บทที่ 10

การเก็บค่าสถิติการเบี่ยงเบนทางไวยากรณ์

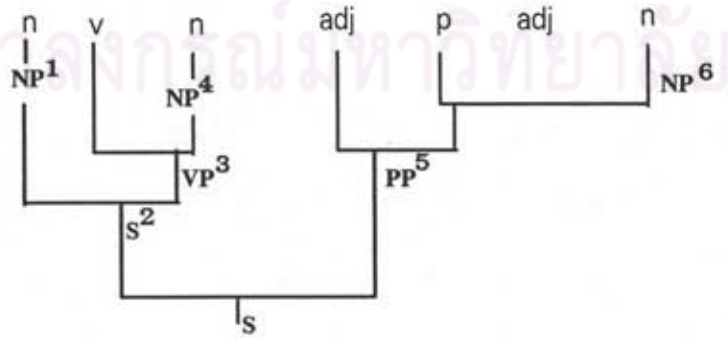
อย่างที่ได้กล่าวในตอนต้นแล้วว่า เป็นการยากที่จะกำหนดไวยากรณ์ทางภาษาให้ครอบคลุมการใช้งานทั้งหมดที่ต้องการ ค่าสถิติที่ได้จากการประมวลผลข้อมูลที่ผิดรูปแบบไวยากรณ์ หรือข้อมูลที่เบี่ยงเบนไปเหล่านี้จะช่วยให้เราสามารถปรับเปลี่ยนไวยากรณ์ที่มีให้เหมาะสมกับการใช้งานมากขึ้น ตัวอย่างต่างๆ ต่อไปนี้จะช่วยแสดงการเก็บค่าสถิติเหล่านี้ กำหนดให้ไวยากรณ์ไม่พึ่งบริบทที่ใช้ในการแจงส่วนคือ

(1) S -> NP VP
(2) S -> S PP
(3) NP -> n
(4) NP -> NP PP
(5) PP -> p NP
(6) VP -> v NP
(7) VP -> v PP

รูปที่ 10.1 แสดงไวยากรณ์ไม่พึ่งบริบทอย่างง่ายของภาษาไทยที่ใช้ในการอธิบาย

ถ้าประโยคผิดรูปแบบไวยากรณ์ที่ส่งเข้าประมวลผลคือ "วาง หนังสือ สีแดง โต๊ะ สีดำ" และประโยคที่ได้จากการแจงส่วนประโยคคือ "[n] วาง หนังสือ {สีแดง} [p] โต๊ะ {สีดำ}" เมื่อ [n] และ [p] เป็นองค์ประกอบที่ตัวแจงส่วนแทรกเข้าไปในประโยค {สีแดง} และ {สีดำ} เป็นองค์ประกอบที่ถูกละไปในการแจงส่วน เพราะเป็นชนิดคำที่ตัวแจงส่วนไม่รู้จัก (เทียบกับไวยากรณ์ปัจจุบันที่ใช้ในการแจงส่วน) โครงสร้างต้นไม้ของประโยคนี้สามารถเขียนได้ดังรูปที่ 10.2

[n] วาง หนังสือ {สีแดง} [p] {สีดำ} โต๊ะ



รูปที่ 10.2 แสดงโครงสร้างต้นไม้ของประโยคตัวอย่างที่กำหนด

พิจารณาจากรูปที่ 10.2 จากกฎการผลิตข้อ 1 เราจะได้ความสัมพันธ์ว่าองค์ประกอบ S^2 ประกอบด้วยส่วนนามวลี NP^1 กับส่วนกริยา VP^3 แต่ส่วนนามวลี NP^1 เกิดจากการแทรกเข้าไปของตัวแจงส่วน ข้อมูลที่มากจากประโยคจริง มีเพียงส่วนกริยา VP^3 เท่านั้น ดังนั้นในกรณีนี้กฎที่ได้จากการเบี่ยงเบนกฎการผลิตข้อ 1. S -> NP VP ก็คือ S -> VP (ตัดส่วน NP ทิ้งเพราะมาจากการแก้ไขของตัวแจงส่วน)

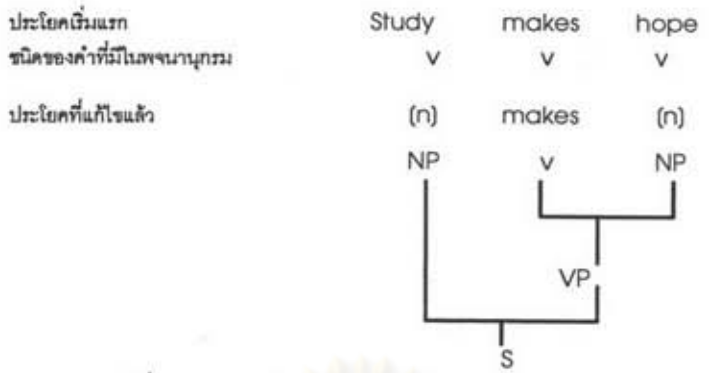
ลองพิจารณาที่ NP^4 ตามกฎการผลิตข้อ 3 (NP -> n) และ PP^5 ตามกฎการผลิตข้อ 5 (PP -> p NP) ในที่นี้ตัวแจงส่วนได้ละ คำวิเศษณ์ (adj) (สีแดง) โดยไม่ทำการแจงส่วนเพราะเป็นชนิดคำที่ไม่รู้จัก (อ้างอิงกับไวยากรณ์ ในรูปที่ 10.1) แต่ในความเป็นจริงส่วนนามวลี กับส่วนบุพบทวลี อาจมีองค์ประกอบเป็นแบบอื่นได้ ดังนั้นเราจึงเก็บค่าสถิติของส่วนนามวลี กับส่วนบุพบทวลี และองค์ประกอบที่เกี่ยวข้องกับการเบี่ยงเบนกฎที่พบในการประมวลผลไว้ด้วย ดังนั้นในกรณีนี้กฎที่บันทึกไว้เป็นค่าสถิติคือ NP -> n adj และ PP -> adj p NP จากตัวอย่างในกรณีนี้เราไม่ทราบแน่นอนว่า adj เกิดจากการเบี่ยงเบนกฎข้อใด เพราะเป็นชนิดคำใหม่ที่ไม่มีในกฎมาก่อน จึงต้องเก็บข้อมูลการเบี่ยงเบนไว้ทั้ง 2 แบบ ในการใช้งานจริง ถ้ามีข้อมูลของการเกิดในแบบใดแบบหนึ่งมากกว่าแบบอื่น ก็จะช่วยให้เราทราบว่าควรจะเพิ่มเติมกฎของคำวิเศษณ์ (adj) นี้อย่างไร

