouse	cc:>13	@<P %NH N NOM SG
16	.	.		
17	<s>	<s>		

2) โปรแกรมสร้างกราฟเชิงความคิด

โปรแกรมในการสร้างกราฟเชิงความคิดจะใช้ผลการแจกแจงไวยากรณ์ที่ได้จาก Machine Syntax เป็นข้อมูลนำเข้าในการสร้างเป็นกราฟเชิงความคิด โดยมีหลักในการทำงาน ดังนี้

ข้อมูลเข้า:

1. ประโยคภาษาอังกฤษ
2. พื้นข้อมูลเกี่ยวกับคำ
3. โปรแกรมแจกแจงไวยากรณ์

ผลลัพธ์ที่ได้:

กราฟแทนความหมาย

ขั้นตอนการทำงาน:

1. แจกแจงไวยากรณ์ของประโยคหนึ่งๆ เพื่อจะได้ข้อมูลเกี่ยวกับ syntax, morphology และ syntactical relations. ซึ่งในวิทยานิพนธ์ฉบับนี้ใช้ซอฟต์แวร์ชื่อ "Machine Syntax" พัฒนาโดย (Voutilainen and Heikkila, 1993) and (Connexor Oy Inc., 2005) ซึ่งตัวอย่างของผลลัพธ์การแจกแจงไวยากรณ์แสดงในรูปที่ 3.4.
2. เลือกคำที่มี syntactic relation เท่ากับ "main" เพื่อเป็นจุดเริ่มต้น
3. แทนคำนั้นๆ ด้วยกรอบสี่เหลี่ยม เรียกว่า concept node
4. ค้นหาสายเชื่อมโยงระหว่างคำนั้นๆ กับคำอื่น
5. ติดตามสายเชื่อมโยงไปยังคำที่เชื่อมกันที่มี Morphological tag เท่ากับ "Noun", "Verb" และ "Adjective".
6. แทนความสัมพันธ์ระหว่าง concept ด้วย semantic roles (ดังแสดงในภาคผนวก ข).
7. ทำซ้ำข้อ 3 จนกระทั่งไม่มีสายเชื่อมโยงเหลือ

ภาพที่ 3.5
แสดงผลที่ได้จากการโปรแกรมการสร้างกราฟ

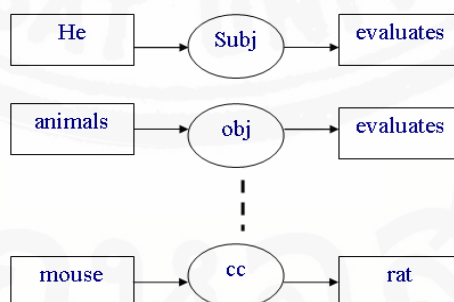
```

He,subj,evaluates
animals,obj,evaluates
based,mod,animals
on,phr,based
official,attr,standards
standards,pcomp,on
for,mod,standards
type,pcomp,for
of,mod,type
rat,pcomp,of
or,cc,rat
mouse,cc,rat
*****

```

จากผล (Output File) ที่ได้สามารถเขียนเป็นกราฟเชิงความความคิดได้ดังนี้

ภาพที่ 3.6
แสดงผลกราฟเชิงความคิด



3.4 ขั้นตอนการแทนความหมายของประโยค

ในขั้นตอนนี้เป็นการแทนความหมายของประโยคคำแปลที่รวบรวมมาโดยใช้ Machine Syntax จากนั้นใช้โปรแกรมในการสร้างกราฟเชิงความคิดโดยมีผลที่ได้จากแจกแจงไวยากรณ์เป็นข้อมูลนำเข้า แล้วนำผลที่ได้จากโปรแกรมสร้างกราฟจัดเก็บในระบบฐานข้อมูล ซึ่งการจัดเก็บข้อมูลกราฟในระบบฐานข้อมูลนั้นมีตารางจัดเก็บ 2 ส่วนด้วยกันคือ

