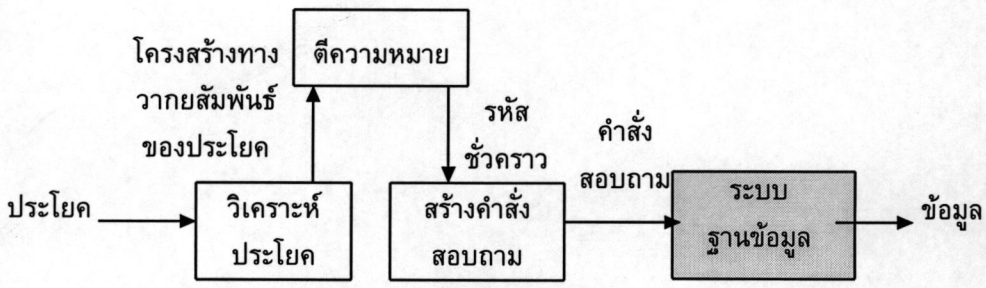
บทที่ 2

การแปลงประโยคภาษาไทยไปเป็นคำสั่งสอบถาม

ส่วนฟรอนต์เอนด์ของระบบฐานข้อมูลที่ใช้ภาษาธรรมชาติเป็นตัวเชื่อมประสาน คือ การใช้ภาษามนุษย์ในการสอบถามแทนภาษาคำสั่งสอบถามที่มีลักษณะที่ยากต่อความเข้าใจในการใช้งาน โดยเป็นการพยายามทำให้การสอบถามข้อมูลจากระบบฐานข้อมูลดูน่าใช้มากขึ้น (User Friendly) วิธีในการรับคำสั่งอาจใช้วิธีพิมพ์ หรือพูด เป็นต้น การแปลงประโยคภาษาไทยให้เป็นคำสั่งสอบถามโดยหลักประกอบด้วย 3 ส่วน คือ ส่วนวิเคราะห์ประโยค ส่วนแปลความหมาย และส่วนสร้างประโยคสอบถาม



รูปที่ 2.1 ส่วนประกอบของตัวแปลงประโยคเป็นคำสั่งสอบถาม

ส่วนวิเคราะห์ประโยคเป็นส่วนที่ทำการกระจายคำในประโยคออกมาเป็นโครงสร้างทางวากยสัมพันธ์ของประโยคในรูปของโครงสร้างต้นไม้ที่ประกอบไปด้วยคลังคำ (Lexicon) โดยกำหนดให้คำกริยาเป็นส่วนรากต้นไม้และให้ส่วนอื่นๆ เช่น นามวลี และส่วนประกอบของกริยาวลี เป็นกิ่งย่อยของคำกริยา ซึ่งทำให้ง่ายต่อการวิเคราะห์ การทำให้เกิดผลของโครงสร้างต้นไม้อาจจัดทำขึ้นมาให้อยู่ในรูปแบบต่างๆกันเช่น ต้นไม้แบบค้นทวิภาค โครงสร้างคุณสมบัติ (Feature Structure) เป็นต้น โดยขึ้นอยู่กับความเหมาะสมในการประยุกต์ใช้

ส่วนแปลความหมายเป็นส่วนที่วิเคราะห์ความหมายของโครงสร้างทางวากยสัมพันธ์ของประโยคโดยมีฐานความรู้ในการช่วยแปลความหมายของคำในประโยคว่ามีความหมายว่าอย่างไรและคำนั้นกล่าวถึงข้อมูลประเภทใดอยู่ ซึ่งการแปลความหมายจะพิจารณาประโยคจากการไล่ไปตามโครงสร้างต้นไม้ แล้วทำการแปลงคำต่าง ๆในประโยคให้เป็นรหัสและจัดให้อยู่ในรูปแบบที่ง่ายต่อการนำไปแปลคำสั่งเพื่อสร้างเป็นคำสั่งสอบถาม

