บทที่ 4

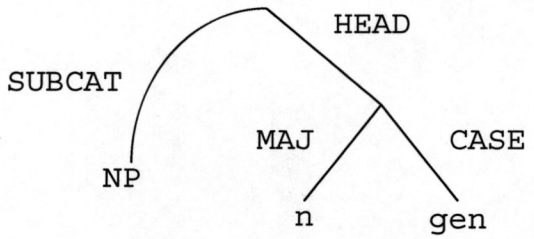
การวิเคราะห์กระจายประโยค

ในงานวิจัยนี้ได้นำทฤษฎีเอชพีเอสจีมาประยุกต์ใช้ในส่วนวิเคราะห์กระจายประโยค โดยทำให้เกิดผลโครงสร้างคุณสมบัติ (Feature Structure) ในรูปของกราฟไร้วงที่มีทิศทาง (Directed Acyclic Graph) มาใช้ในการเก็บข้อมูลของคลังคำ ทั้งนี้กลไกหลักในการกระจายประโยคคือการวิเคราะห์กระจายแบบแอลอาร์ (LR Parser) ซึ่งใช้ในการวิเคราะห์กระจายไวยากรณ์แบบไม่พึ่งบริบท (Context-Free Grammar) โดยมีตัวดำเนินการยูนิฟาย (Unify Operator) มาใช้ในการประกอบคลังคำที่อยู่ในรูปของโครงสร้างคุณสมบัติมาประกอบเป็นโครงสร้างองค์ประกอบตามกฎในแต่ละสถานะของการเกิดวลีหรือประโยค ซึ่งผลสุดท้ายที่ได้จากการวิเคราะห์กระจายประโยคคือโครงสร้างองค์ประกอบของประโยค

การทำให้เกิดผลของโครงสร้างคุณสมบัติ

เอชพีเอสจีเป็นไวยากรณ์ที่ใช้พื้นฐานของการยูนิฟิเคชัน (Unification-based Grammar) ชนิดหนึ่งที่เก็บสารสนเทศคลังคำในโครงสร้างคุณสมบัติ โดยที่โครงสร้างมีลักษณะคล้ายแมททริกซ์ที่จัดเก็บสารสนเทศของคำอันประกอบด้วยลักษณะเฉพาะ (Attribute) และค่าของลักษณะเฉพาะ (Value) ซึ่งค่าของลักษณะเฉพาะอาจเป็นได้ทั้งค่าอะตอมมิก (Atomic Value) หรือโครงสร้างคุณสมบัติเฉพาะ ทำให้เรียกโครงสร้างคุณสมบัติได้อีกอย่างว่าแมททริกซ์ค่าลักษณะเฉพาะหรือเอวีเอ็ม (AVM - Attribute Value Metrix) การนำเสนอโครงสร้างคุณสมบัติอาจจัดทำให้อยู่ในรูปของโครงสร้างต้นไม้ หรือกราฟไร้วงที่มีทิศทางเป็นต้น ซึ่งขึ้นอยู่กับการทำให้เกิดผลของแต่ละระบบที่จัดทำขึ้น

เนื่องจากคำแต่ละคำมีจำนวนลักษณะเฉพาะในโครงสร้างคุณสมบัติไม่เท่ากันอันเป็นคุณลักษณะแบบ n-ary ทำให้การนำเสนอด้วยโครงสร้างต้นไม้เพียงอย่างเดียวนั้นไม่ง่ายต่อการทำให้เกิดผล ในทางตรงข้ามการนำเสนอด้วยทฤษฎีกราฟสามารถทำได้ง่ายกว่ากล่าวคือด้วยทฤษฎีกราฟจะสามารถนำเสนอรูปแบบจำลองของทฤษฎีทางภาษาศาสตร์ที่ใช้ระบบโครงสร้างคุณสมบัติได้ง่ายกว่าและสร้างตัวดำเนินการรวมโครงสร้างที่มีศักยภาพได้ การทำให้เกิดผลของโครงสร้างคุณสมบัติด้วยกราฟไร้วงที่มีทิศทางทำได้โดยจัดให้เส้นของกราฟเป็นชื่อของลักษณะเฉพาะ และแต่ละเส้นจะชี้ไปยังโครงสร้างคุณสมบัติย่อยหรือค่าอะตอมมิก ซึ่งโครงสร้างคุณสมบัติที่ถูกแสดงแทนด้วยโครงสร้างกราฟสามารถแสดงได้ดังรูปที่ 4.1

