บทที่ 2
สรุปสาระสำคัญจากเอกสารที่เกี่ยวข้อง

การคัดแยกข้อมูล (Information Extraction) เป็นกระบวนการในการคัดแยกข้อมูลที่เป็นประโยชน์จากเอกสารที่เกี่ยวข้องกับข้อมูลที่สนใจโดยไม่จำเป็นต้องเข้าใจความหมายอย่างสมบูรณ์ โดยได้มีการนำเสนองานทางด้านนี้เป็นครั้งแรกที่งาน Message Understanding Conferences (MUCs) (Cowie and Lehnert, 1996) เทคโนโลยีเกี่ยวกับการคัดแยกข้อมูลจากภาษาธรรมชาติ (Natural Language) นั้นเกิดขึ้นจากความต้องการให้กระบวนการในการคัดแยกความรู้ที่มีการศึกษาเป็นพิเศษนั้นมีประสิทธิภาพมากขึ้น จึงได้มีความพยายามเป็นอย่างมากเพื่อทำการแก้ปัญหาในส่วนของการวิเคราะห์คำในประโยคสำหรับประโยคที่มีความยาวมาก ๆ โดยใช้พาสเซอร์ (parser) ซึ่งจะมีประโยชน์ในการตัดคำที่ไม่มีความเกี่ยวข้องกับเรื่องที่ต้องการได้ ดังนั้นเทคโนโลยีในการคัดแยกข้อมูลจึงมุ่งเน้นไปยังส่วนที่มีความเกี่ยวข้องกับข้อมูลที่ต้องการโดยจะไม่สนใจในส่วนที่ไม่เกี่ยวข้อง การคัดแยกข้อมูลนั้นสามารถประเมินผลได้ด้วยค่ารีคอล (recall) ซึ่งเป็นการวัดความสมบูรณ์ของข้อมูลที่มีความเกี่ยวข้องที่สามารถคัดแยกมาได้และค่าพรีซิสชัน (precession) ซึ่งเป็นการวัดความถูกต้องของข้อมูลที่คัดแยกได้

ในส่วนต่อไปนี้จะกล่าวถึงงานทางด้านการคัดแยกข้อมูลที่ได้ศึกษาและมีความเกี่ยวข้องกับวิทยานิพนธ์ที่นำเสนอ มีเนื้อหาดังนี้

ในปี ค.ศ.1993 ได้มีการนำเสนอระบบฟาสตัส (FASTUS: Finite State Automata-base Text Understanding System) ซึ่งมีโมเดลเป็นแบบนอนดีเทอร์มินิสติกไฟไนท์สเทต (nondeterministic finite-state) ทำให้สามารถแบ่งประโยคออกเป็นองค์ประกอบต่าง ๆ ของประโยคได้ เช่น นามวลี กริยาวลี ฟาสตัสได้ถูกพัฒนาขึ้นเพื่อทำการปรับปรุงระบบการคัดแยกข้อมูลให้มีความรวดเร็วและถูกต้องยิ่งขึ้น

กระบวนการทำงานของระบบฟาสตัสประกอบไปด้วย 4 ส่วนคือ

1. Triggering เป็นส่วนที่เกี่ยวข้องกับการกำหนดคำที่เป็นทริกเกอร์เวิร์ด จากหนึ่งรูปแบบ (pattern) ของข้อมูลที่สนใจ จะมีเพียงหนึ่งทริกเกอร์เวิร์ดเท่านั้นที่จะถูกกำหนดขึ้นเพื่อใช้ในการค้นหาข้อมูลที่ต้องการ
2. Recognizing phrases เป็นส่วนที่จะทำการกำหนดกลุ่มคำต่าง ๆ เช่น กลุ่มคำนาม กลุ่มคำกริยา กลุ่มคำบุพบท เป็นต้น

3. Recognizing Patterns สำหรับข้อมูลที่ป้อนเข้าในส่วนนี้คือรายการของวลีต่าง ๆ ที่ได้จากส่วนที่สอง โดยจะได้ผลลัพธ์เป็นคู่ของคำและชนิดของกลุ่มคำนั้น