1. ตารางคำ เก็บข้อมูล

- เลขที่ของคำ
- คำ

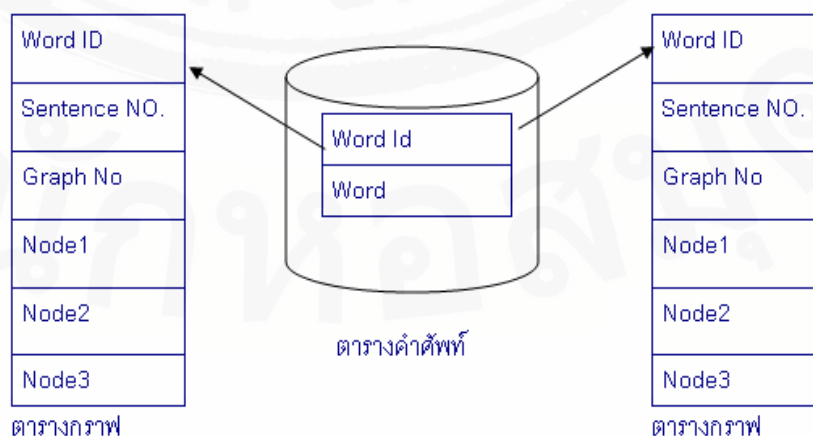
2. ตารางกราฟ เก็บข้อมูลกราฟเชิงความคิดซึ่งสร้างจากประโยคคำแปลของคำประกอบด้วย

- เลขที่คำ
- เลขที่ประโยค
- เลขที่กราฟ
- ข้อมูลในโหนดที่ 1
- ข้อมูลในโหนดที่ 2
- ข้อมูลในโหนดที่ 3

โดยมีโครงสร้างตารางดังรูป

ภาพที่ 3.7

แสดงรูปแบบการจัดเก็บข้อมูลกราฟในระบบฐานข้อมูล



3.5 ขั้นตอนการเปรียบเทียบหาค่าความสัมพันธ์

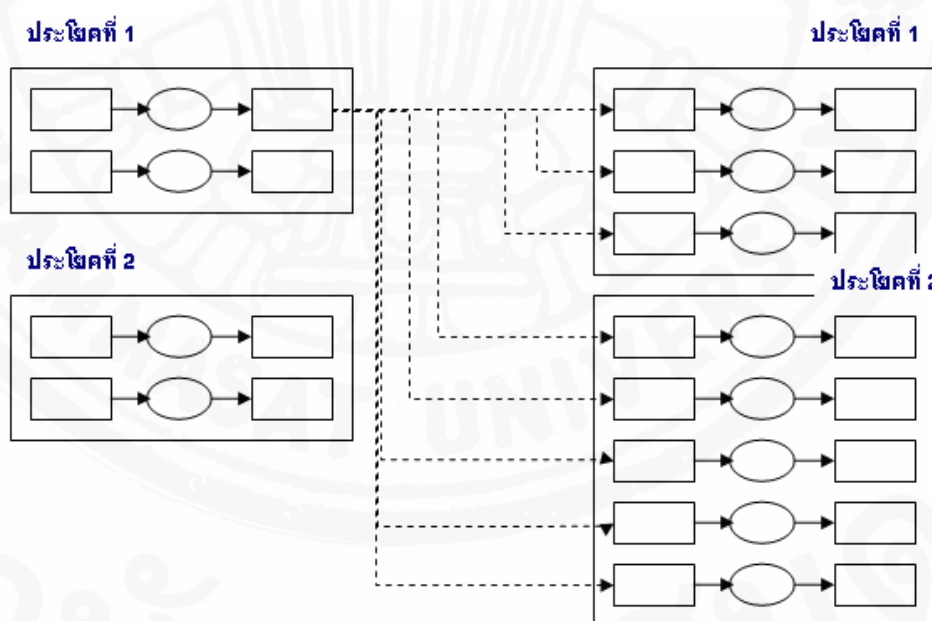
3.5.1 การเปรียบเทียบกราฟเชิงความคิด

การเปรียบเทียบกราฟเชิงความคิด คือ เป็นการแทนความหมายของประโยคคำแปลเพื่อการวัดค่าความสัมพันธ์ในเชิงความหมายของคำนั้น กราฟที่ได้จากการแทนความหมายของประโยคในขั้นตอนที่ 3.4 นั้นมักจะมีมากกว่า 1 กราฟ ดังนั้นในการเปรียบเทียบกราฟของประโยคมีหลักในการเปรียบเทียบดังนี้