ส่วนสร้างคำสั่งสอบถามเป็นขั้นตอนสุดท้ายที่ทำการแปลงประโยคไปเป็นคำสั่งสอบถาม ก่อนที่จะส่งคำสั่งสอบถามไปยังฐานข้อมูล โดยจะทำการแปลงรหัสที่ได้จากการแปลความหมายเป็นคำสั่งสอบถาม การที่กำหนดให้ส่วนสร้างคำสั่งสอบถามและส่วนแปลความหมายแยกออกจากกันเนื่องจากคำสั่งสอบถามสำหรับระบบการจัดการฐานข้อมูลมีความแตกต่างกันแต่ที่นิยมใช้กันในท้องตลาดและเป็นมาตรฐานคือภาษาสอบถามที่ชื่อว่าเอสคิวแอล (SQL - Structure Query Language) ถึงแม้ว่าเอสคิวแอลจะเป็นภาษาสอบถามมาตรฐานก็ตามแต่ระบบการจัดการฐานข้อมูลที่รองรับเอสคิวแอลแต่ละตัวมีรายละเอียดของเอสคิวแอลแตกต่างกันเล็กน้อย ทั้งนี้เพื่อต้องการเพิ่มประสิทธิภาพให้กับคำสั่งสอบถามนั้น

การวิเคราะห์ประโยคภาษาไทย

การวิเคราะห์ประโยคเป็นการนำคำที่ประกอบเป็นประโยคมากระจายให้เป็นโครงสร้างทางวากยสัมพันธ์ซึ่งอยู่ในรูปโครงสร้างต้นไม้ของคลังคำที่มีสารสนเทศเพื่อใช้ในการวิเคราะห์ในขั้นตอนต่อไป แต่เนื่องจากการเขียนประโยคภาษาไทยไม่มีการเว้นช่องว่างระหว่างคำเพื่อเป็นการกำหนดเขตของคำไว้อย่างชัดเจน ดังนั้นก่อนทำการวิเคราะห์กระจายต้องมีการประมวลผลด้วยการแบ่งคำเสียก่อน แล้วจึงทำการวิเคราะห์กระจาย

ส่วนการแบ่งคำ

ตั้งแต่อดีตได้มีการวิจัยการแบ่งคำในประโยคภาษาไทยอย่างมากมาย ทำให้เกิดวิธีการแบ่งคำหลากหลายวิธี เช่น การแบ่งคำด้วยกฎ การแบ่งคำด้วยพจนานุกรม การแบ่งคำด้วยพจนานุกรมและกฎ เป็นต้น การแบ่งคำดังกล่าวนั้นเป็นลักษณะการแบ่งคำเพื่อใช้ในงานจัดทำเอกสารเสียส่วนใหญ่ ซึ่งการแบ่งคำประเภทนี้จะเป็นการแบ่งคำให้ดูสวยงามและง่ายต่อการอ่านผลลัพธ์ที่ได้จึงมีทั้งคำและที่เป็นส่วนหนึ่งของคำ แต่ในการกระจายประโยคเป็นนำคำมาใช้ในการวิเคราะห์มากกว่าที่จะนำส่วนประกอบของคำมาใช้ในการวิเคราะห์ ดังนั้นส่วนวิเคราะห์กระจายจึงต้องการส่วนแบ่งคำที่ทำการแบ่งคำทั้งคำออกมามากกว่าส่วนประกอบของคำ

คำทุกคำที่ผ่านการแบ่งคำแล้วจำเป็นอย่างยิ่งที่จะต้องมีข้อมูลในลักษณะของคลังคำ ทั้งนี้เพราะการวิเคราะห์กระจายต้องใช้สารสนเทศในคลังคำของคำนั้นประกอบในการวิเคราะห์ โดยทั่วไปขนาดของฐานข้อมูลของคลังคำต่องานของระบบฐานข้อมูลหนึ่งจะมีขนาดไม่ใหญ่มากนัก เนื่องจากการวิเคราะห์กระจายประโยคเพื่อการสอบถามนั้นจำกัดแค่เฉพาะเรื่องที่มีอยู่ในฐานข้อมูลเท่านั้น