4. Merging Incidents รวมเหตุการณ์ที่พบเข้ากับอีกเหตุการณ์จากประโยคที่เหมือนกัน

การสืบค้นหากลุ่มของคำนามในประโยคนั้นจะทำการสืบค้นจากด้านซ้ายของประโยคไปด้านขวาของประโยค โดยการทำงานของระบบนี้จะต้องมีชุดข้อมูลที่ใช้ในการเรียนรู้ (training data) เพื่อใช้ในการจดจำวลี (recognizing phrases) และรูปแบบ (recognizing patterns) ของข้อมูลที่สนใจ การทดสอบใช้ทดสอบกับคลังข้อมูล MUC-4 ซึ่งเป็นแหล่งข้อมูลที่เป็นบทความเกี่ยวกับขบวนการก่อการร้าย และทำการประเมินผลความถูกต้องที่ได้ด้วยค่ารีคอล (recall) ซึ่งได้ผล 44 เปอร์เซ็นต์ และค่าพรีซิสชัน (precision) ได้ผล 52 เปอร์เซ็นต์ จากผลการประเมินที่ได้สรุปได้ว่าเกิดจากข้อมูลที่นำมาใช้ในการทดสอบนั้นมีประโยคที่มีความยาวมาก ๆ อยู่เป็นจำนวนมากในบทความนั้นและปัญหาเกี่ยวกับโครงสร้างของการบรรยายทำให้ได้ข้อมูลที่ผิดพลาดได้

(Douglas E. Appelt et al, 1993)

ในระยะเวลาสิบปีที่ผ่านมานั้นงานในด้านการคัดแยกข้อมูลจากภาษาธรรมชาตินั้นมีความก้าวหน้าเป็นอย่างมาก จึงได้มีการนำไปประยุกต์ใช้กับข้อมูลในด้านต่างๆ เช่น ด้านเศรษฐกิจ ด้านการสงคราม ซึ่งในปัจจุบันนี้ได้เริ่มมีความพยายามในการจะนำมาประยุกต์ใช้กับข้อมูลทางด้านไบโอเมดดิคอล (Biomedical) มากขึ้น ดังนั้นระบบคัดแยกที่พัฒนาขึ้นจึงมีความเหมาะสมกับแหล่งข้อมูลที่ต่างกันออกไปตามจุดประสงค์ที่จะนำไปใช้คัดแยกกับข้อมูลประเภทใด

จากกระบวนการทางภาษาธรรมชาติ การคัดแยกข้อมูลเป็นเรื่องที่น่าสนใจด้วยเหตุผลหลายประการด้วยกัน ซึ่งเหตุผลข้อหนึ่งคือ การทำงานในระบบการคัดแยกข้อมูลนั้นสามารถประเมินผลและเปรียบเทียบกับเกณฑ์มาตรฐานที่ได้จากการทำงานของมนุษย์ได้ ซึ่งเป็นโอกาสที่ดีสำหรับผู้ที่ทำการวิจัยเกี่ยวกับภาษาธรรมชาติ ระบบการคัดแยกข้อมูลนั้นจะมีลักษณะของกระบวนการก่อนการทำงานที่เป็นพื้นฐานของระบบดังนี้

- กระบวนการในการระบุชนิดของคำ ซึ่งเป็นขั้นต้นในการจดจำกลุ่มต่าง ๆ ของวลีในประโยค
- กฎที่ใช้สำหรับจุดประสงค์พิเศษในการจดจำความหมายประเภทต่าง ๆ ของวลี ซึ่งรวมถึงชื่อบริษัท สถานที่ ชื่อผู้คน เงินตรา และชื่อเครื่องมือต่าง ๆ เป็นต้น