ด้วยวิธีการต่างๆ เหล่านี้ สามารถเก็บค่าสถิติได้ดังตารางที่ 10.1

ข้อมูลทางสถิติที่เก็บ	ความถี่	กฎที่ถูกเบี่ยงเบน
S => VP	1	S => NP VP
NP => n adj	2	NP => n
PP => adj p NP	1	PP => p NP
PP => NP	1	PP => p NP

ตารางที่ 10.1 แสดงค่าสถิติที่ได้จากการแก้ไขประโยค "วาง หนังสือ สีแดง โต๊ะ สีดำ"

ตัวอย่างที่ผ่านมาแสดงการเก็บค่าสถิติที่เกิดจากการแก้ไขการผิดไวยากรณ์ที่เกิดจากการแทรก และการขาดหายไปขององค์ประกอบในประโยค ตัวอย่างต่อไปข้างล่างแสดงการเก็บข้อมูลจากการแก้ไขข้อผิดพลาดจากการแทนที่องค์ประกอบในประโยค ข้อมูลที่ได้จากการแก้ไขประโยคในลักษณะนี้ นอกจากจะเป็นค่าสถิติของกฎที่ถูกเบี่ยงเบนแล้ว ยังใช้ช่วยในการปรับข้อมูลชนิดของคำในพจนานุกรมอีกด้วย ตัวอย่างเช่น ถ้าประโยคที่ทำการแจงส่วนคือ "Study makes hope" (อ้างอิงกับไวยากรณ์ ในรูป 10.1 เช่นกัน) ถ้าข้อมูลในพจนานุกรมของเราเก็บชนิดคำของคำทั้ง 3 ไว้ว่าเป็นคำกริยาทั้งหมด ประโยคนี้จะผิดรูปแบบไปจากไวยากรณ์ที่กำหนดทันที และประโยคที่ผ่านการแก้ไขโดยวิธีการแทนที่องค์ประกอบในประโยค จะแก้ไขประโยคใหม่เป็น "[n] makes [n]" (ประโยคที่ผ่านการแก้ไขแล้ว อาจจะเป็นไปได้หลายแบบ แต่เลือกกรณีนี้เพื่อใช้ในการอธิบาย) โครงสร้างต้นไม้ของประโยคนี้แสดงในรูปที่ 10.3



รูปที่ 10.3 แสดงโครงสร้างต้นไม้ของประโยค "[n] makes [n]"

ในกรณีนี้ค่าสถิติที่เก็บต้องใช้ข้อมูลจริงที่ได้จากประโยคที่นำมาประมวลผล จะไม่ใช้ข้อมูลใหม่ที่ได้จากการแทนที่ชนิดของคำลงในประโยค ค่าสถิติที่ได้จากรูปที่ 10.3 บันทึกในตารางที่ 10.2

ข้อมูลทางสถิติที่เก็บ	ความถี่	กฎที่ถูกเบี่ยงเบน
S => vVP	1	S => NP VP
VP => vv	1	VP => v NP

ตารางที่ 10.2 แสดงค่าสถิติที่ได้จากการแก้ไขประโยค "Study makes hope"

และค่าสถิติที่จะนำไปใช้ในการปรับปรุงพจนานุกรมคือ

ข้อมูลทางสถิติที่เก็บ	ความถี่	ข้อมูลเดิมที่มี
Study (n)	1	Study (v)
hope (n)	1	hope (v)

ตารางที่ 10.3 แสดงค่าสถิติของชนิดคำใหม่ที่พบจากการแก้ไขประโยค "Study makes hope"

จะเห็นได้ว่าข้อมูลทางสถิติเหล่านี้สามารถบอกเราได้ว่า กฎเกณฑ์ทางไวยากรณ์ที่เราใช้ในการแจงส่วนในปัจจุบันนั้นเหมาะสมกับงานที่ใช้อยู่หรือไม่ เช่นจากตารางที่ 10.1 ถ้าข้อมูลสถิติการพบการผิดไวยากรณ์ในลักษณะของ S -> VP มีความถี่สูงมาก ก็แสดงว่าไวยากรณ์ที่เราใช้อยู่ในขณะนี้ยังไม่ครอบคลุมกฎที่เหมาะสมทั้งหมด ยังขาดกฎที่ประโยคขึ้นต้นด้วยคำกริยา เช่นประโยคคำสั่ง เป็นต้น หรือในกรณีที่มีชนิดคำใหม่ๆ เข้ามาเช่นคำวิเศษณ์ (adj) ข้อมูลสถิติเหล่านี้ก็จะช่วยแนะนำเราได้ว่า ควรปรับปรุงกฎการผลิตที่มีอย่างไร เพื่อให้เหมาะสมกับชนิดคำใหม่นี้ เพื่อนำไปสู่การทำงานที่มีประสิทธิภาพมากขึ้น