ส่วนวิเคราะห์กระจาย

ส่วนวิเคราะห์กระจายเป็นส่วนที่ทำการกระจายประโยคที่ถูกแบ่งคำแล้วจากส่วนแบ่งคำ ในที่นี้อนุมานว่าการแบ่งคำสามารถแบ่งได้อย่างถูกต้อง โดยจะทำการวิเคราะห์ 2 ส่วนด้วยกันคือ ส่วนวากยสัมพันธ์ (Syntactics) และส่วนอรรถศาสตร์ (Semantics) สำหรับทฤษฎีเอชพีเอสจี (HPSG - Head-driven Phrase Structure Grammar) Pollard C. (1987) ซึ่งเป็นไวยากรณ์แบบฟอร์มัลลิซึมประเภทหนึ่ง (Formalism Grammar) จะทำการพิจารณาทั้งสองส่วนไปพร้อมกัน เนื่องจากการเก็บสารสนเทศในคลังคำอยู่ในรูปของวัตถุเชิงวากยสัมพันธ์และเชิงอรรถศาสตร์ (SYNSEM - Syntactics Semantics Object) ในขณะที่โครงสร้างองค์ประกอบของประโยค (Constituent Structure) เกิดขึ้นจากการประกอบคำตามกฎไวยากรณ์ที่กำหนดไว้โดยมีการนำข้อมูลในคลังคำมาประมวลผลด้วยการดำเนินการที่เรียกว่า ยูนิฟิเคชัน (Unification) ซึ่งเป็นการดำเนินการในการรวบรวมโครงสร้างคลังคำมาประกอบเป็นโครงสร้างองค์ประกอบ ซึ่งจะมีลักษณะคล้ายกับการตรวจสอบข้อมูลในคลังคำว่าสอดคล้องกับประโยคหรือไม่ ถ้าไม่สามารถดำเนินการได้จะทำให้ทราบว่าคลังคำที่นำมาประกอบเป็นประโยคนั้นไม่ถูกต้อง สาเหตุอาจเนื่องมาจากการเลือกชนิดของคลังคำไม่เหมาะสม หรือการแบ่งคำไม่ถูกต้อง หรือประโยคที่รับเข้ามาอยู่ในรูปที่ไม่ถูกต้อง (ill-formed) เป็นต้น ดังนั้นส่วนที่ทำการวิเคราะห์กระจายประโยคจึงเป็นส่วนที่สามารถบอกให้ระบบได้รับทราบว่าประโยคนั้นๆสามารถจัดเป็นประโยคได้หรือไม่ โดยมีความถูกต้องทั้งในเชิงวากยสัมพันธ์และเชิงอรรถศาสตร์

ผลลัพธ์ที่ได้จากการวิเคราะห์ประโยคคือโครงสร้างองค์ประกอบประโยค ซึ่งถูกจัดให้อยู่ในรูปโครงสร้างต้นไม้ เหตุที่เป็นเช่นนี้เนื่องจากโครงสร้างต้นไม้สามารถนำมาอธิบายความสัมพันธ์ของคำในประโยคได้ง่าย และยังนำมาใช้กับงานด้านภาษาศาสตร์คอมพิวเตอร์ (Computational Linguistics) ได้อีกด้วย

การแปลความหมาย

การแปลความหมายเป็นการแปลงคำในประโยคโดยใช้ฐานความรู้ที่อาจอยู่ในรูปของตารางหรือระบบฐานข้อมูลเพื่อใช้ในการเปรียบเทียบคำสั่ง หรือค่าคงที่ หรือการอ้างอิงถึงรายการในระบบฐานข้อมูลที่ใช้อยู่ เป็นต้น แล้วทำการเปลี่ยนเป็นรหัสที่เหมาะสมเพื่อนำไปใช้ในส่วนของการสร้างคำสั่งสอบถามต่อไป การที่จะแปลคำแต่ละคำในประโยคที่อยู่ในรูปของโครงสร้างองค์ประกอบของประโยคซึ่งในที่นี้คือรูปของโครงสร้างต้นไม้ ตัวอัลกอริทึมที่ทำการแปลจะต้องทำการไล่ไปตามจุดแตกกิ่ง (Node) ในโครงสร้างต้นไม้เพื่อทำการวิเคราะห์ความสัมพันธ์ของคำต่างๆและลักษณะเฉพาะของคำที่มีการอธิบายไว้ด้วยคำแวดล้อมที่ประกอบในประโยค เพื่อค้นหาความหมายที่แท้จริงว่าเป็นการกล่าวอ้างถึงอะไรในระบบฐานข้อมูล และมีเงื่อนไขอะไรบ้างในการขอค้นคืนข้อมูล