สำหรับเอกสารที่มีอยู่ทั่วไปนั้นพบว่ามีข้อความที่ปรากฏชื่อเฉพาะต่าง ๆ รูปแบบของวันที่ รูปแบบของค่าวัดต่าง ๆ ปรากฏอย่างมากมายในเอกสาร กลุ่มของวลีดังกล่าวก็จะต้องถูกนำไปใช้

ซึ่งโดยปรกติแล้วก็ไม่ได้เป็นปัญหาสำหรับการอ่านของมนุษย์ แต่พบว่าเป็นปัญหาให้เกิดความกำกวมต่อระบบคัดแยกข้อมูล ลักษณะและโครงสร้างของแหล่งข้อมูลที่นำมาใช้ในการคัดแยกข้อมูลจากภาษาธรรมชาตินั้นจะมีผลต่อความถูกต้องในการคัดแยกข้อมูล ดังนั้นระบบคัดแยกข้อมูลที่พัฒนาขึ้นในแต่ละยุคจึงมีความเหมาะสมที่จะใช้กับแหล่งข้อมูลที่แตกต่างกันออกไปตามลักษณะและโครงสร้างของแหล่งข้อมูลนั้น ๆ เนื่องจากข้อมูลแต่ละชนิดย่อมมีโครงสร้างและลักษณะของประโยคแตกต่างกันออกไป

โดยทั่วไปแล้วจุดประสงค์ของระบบการคัดแยกข้อมูลนั้นมีมากกว่าการเติมข้อมูลลงในแบบเท่านั้น ดังนั้นจึงจะแสดงการพัฒนาระบบคัดแยกตั้งแต่ยุคแรก ๆ ดังนี้

- ครั้งแรกที่มีการนำเสนองานของระบบคัดแยกข้อมูลนั้น เป็นการทำงานกับเอกสารที่ไม่มีการจำกัดขอบเขตของหัวข้อที่ใช้ โดยใช้ newswire network เป็นแหล่งข้อมูลเกี่ยวกับข่าวต่าง ๆ โปรแกรมจะทำการควบคุมข้อมูลโดยใช้สคริปต์แบบง่าย ๆ ในการออกแบบเพื่อครอบคลุมเรื่องราวทั้งหมด โปรแกรมจะทำการค้นหาแต่ละเรื่องที่สามารถเข้ากันได้กับสคริปต์ที่เกี่ยวข้องโดยใช้คำสำคัญและการวิเคราะห์ประโยคเป็นหลักพื้นฐาน
- ก่อนปี ค.ศ. 1970 ได้มีการนำเสนอการคัดแยกข้อมูลจากเอกสาร โดยจะทำการค้นหาเพื่อแปลงข้อมูลสรุปของผู้ป่วยที่ได้มีการจำหน่ายออกให้อยู่ในรูปแบบที่เหมาะสมสำหรับการใช้เป็นข้อมูลนำเข้าในฐานข้อมูล
- ในปี ค.ศ. 1980 มีการคัดแยกข้อมูลเกี่ยวกับการบินของดาวเทียมจากรายงานที่มีการทำขึ้นจากทั่วโลก แต่ระบบนี้ยังมีข้อจำกัด เกี่ยวกับประโยคเดี่ยวและไม่มีวิธีการที่เพียงพอในการคัดแยกให้ได้การอธิบายเหตุการณ์ที่สมบูรณ์
- ในปี ค.ศ. 1981 ได้มีการพัฒนาระบบการคัดแยกที่ทำการคัดแยกโครงสร้างซึ่งเป็นที่ยอมรับจากเอกสารที่มีการอธิบายเกี่ยวกับพืชและสัตว์

ความแตกต่างที่สำคัญระหว่างระบบที่ได้รับการพัฒนาในปี ค.ศ. 1980 และปีค.ศ. 1990 คือจำนวนรวมของเวลาที่ใช้เป็นจำนวนมากและพลังงานที่ต้องใช้ในการเก็บรวบรวมเอกสารที่เกี่ยวข้อง และสร้างเซ็ตของรูปแบบเพื่อนำไปสร้างแหล่งข้อมูลที่ใช้ในการทดสอบเนื้อหาและตรวจสอบคำตอบที่ถูกต้อง ซึ่งทำให้สามารถประเมินค่าการทำงานของระบบการคัดแยกข้อมูลได้