1. กราฟทุกกราฟของประโยคซึ่งแทนความหมายของคำแรกจะต้องเปรียบเทียบกับกราฟทุกกราฟของประโยคซึ่งแทนความหมายของคำที่สองที่จะเปรียบเทียบด้วย

ภาพที่ 3.8

แสดงการเปรียบเทียบกราฟซึ่งแทนความหมายของประโยค

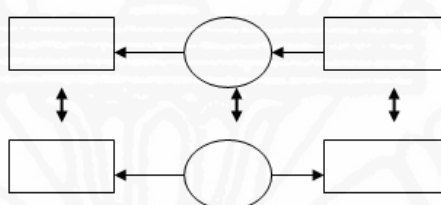


การเปรียบเทียบกราฟในแต่ละประโยคตามภาพที่ 3.8 คือ ประโยคในคำนิยามของคำสำคัญมี 2 ประโยค และแต่ละประโยคประกอบด้วย 2 กราฟเชิงความคิด ส่วนประโยคในคำนิยามของคำที่กำหนดให้ประโยคแรกประกอบด้วยกราฟเชิงความคิด 3 กราฟ และประโยคที่สองมี 5 กราฟ โดยกราฟทุกกราฟในฝั่งซ้ายจะต้องเปรียบกับทุกกราฟในฝั่งขวา

2. แต่ละ node ภายในกราฟจะถูกเปรียบเทียบกับ concept node จะถูกเปรียบกับ concept node และ relation node จะถูกเปรียบเทียบกับ relation node เช่นกัน ดังรูป

ภาพที่ 3.9

แสดงการเปรียบเทียบระหว่างโหนด



3.5.2 การให้น้ำหนักความสัมพันธ์

ในการพิจารณาหาความสัมพันธ์กันของโหนดซึ่งอยู่ในกราฟที่ทำการเปรียบเทียบกับนั้น ในงานวิจัยนี้พิจารณาความสัมพันธ์ที่เป็น

1. Synonym
2. Hypernym
3. Hyponym
4. Antonym
5. Meronym
6. Holonym

วิทยานิพนธ์ฉบับนี้ได้ให้น้ำหนักความสำคัญของความสัมพันธ์ในแต่ละประเภทที่แตกต่างกัน โดยในการเปรียบเทียบระหว่างโหนดในขั้นแรกจะพิจารณาความเหมือนกันก่อน หากเป็นคำคำเดียวกันก็จะให้น้ำหนักเท่ากับ 1 หากไม่เหมือนกันจึงค่อยมาพิจารณาว่ามีความสัมพันธ์แบบใดโดยอาศัยฐานความรู้จากฐานข้อมูลเวิร์ดเน็ตที่ได้สร้างไว้ ซึ่งในงานวิจัยนี้ได้ให้น้ำหนักความสัมพันธ์กันในแต่ละประเภท ดังนี้

ตารางที่ 3.3
แสดงการให้น้ำหนักความสัมพันธ์แต่ละประเภท

Synonym	score = 1
Hyponym	score = 0.8
Hypernym	score = 0.7
Antonym	score = 0.5
Meronym	score = 0.5
Holonym	score = 0.5