สิ่งที่ได้จากการแปลความหมายคือรหัสชั่วคราว (Intermediate Code) เพื่อที่จะนำไปเป็นข้อมูลในการสร้างคำสั่งสอบถามต่อไป โดยรหัสนี้สามารถจัดให้อยู่ในรูปโครงสร้างอื่นที่แตกต่างจากโครงสร้างองค์ประกอบ หรือโครงสร้างองค์ประกอบที่ประกอบด้วยสารสนเทศที่ได้จากการแปลความหมายแล้ว ซึ่งไม่ว่าจะอยู่ในรูปแบบของโครงสร้างใดก็ตามคุณสมบัติที่สำคัญของรหัสที่จะต้องมีคือขนาดของสารสนเทศที่เพียงพอแก่การนำไปสร้างคำสั่งสอบถามแล้วได้ผลลัพธ์ตรงกับคำสั่งหรือคำถาม วิธีที่ดีที่สุดที่สามารถนำไปสร้างคำสั่งสอบถามได้หลากหลายประเภทคือการส่งข้อมูลทั้งหมดเท่าที่มีไปยังส่วนการสร้างคำสั่งสอบถาม ซึ่งวิธีนี้จะไม่เหมาะกับระบบที่ต้องการประสิทธิภาพในด้านของเวลาในการประมวลผล ทั้งนี้เนื่องจากระบบจะต้องมีการประมวลผลสารสนเทศที่มีทั้งหมดก่อนที่จะพิจารณาว่าควรสร้างคำสั่งสอบถามอย่างไร แต่จุดเด่นคือส่วนของการสร้างคำสั่งสอบถามสามารถทำการเพิ่มประสิทธิภาพ (Optimization) คำสั่งสอบถามได้อย่างถูกต้อง อันมีผลต่อการค้นคืนข้อมูลของระบบฐานข้อมูลเป็นอย่างมาก ในทางตรงข้ามการที่ละข้อมูลที่ไม่เป็นจุดสำคัญบางอย่างที่ได้จากโครงสร้างองค์ประกอบจะช่วยลดเวลาในการประมวลผลได้เป็นอย่างดี แต่วิธีนี้จะเหมาะกับการสร้างคำสั่งสอบถามได้น้อยประเภทเท่านั้น เนื่องจากสารสนเทศบางอย่างที่ละไว้อาจมีความสำคัญต่อการสร้างคำสั่งสอบถามประเภทอื่นๆ

การสร้างคำสั่งสอบถาม

ในส่วนนี้จะเป็นการแปลงรหัสที่ได้จากส่วนแปลความหมายให้เป็นคำสั่งสอบถามเฉพาะของแต่ละระบบฐานข้อมูล ขั้นตอนการทำงานของการสร้างคำสั่งสอบถามจะมีลักษณะ

คล้ายกับส่วนแปลความหมายในแง่ของการพิจารณาโครงสร้าง เพียงแต่มีการทำงานที่ง่ายกว่า คือส่วนแปลความหมายจะต้องทำการพิจารณาคำและคำแวดล้อมที่มีอยู่ในประโยคที่อยู่ในรูปโครงสร้างต้นไม้โดยอาจเกิดการย้อนรอยเพื่อที่จะสามารถแปลความหมายได้ ซึ่งต่างกับส่วนการสร้างคำสั่งสอบถามที่ทำการเปรียบเทียบความหมายในรหัสกับคำสั่งสอบถามได้โดยตรงจึงไม่มีความจำเป็นที่จะต้องทำให้อัลกอริทึมมีความซับซ้อนมาก สิ่งที่แตกต่างจากส่วนแปลความหมายคือความสามารถในการเพิ่มประสิทธิภาพคำสั่งสอบถาม ซึ่งส่วนสร้างคำสั่งสอบถามจะต้องทราบว่าคำสั่งสอบถามรูปแบบใดที่มีประสิทธิภาพในการค้นคืนข้อมูลจากระบบฐานข้อมูล และจะต้องทราบว่าวิธีการจัดลำดับในการค้นคืนอย่างไรให้มีประสิทธิภาพเมื่อต้องมีการค้นคืนมากกว่า 1 ครั้ง

ในงานวิจัยนี้คำสั่งสอบถามที่ใช้คือเอสคิวแอลซึ่งมีรูปแบบดังนี้

```
SELECT * | {[DISTINCT | ALL] <VALUE EXPRESSION>,…}
   FROM {<TABLE NAME> [<ALIAS>]},…
   [ [WHERE <PREDICATE>]
     [GROUP BY {<COLUMN NAME> | <INTEGER>},…]
     [HAVING <PREDICATE>]
     [ORDER BY {<COLUMN NAME> | <INTEGER>},…] ];
```

ส่วนที่เป็นเงื่อนไขของคำสั่งสอบถามคือ {WHERE} {GROUP BY} {HAVING} และ {ORDER BY} จะมีการใช้ที่แตกต่างกัน และอาจจะได้หากไม่มีเงื่อนไขในการขอค้นคืนข้อมูล ซึ่งต่างจากส่วนที่เป็นข้อมูลที่ต้องการคือต้องมีการระบุว่าการข้อมูลที่อยากทราบได้มาจากระเบียนใดในตารางข้อมูลใดในฐานข้อมูล