(Jim Cowie and Wendy Lehnert, 1996)

เอกสารนี้ได้ทำการรวมรวมชนิดของรูปแบบ (extraction pattern) ที่ใช้สำหรับการคัดแยกข้อมูล ระบบการคัดแยกข้อมูลนั้นจะขึ้นอยู่กับกลุ่มของรูปแบบ (pattern) ที่ใช้ในการคัดแยกข้อมูล

ซึ่งจะมีรูปแบบแตกต่างกันออกไป จึงได้มีการรวบรวมชนิดของรูปแบบขึ้นเพื่อพิจารณาข้อเหมือนหรือข้อแตกต่างของรูปแบบในแต่ละกลุ่ม ซึ่งรูปแบบที่ใช้ในการคัดแยกข้อมูลนั้นถือว่าเป็นส่วนที่มีความสำคัญต่อระบบการคัดแยกข้อมูล

ระบบ AutoSlog นั้นจะมีรูปแบบที่เรียกว่า คอนเซ็ป (concepts) หรือ คอนเซ็ปโหนด (concept node) ซึ่งแต่ละคอนเซ็ปจะต้องมีทริกเกอร์เวิร์ด (trigger word) เพื่อใช้ในการคัดแยกข้อมูลที่ต้องการและตรงตามรูปแบบทางภาษา (linguistic pattern) ซึ่งสำหรับ AutoSlog จะมีทั้งหมด 13 รูปแบบด้วยกัน โดยทั่วไปแล้วทริกเกอร์เวิร์ดจะเป็นคำกิริยา และข้อมูลที่ต้องการจะเป็นคำนามที่อาจทำหน้าที่เป็นประธานหรือกรรม

ระบบ LIEP เป็นระบบการเรียนรู้ระบบหนึ่งที่สามารถสร้างกฎเพียงกฎเดียวสำหรับทุกรายการที่สนใจได้ เช่น ต้องการคัดแยกรายการที่เป็นเหยื่อและผู้กระทำ โดยที่รูปแบบที่ใช้นั้นจะประกอบไปด้วยกฎที่เป็นข้อบังคับเกี่ยวกับการสร้างประโยคและข้อบังคับที่เกี่ยวข้องกับความหมาย

ระบบ PALKA จะแสดงรูปแบบของโครงสร้างทางภาษาที่เรียกว่า “frame-phrasal pattern structure” ซึ่งประกอบไปด้วย meaning frame เป็นส่วนที่ใช้ในการแสดงรายการของข้อมูลที่ต้องการพร้อมกับความหมายที่เกี่ยวข้องกับแต่ละรายการ และ phrasal pattern เป็นส่วนที่แสดงการจัดลำดับเหตุการณ์ที่เกิดขึ้น รูปแบบของโครงสร้างทางภาษา (frame-phrasal pattern structure) ที่กำหนดขึ้นนี้ จะทำการรวมทั้งสองส่วนเข้าด้วยกันในภายหลัง ถ้า phrasal pattern นั้นตรงกับประโยคและมีความหมายเกี่ยวข้องกับ meaning frame ที่มีก็จะทำการคัดแยกข้อมูลที่ถูกต้องได้ ระบบ PALKA นั้นจะสามารถแสดงข้อมูลที่ถูกคัดแยกโดยใช้ทริกเกอร์เวิร์ดที่เป็นคำกิริยาเท่านั้นแต่ระบบ AutoSlog ทริกเกอร์เวิร์ดจะเป็นได้ทั้งคำกิริยาและคำนาม