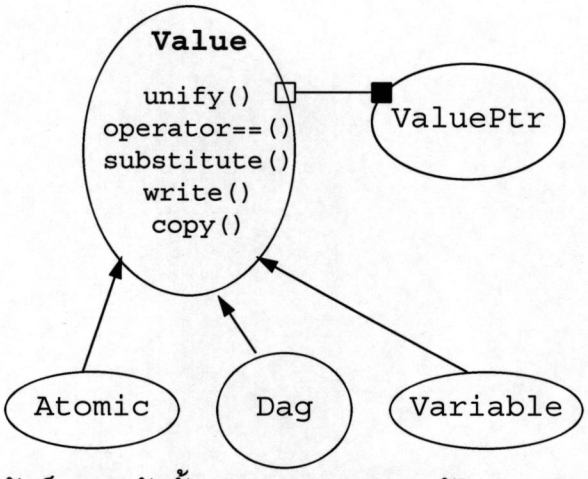


รูปที่ 4.1 โครงสร้างคุณสมบัติที่ถูกแสดงแทนด้วยโครงสร้างกราฟ

การแทนด้วยกราฟประเภทนี้เมื่อนำมาให้เกิดผลด้วยทฤษฎีเชิงวัตถุ (Object Oriented Theory) ก็จะสามารถทำให้ง่ายต่อการทำให้เกิดผลทั้งในแง่ของการสร้างรูปแบบจำลองและการทำการรวมโครงสร้างคุณสมบัติ เนื่องจากทฤษฎีเชิงวัตถุมีคุณสมบัติในการถ่ายทอดคุณสมบัติ (Inheritance Object) และคุณสมบัติแบบโพลิมอร์ฟิซึม (polymorphism) ซึ่งเพียงพอต่อการนำมาทำให้เกิดผลกับโครงสร้างคุณสมบัติที่แทนด้วยทฤษฎีกราฟนี้ และยังง่ายต่อการทำให้เกิดผลกับการดำเนินการยูนิฟิเคชันในไวยากรณ์ได้ด้วย นอกจากนี้ประโยชน์ที่นำทฤษฎีเชิงวัตถุมาใช้ในการทำให้เกิดผลนี้คือความง่ายในการพัฒนาโครงสร้างคุณสมบัติเพื่อใช้กับทฤษฎีไวยากรณ์อื่นๆที่ใช้โครงสร้างคุณสมบัติเป็นที่เก็บสารสนเทศของคลังคำ และแนวโน้มของการพัฒนาระบบในปัจจุบันส่วนใหญ่ได้หันมาใช้ทฤษฎีเชิงวัตถุมากขึ้นมีผลทำให้เกิดเครื่องมือในการพัฒนาให้สามารถเลือกใช้ได้มากมายในท้องตลาดมากขึ้น

ในงานวิจัยนี้ได้ทำให้เกิดผลกราฟไร้วงที่มีทิศทางด้วย C++ โดยได้นำโปรแกรมของ Perelman-Hall (1995) ที่ใช้ทำให้เกิดผลกราฟไร้วงที่มีทิศทางเพื่อใช้กับตัวดำเนินการยูนิฟิเคชันมาทำการปรับปรุงให้สามารถใช้ได้กับทฤษฎีเอชพีเอสจีได้มากยิ่งขึ้น โดยในตัวโปรแกรมได้นำเสนอตัวอะตอมมิก ตัวแปร และกราฟไร้วงที่มีทิศทาง ในรูปของ class ด้วย

การถ่ายทอดคุณสมบัติจาก virtual class ที่ชื่อว่า Value ซึ่งมีการกำหนดเป็น virtual function ไว้และส่วนใหญ่จะถูกกำหนดให้เป็น pure virtual function เพื่อให้ class ทั้งสามประเภทสามารถนำไปกำหนดใหม่ให้เหมาะสม ซึ่งเป็นคุณสมบัติการดำเนินการหลายประเภท นอกจากนี้ยังมีการกำหนดให้สามารถอ้างอิงได้ด้วย class ที่ชื่อว่า ValuePtr ดังแสดงไว้ในรูปที่ 4.2



รูปที่ 4.2 ความสัมพันธ์ของลำดับขั้นการถ่ายทอดคุณสมบัติของ Value และ ValuePtr

ด้วยการออกแบบในลักษณะเช่นนี้ เมื่อมีความต้องการแก้ไขหรือเพิ่มเติมอะไรก็ตามให้กับกราฟไร้วงที่มีทิศทางนี้มีคุณสมบัติในการใช้งานที่เหมาะสมกับทฤษฎีเอชพีเอสจีหรือกลุ่มไวยากรณ์ที่มีการใช้โครงสร้างคุณสมบัติแบบเดียวกันก็สามารถทำได้โดยไม่ยากนัก