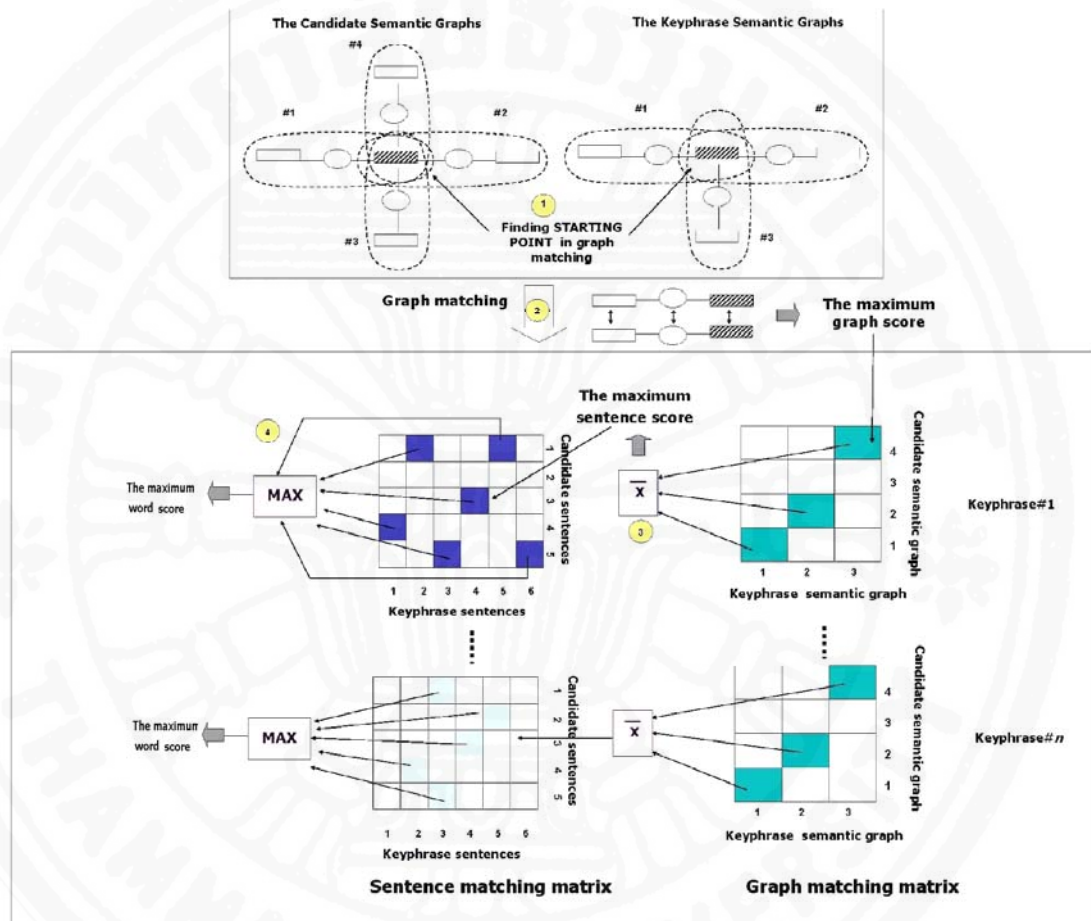
จากตารางที่ 3.3 จะเห็นได้ว่าเราจะให้น้ำหนักความสัมพันธ์ Synonym ซึ่งจะมี
ความหมายเหมือนกันมากที่สุด ส่วนความสัมพันธ์แบบ Hypernym นั้นจะให้น้ำหนักน้อยกว่า
ความสัมพันธ์แบบ Hyponym

3.5.3 การคำนวณหาค่าความสัมพันธ์

ในการวัดความสัมพันธ์เชิงความหมายระหว่างคำสองคำ สมมติว่าเป็นคำว่า “car”
และ “vehicle” เราจะนำกราฟเชิงความคิดจากขั้นตอนที่ 2 มาคำนวณหาคะแนนความสัมพันธ์
ของกราฟ, ประโยค และ คำ ตามลำดับ จากรูปที่ 3.10 ทางด้านซ้ายมือเป็นกราฟเชิงความคิดของ
ประโยคที่อธิบายความหมายของคำว่า “car” ซึ่งประกอบไปด้วย กราฟย่อยๆ จำนวน 4 กราฟ
ส่วนทางด้านขวามือเป็น กราฟเชิงความคิดของประโยคที่อธิบายความหมายของคำว่า “vehicle”
ซึ่งประกอบไปด้วยกราฟย่อยๆ จำนวน 3 กราฟ ขั้นตอนในการคำนวณหาค่าความสัมพันธ์ระหว่าง
คำว่า “car” กับ “vehicle” มีดังนี้

1. กำหนดจุดเริ่มต้นในการเปรียบเทียบ เนื่องจากในหนึ่งประโยคประกอบด้วยกราฟ
ย่อยๆ จำนวนหลายกราฟ เราจะเริ่มต้นการเปรียบเทียบด้วยการเลือกโหนดที่มี
จำนวนเส้นเชื่อมต่อมากที่สุด ดังแสดงด้วยกรอบแรเงาในรูปที่ 3.10