ระบบ CRYSTAL จะมีคอนเซ็ปโหนดที่แสดงทั้งประธานและบุพบทวลีที่สามารถคัดแยกออกมาได้ถ้าตรงกับคำกิริยาและบุพบท สำหรับรูปแบบที่ใช้ใน CRYSTAL นั้นจะมีความชัดเจนมากกว่ารูปแบบที่ใช้ใน PALKA เพราะจะทำการคัดแยกข้อมูลเฉพาะคำที่ตรงกับรากของคำกิริยาเท่านั้น ต่อมาจึงได้มีการพัฒนาระบบ Webfoot ขึ้นมา เพื่อให้ CRYSTAL สามารถทำงานบนเว็บเพจได้ โดยทำการส่งเว็บเพจเข้าสู่ระบบ Webfoot แล้วจึงนำเอาระบบ CRYSTAL มาประยุกต์ใช้

ระบบ HASTEN จะทำการสร้างรูปแบบที่เรียกว่า “Egraphs” ซึ่งจะมีลักษณะคล้ายกันกับรูปแบบที่ใช้ในระบบ AutoSlog โดยที่เป้าหมายนั้นจะต้องเป็นนามวลีที่มีความหมายเป็นกรรมของประโยค ซึ่งจะต้องมีการตรงกันของส่วนที่เป็นโครงสร้างและส่วนแสดงความหมาย

ทุกระบบที่ได้มีการนำเสนอขึ้นนั้นจะต้องมีการใช้ข้อกำหนดที่เกี่ยวกับการสร้างประโยคและความหมายในการระบุข้อมูลที่สนใจ ซึ่งแต่ละวิธีก็มีข้อดีข้อเสียแตกต่างกันออกไป ระบบ LIEP และระบบ HASTEN นั้นจะทำการระบุวลีที่สนใจในขณะที่ระบบ AutoSlog ระบบ PALKA และระบบ CRYSTAL นั้นจะทำการกำหนดเฉพาะส่วนที่เกี่ยวข้องกับการสร้างประโยคที่มีวลีที่เป็นเป้าหมายเท่านั้น

(Ion Muslea, 1999)

เมื่อมีความต้องการในการคัดแยกข้อมูลที่มีอยู่เพิ่มมากขึ้น และมีวิธีหลากหลายวิธีที่จะใช้ในการทำงานร่วมกับข้อมูลที่เป็นภาษาธรรมชาติ จึงได้มีการนำเสนอวิธีในการคัดแยกข้อมูลที่มีพื้นฐานอยู่บนการทำพาสเซอร์ สำหรับระบบนี้พาสเซอร์จะทำการแปลงประโยคต่าง ๆ ที่ทำการอธิบายเหตุการณ์ให้อยู่ในรูปของโครงสร้างที่สร้างขึ้น โดยพิจารณาจากคำกิริยาที่แสดงเหตุการณ์ และความหมายที่ปรากฏ เช่น ประธาน และกรรม เป็นต้น การคัดแยกข้อมูลโดยทั่วไปสามารถทำได้โดยใช้วิธีแพตเทิร์นแมชชิง เมื่อมีการเปลี่ยนแปลงการนำเสนอโดยใช้พาสเซอร์ จึงใช้จำนวนแพตเทิร์นในการคัดแยกข้อมูลน้อยลง งานวิจัยนี้ได้มีการนำเสนอกระบวนการก่อนการเริ่มต้นในการแก้ปัญหาของการทำพาสเซอร์ วิธีที่นำเสนอคือกระบวนการ term recognize โดยการจัดคำเป็นนามวลี ทำให้พาสเซอร์สามารถจัดการกับวลีนั้นเหมือนกับเป็นคำหนึ่งคำ