การนำเสนอคลังคำเอชพีเอสจี

ในงานวิจัยนี้ได้นำเสนอโครงสร้างคลังคำตามแบบทฤษฎีของเอชพีเอสจีบางส่วนเท่านั้น ซึ่งได้ตัดลักษณะเฉพาะบางส่วนออกไปเพื่อให้ง่ายต่อการทำวิจัย โดยได้นำเสนอรูปแบบของโครงสร้างคลังคำ ดังนี้

$$\begin{bmatrix} \text{PHON} : & \\ \text{SYN} : & \begin{bmatrix} \text{HEAD} : & \\ \text{SUBCAT} : & \end{bmatrix} \\ \text{SEM} : & \end{bmatrix}$$

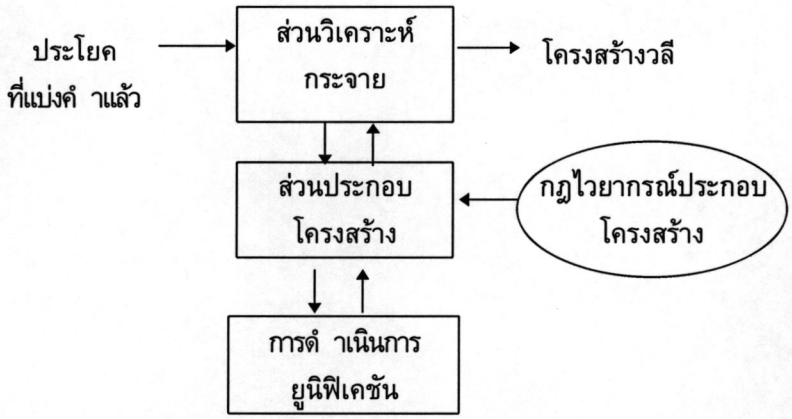
โดยที่

- PHON เป็นส่วนที่ใช้ในการระบุถึงคำหรือการออกเสียงของคำนั้น
- SYN เป็นส่วนที่เก็บข้อมูลของส่วนวากยสัมพันธ์
- HEAD เป็นส่วนที่ระบุสารสนเทศในส่วนวากยสัมพันธ์ของหน่วยหลัก
- SUBCAT เป็นส่วนที่หน่วยภาษาที่ต้องเกิดร่วมด้วย
- SEM เป็นข้อมูลของคลังคำในส่วนอรรถศาสตร์

การกำหนดประเภทของคำว่าเป็นคำนาม คำกริยา คำบุพบท หรือคำวิเศษณ์ ถูกระบุไว้ในส่วนหนึ่งของลักษณะเฉพาะที่มีชื่อว่า HEAD และสารสนเทศที่เป็นส่วนประกอบของคำที่จะทำให้การรวมโครงสร้างเป็นโครงสร้างองค์ประกอบจะขึ้นอยู่กับลักษณะเฉพาะของแต่ละคำ ซึ่งคำแต่ละคำจะมีลักษณะเฉพาะย่อยเหล่านี้ไม่เหมือนกัน

อัลกอริทึมวิเคราะห์กระจาย

ในการวิเคราะห์กระจายในรูปแบบของเอชพีเอสจีเป็นการพิจารณาคำทีละคำโดยผ่านขั้นตอนการดำเนินการ 3 ส่วนด้วยกันคือ ส่วนวิเคราะห์กระจายไวยากรณ์ ส่วนประกอบโครงสร้างองค์ประกอบ (Constituent Structure) และส่วนดำเนินการยูนิฟิเคชัน (Unification) โดยมีส่วนวิเคราะห์กระจายไวยากรณ์เป็นโครงสร้างหลักทำหน้าที่พิจารณาคำที่เข้ามาว่าถูกต้องตามกฎไวยากรณ์หรือไม่ โดยดูจากการเรียงลำดับของคำในภาษาว่าถูกต้องหรือไม่ ส่วนประกอบโครงสร้างทำหน้าที่ประกอบโครงสร้างของคลังคำหรือวลีเข้าไปในโครงสร้างวลีหรือประโยคตามกฎไวยากรณ์ที่เกิดจากแต่ละสถานะในส่วนวิเคราะห์ประโยค ซึ่งเมื่อพิจารณาให้ดีแล้วส่วนประกอบโครงสร้างก็คือส่วนที่กำหนดว่าโครงสร้างคุณสมบัติของคำกับคำ หรือคำกับวลี หรือวลีกับวลี จะประกอบกันเป็นโครงสร้างองค์ประกอบที่อยู่ในรูปของโครงสร้างคุณสมบัติอย่างไรในแต่ละกฎ ทั้งนี้มีส่วนการดำเนินการยูนิฟิเคชันไว้ใช้เป็นตัวดำเนินการประกอบโครงสร้างคุณสมบัติที่เข้ามาลงในโครงสร้างวลีหรือประโยคตามที่ส่วนประกอบโครงสร้างกำหนด