ภาพที่ 3.10
แสดงการคำนวณค่าสัมพันธ



2. เปรียบเทียบกราฟย่อยแต่ละกราฟ กราฟที่เชื่อมต่อกับโหนดเริ่มต้นจากข้อ 1 จะถูกเปรียบเทียบด้วยสมการ

$$graph_score(i, j) = \frac{\sum_{n=1}^3 hit(n)}{3} \quad (3.1)$$

กำหนดให้ $graph_score(i,j)$ คือ คะแนนที่เปรียบเทียบกราฟที่ i ของประโยคที่อธิบายคำว่า “car” กับ คะแนนที่เปรียบเทียบกราฟที่ j ของประโยคที่อธิบายคำว่า “vehicle” และ $hit(n)$ คือ สถานะที่บอกว่าข้อมูลในโหนดที่เปรียบเทียบกันเท่ากันหรือไม่ ในการพิจารณาว่ามีความสัมพันธ์แบบใดโดยอาศัยฐานความรู้จากฐานข้อมูลเว็บริดเน็ตที่ได้สร้างไว้และให้น้ำหนักความสัมพันธ์ในแต่ละแบบตามที่กำหนดในข้อ 3.5.2 โดยน้ำหนักเหล่านี้จะนำมาคูณกับ $hit(n)$ เช่น ถ้าค่าของข้อมูลที่อยู่ในแต่ละโหนดเป็น Synonym กัน ค่าของ $hit(n)$ เท่ากับ $hit(n)$ แต่ถ้าค่าที่อยู่ในโหนดที่นำมาเปรียบเทียบเป็น Hyponym ของค่าที่อยู่ในโหนดแรก ค่าของ $hit(n)$ จะเท่ากับ $hit(n) \times 0.8$ คะแนนที่ได้จากการเปรียบเทียบแต่ละครั้งเป็นของกราฟย่อยที่ประกอบไปด้วยโหนดจำนวน 3 โหนด ซึ่งเราจะนำมาคำนวณหาค่าเฉลี่ย จากนั้นจะนำไปเก็บไว้ใน graph matching matrix ดังแสดงในภาพที่ 3.10

3. หาคะแนนของการเปรียบเทียบประโยค เนื่องจากในการแทนความหมายของแต่ละประโยค เราสามารถแทนได้ด้วยกราฟย่อยหลายๆ กราฟ ดังนั้น คะแนนของการเปรียบเทียบของประโยคจะมาจากการรวมค่าที่ได้จากการเปรียบเทียบกราฟย่อย ดังในสมการที่ 2 เนื่องจากในการเปรียบเทียบกราฟย่อยแต่ละรอบเราได้ค่าหลายค่า จากบทความนี้เราจะเลือกคะแนนกราฟย่อยที่มีค่าเฉลี่ยสูงสุด มารวมกันเพื่อเก็บเป็นคะแนนของประโยคโดยเก็บไว้ใน Sentence matching matrix ดังแสดงในภาพที่ 3.10

$$sentence_score(i, j) = \frac{\sum_{i=1}^k (\max_{j=1}^k (graph_score(i, j)))}{k} \quad (3.2)$$

4. หาคะแนนของการเปรียบเทียบคำ ในทำนองเดียวกัน คำหนึ่งคำ อาจมีประโยคที่อธิบายความหมายได้มากกว่า หนึ่งประโยค เมื่อนำเอาแต่ละประโยคมารวมกันแล้ว จะได้คะแนนของประโยคที่อยู่ใน Sentence matching matrix ซึ่งในการพิจารณาความเหมือนของคำในงานวิจัยนี้กำหนดว่า คำสองคำ

ถ้ามีประโยคที่อธิบายความหมายที่เหมือนกันอย่างน้อยหนึ่ง ความสัมพันธ์
ระหว่างคำ โดยใช้ความหมายของประโยคเป็นตัวกำหนดของเซตของความหมาย

$$\text{word_score}(i, j) = \max_{i=1}^k \max_{j=1}^c (\text{sentence_score}(i, j)) \quad (3.3)$$