ปัญหาของการคัดลอกข้อมูลจากเอกสารทางไบโอคือการที่คำกิริยานั้นสามารถมีได้หลายรูปแบบด้วยกัน ความสัมพันธ์กันระหว่างสารสองชนิดนั้นไม่ได้อยู่ในรูปแบบของ ประธาน-กิริยา-กรรม เสมอไป ดังนั้นถ้าเราต้องการที่จะคัดลอกข้อมูลที่เป็นการกระทำใด ๆ โดยตรงจะต้องใช้กฎจำนวนมากในการทำการคัดลอกข้อมูลนั้น รวมทั้งการแสดงลำดับขั้นตอนในการเกิดปฏิกิริยาต่อกัน ในแต่ละประโยคก็ทำให้ยากที่จะพบปฏิกิริยาที่เกิดขึ้นในประโยคจากการทำแพตเทิร์นแมชชิง เพราะว่าแต่ละปฏิกิริยาในประโยคต่างก็มีคำกิริยาในแพตเทิร์นที่แตกต่างกันออกไป ความสัมพันธ์กันในแต่ละปฏิกิริยานั้นสามารถคัดลอกออกมาได้ถ้ามองจากโครงสร้างของประโยคที่สมบูรณ์ จึงได้มีการนำเอาพาสเซอร์มาใช้ในการแก้ปัญหาดังกล่าวโดยการวิเคราะห์โครงสร้างทั้งหมดของประโยค ทำให้สามารถหาวลีในประโยคที่มีความสัมพันธ์กันได้ด้วยทำให้สามารถคัดลอกความสัมพันธ์กันระหว่างปฏิกิริยาในประโยคได้ งานนี้ใช้โพรเซสเซอร์สองตัวในการแก้ปัญหาความกำกวมในประโยคและปรับปรุงประสิทธิภาพ โพรเซสเซอร์ตัวหนึ่งจะใช้ในกระบวนการ term recognize ซึ่งจะใช้ในการระบุคำที่เป็นศัพท์เฉพาะทางเทคนิคและการจัดกลุ่มคำตามความหมาย

สำหรับระบบที่ทำการพัฒนาขึ้นนั้นจะนำไปใช้กับฐานข้อมูล MEDLINE โดยนำ HPSG พาสเซอร์มาใช้โดยไม่มีการปรับหลักไวยากรณ์หรือศัพท์เฉพาะสาขา ทำการทดลองกับคลังข้อมูลที่มีหมายเหตุอธิบายชื่อประกอบโดยนักชีววิทยา ทำการพาสเซอร์ประโยคจากคลังข้อมูล MEDLINE จำนวน 179 ประโยคด้วยเอ็กซ์เอชพีเอสจี (XHPSG) พาสเซอร์โดยไม่ผ่านกระบวนการก่อนการเริ่มต้น ได้ผลลัพธ์ที่ถูกต้องทั้งหมด 66 ประโยค สำหรับประโยคเหล่านี้จะมีจำนวนในการสร้างองค์ประกอบและเวลาในการพาสเซอร์ลดลง

(Akane Yakushiji et al, 2001)

เอกสารนี้ได้ทำการเปรียบเทียบหลักการระหว่างลิงค์แกรมมาพาสเซอร์และคอนเซอร์ดีเอฟจีพาสเซอร์ (Conexor FDG: Conexor Functional Dependency Grammar Parser) ซึ่งการทำพาสเซอร์นั้นเป็นส่วนสำคัญในกระบวนการที่มีความเกี่ยวข้องกับภาษาธรรมชาติ โดยพาสเซอร์ที่ดีนั้นจะต้องมีความสามารถในการวิเคราะห์โครงสร้างของประโยคได้อย่างถูกต้อง จึงได้มีการวิเคราะห์ความแตกต่างระหว่างพาสเซอร์ทั้งสองชนิดนี้ซึ่งสามารถจำแนกได้ตามหัวข้อดังนี้

Direction of dependency ลิงค์ของลิงค์แกรมมาถึงแม้ว่าจะเหมือนกับว่าขึ้นต่อกันจริง แต่อย่างไรก็ตามลิงค์แกรมมาจะมีชนิดของลิงค์ทางซ้ายและทางขวาที่แตกต่างกัน สำหรับคอนเซอร์-เอฟดีจีนั้นจะมีการแสดงทิศทางที่มีการขึ้นต่อกันเสมอ