รูปที่ 4.4 แสดงส่วนประกอบในการวิเคราะห์ประโยค

กลไกหลักในอัลกอริทึมวิเคราะห์กระจายคือส่วนวิเคราะห์กระจายไวยากรณ์ที่ใช้กลไกของตัววิเคราะห์กระจายแบบแอลอาร์ (LR Parsing) ในการกระจายไวยากรณ์แบบไม่พึ่งบริบท (CFG - Context-Free Grammar) ซึ่งเป็นการกระจายคำเชิงพื้นผิว (Surface Parsing) เท่านั้น โดยในแต่ละสถานะของการเกิดการกระทำ reduce เป็นวลีหรือประโยค จะมีการดำเนินการประกอบโครงสร้างของคลังคำให้เป็นโครงสร้างองค์ประกอบตามแต่ละฟังก์ชันที่กำหนดไว้ในกฎของการเกิดวลีหรือประโยคนั้นๆ ซึ่งจะมีการใช้ตัวดำเนินการยูฟิเคชันมาทำการผูกโครงสร้างคุณสมบัติ (Feature Binding) และเป็นการตรวจสอบความเข้ากันได้ของโครงสร้างด้วย ผลสุดท้ายที่ได้จากอัลกอริทึมนี้คือโครงสร้างองค์ประกอบขนาดใหญ่ของประโยค หลักไวยากรณ์ที่ใช้ในงานวิจัยมีดังนี้

```
S    →   NP  VP
S    →   VP
S    →   S  PP
NP   →   n
NP   →   n  S
NP   →   n  AP
NP   →   n  PP
PP   →   p  NP
AP   →   adj
AP   →   adj  n
VP   →   v  NP
VP   →   v  PP
```

โดยที่

n	หมายถึง	คำนาม
v	หมายถึง	กริยา
p	หมายถึง	บุพบท

adj หมายถึง นามวิเศษณ์
S หมายถึง ประโยค
NP หมายถึง นามวลี
VP หมายถึง กริยาวลี
PP หมายถึง บุพบทวลี
AP หมายถึง วิเศษณ์วลี

State	Action Table					Goto Table				
	n	v	p	adj	$	S	NP	VP	PP	AP
0	sh4	sh5				1	2	3		
1			sh7		acc				6	
2		sh5						8		
3			re1		re1					
4	sh4/re3	sh5/re3	sh7	sh12	re3	9	2	3	11	10
5	sh4		sh7				13		14	
6			re2		re2					
7	sh4						15			
8			re0		re0					
9		re4	sh7		re4				16	
10		re5			re5					
11		re6			re6					
12	sh17	re8			re8					
13			re10		re10					
14			re11		re11					
15			re7		re7					
16			re12		re12					
17		re9			re9					

ตาราง 4.1 ตารางไวยากรณ์ไม่พึ่งบริบทสำหรับตัวกระจายไวยากรณ์แบบแอลอาร์

จากไวยากรณ์ที่ได้สามารถเขียนเป็นตารางไวยากรณ์สำหรับกลไกของตัววิเคราะห์กระจายแบบแอลอาร์ได้ดังในตาราง 4.1 ซึ่งไวยากรณ์ดังกล่าวใช้ในการวิเคราะห์ประโยคที่มีไวยากรณ์ไม่ซับซ้อนมากนัก ซึ่งถ้าหากในอนาคตมีการศึกษาเพิ่มเติมในภายหลัง ไวยากรณ์ที่ได้จะมีลักษณะที่แตกต่างจากที่กำหนดไว้ข้างต้น และสามารถทำการแก้ไขตารางไวยากรณ์ที่ใช้ในตัววิเคราะห์กระจายแบบแอลอาร์ได้

การประกอบเป็นโครงสร้างองค์ประกอบ

ส่วนการประกอบโครงสร้างนี้เป็นส่วนที่เกิดขึ้นเมื่อพบว่ามีการนำคำมาประกอบเป็นวลีหรือประโยค ซึ่งสิ่งที่ได้คือโครงสร้างองค์ประกอบของวลีหรือประโยคดังที่กล่าวไว้ข้างต้น วิธีที่จะทำการประกอบโครงสร้างเหล่านี้สามารถทำได้โดยสร้างโครงสร้างองค์ประกอบขึ้นมา ซึ่งจะทำการเพิ่มลักษณะเฉพาะอีกตัวหนึ่งที่ชื่อ DTRS (Daughters) เพื่อใช้เป็นส่วนที่จะนำโครงสร้างของคำหรือโครงสร้างวลีมาประกอบกัน ซึ่งในค่าลักษณะเฉพาะของ DTRS จะมีลักษณะเฉพาะอีกสองตัวคือ HEAD-DTR (Head Daughter) และ COMP-DTRS (Complement Daughters) โดยให้ HEAD-DTR เป็นตัวนำของวลีหรือประโยค และให้ COMP-DTRS เป็นส่วนประกอบของวลีหรือประโยคนั้น

$$\begin{bmatrix} \text{PHON} : \\ \text{SYN} : \\ \text{DTRS} : \begin{bmatrix} \text{PHON} : \\ \text{HEAD - DTR} : \\ \text{COMP - DTRS} : \end{bmatrix} \\ \text{SEM} : \end{bmatrix}$$

ซึ่งขณะที่มีการประกอบเป็นโครงสร้างองค์ประกอบ อัลกอริทึมจะทำการตรวจสอบคำที่เข้าประกอบว่าถูกต้องตามวากยสัมพันธ์หรือไม่ โดยนำข้อมูลของ SUBCAT จากตัวที่จะนำมาประกอบในส่วนของ HEAD-DTR มาดำเนินการยูนิฟาย ถ้าหากไม่ผ่านขั้นตอนการยูนิฟายนั่นคือคำที่นำมาประกอบนั้นไม่เหมาะสมหรือผิดหลักวากยสัมพันธ์ตามกฎในไวยากรณ์ที่กำหนดไว้

ข้อผิดพลาดในการวิเคราะห์กระจาย

ในการดำเนินการทั้งสามส่วนของส่วนวิเคราะห์กระจายดังที่กล่าวไว้ข้างต้นอาจเกิดข้อผิดพลาดขึ้นได้ในแต่ละส่วน เช่น ไม่สามารถทำการกระจายประโยคได้อันเนื่องมาจากการข้อผิดพลาดในส่วนของการทำยูนิฟิเคชัน เป็นต้น ข้อผิดพลาดเหล่านี้มีสาเหตุหลักดังต่อไปนี้

1. การเลือกชนิดของคำจากคลังคำไม่เหมาะสม ข้อผิดพลาดประเภทนี้เกิดจากการที่คำหนึ่งมีชนิดของคำได้หลายประเภท เช่น "ที่" มีชนิดของคำเป็นนาม ลักษณะนาม สรรพนาม วิเศษณ์ และสันธาน เป็นต้น วิธีแก้ข้อผิดพลาดประเภทนี้คือ การให้น้ำหนักในการเลือกคำในทางสถิติ หรือการทดลองเลือก (Trial and Error) เป็นต้น

2. การแบ่งคำที่ไม่ถูกต้อง ข้อผิดพลาดประเภทนี้เกิดจากส่วนแบ่งคำเนื่องจากการแบ่งคำด้วยพจนานุกรมเพียงอย่างเดียวไม่ช่วยในการกำหนดได้ว่าการแบ่งคำนั้นถูกต้องตามกฎไวยากรณ์และความหมายหรือไม่ ส่วนแบ่งคำเพียงแต่หาจุดแบ่งคำที่คิดว่าเหมาะสมออกมาเท่านั้น และการแบ่งคำที่ได้มักเป็นคำที่ยาวที่สุดเท่าที่อัลกอริทึมแบ่งคำจะสามารถครอบคลุมได้ ดังนั้นส่วนแบ่งคำจึงไม่สามารถแบ่งได้ในทุกกรณี เช่น คำกำกวม เป็นต้น ปัญหาเหล่านี้ต้องทำการแก้ไขต่อไป

3. ไวยากรณ์ที่มีไม่เพียงพอต่อการวิเคราะห์กระจาย ในกรณีนี้เกิดจากการการกำหนดไวยากรณ์ไว้น้อยเกินไปหรือไม่ครอบคลุมประโยคทั้งหมดที่อาจจะเกิดขึ้น ทำให้ระบบไม่สามารถวิเคราะห์ประโยคที่อยู่นอกเหนือกฎเกณฑ์นั้นได้