Clausal heads โดยทั่วไปแล้วลิงค์แกรมมาจะเลือกส่วนที่อยู่หน้าสุดเป็นส่วนหัว (head) ของอนุประโยค แต่สำหรับคอนเซอร์เอฟดีจีนั้นจะเลือกกิริยาหลักเป็นส่วนหัวของอนุประโยค

Graph structures ลิงค์ของลิงค์แกรมมาจะมีการขึ้นต่อกันในการสร้างประโยคทั้งในระดับที่สามารถแสดงให้เห็นได้ (surface-syntactic) และในระดับที่มีความสำคัญ (deep-syntactic) ซ่อนอยู่ด้วย ลักษณะของโครงสร้างที่ได้จากลิงค์แกรมมาจะเป็นโครงสร้างต้นไม้ ซึ่งสำหรับคอนเซอร์ดีเอฟจีพาสเซอร์นั้นจะไม่มีลักษณะเป็นกราฟ

Conjunction ลิงค์แกรมมาจะวิเคราะห์ให้คำที่มีการเชื่อมกันนั้นเป็นส่วนหัวของวลีร่วมกัน แต่คอนเซอร์เอฟดีจีนั้นจะวิเคราะห์ให้ตัวแรกหรือตัวสุดท้ายของคำเชื่อมนั้นเป็นส่วนหัวของวลี

Dependency type ลิงค์แกรมมามีชนิดของลิงค์ประมาณ 90 ลิงค์ด้วยกัน คอนเซอร์เอฟดีจีมีความสัมพันธ์ทั้งหมด 32 ความสัมพันธ์ด้วยกัน

ในการเปรียบเทียบความถูกต้องระหว่างลิงค์แกรมมาและคอนเซอร์เอฟดีจีนั้นจะใช้ความสัมพันธ์กันตามหลักไวยากรณ์ (grammatical relation) เช่น ประธานและกรรมของประโยค

ซึ่งประกอบไปด้วยการประเมินความถูกต้องและความเร็ว สำหรับการประเมินความถูกต้องนั้นจะอยู่ภายใต้สมมติฐานที่ว่าความสัมพันธ์กันตามหลักไวยากรณ์นั้นจะสามารถสรุปได้จากโครงสร้างของประโยคหนึ่งที่สร้างขึ้นโดยขึ้นอยู่กับประโยคอื่นด้วยซึ่งจะได้จากพาสเซอร์ ในการประเมินผลจะให้ความสำคัญกับความสัมพันธ์ของ ประธาน กรรม ประโยคย่อยที่เป็นองค์ประกอบที่ทำให้สมบูรณ์โดยไม่มีประธานที่ชัดเจนและส่วนขยาย การเปรียบเทียบพาสเซอร์ทั้งสองชนิดในครั้งนี้จะไม่มีการเปรียบเทียบในส่วนของการให้ความหมาย สำหรับผลที่ได้จากการพาสเซอร์นั้น ในการพาสเซอร์ด้วยคอนเซอร์เอฟดีจีนั้นจะให้ผลเพียงหนึ่งโครงสร้างต่อหนึ่งประโยคแต่ลิงค์แกรมมาจะได้หลายโครงสร้างด้วยกัน ดังนั้นในการประเมินจึงเลือกเพียงโครงสร้างแรกมาใช้เท่านั้น

ผลจากการประเมินพบว่าค่ารีคอลและค่าพรีซิสชันของลิงค์แกรมมานั้นมีค่าต่ำกว่าคอน-เซอร์เอฟดีจี ซึ่งบางส่วนอาจเกิดจากลิงค์ของลิงค์แกรมมาส่วนมากอาจไม่ได้มีการเชื่อมต่อกับส่วนหัวของประโยคย่อย สำหรับการประเมินในส่วนของความเร็วในการพาสเซอร์นั้นลิงค์แกรมมาก็ใช้เวลามากกว่าที่คอนเซอร์เอฟดีจีใช้เนื่องจากว่าลิงแกรมมานั้นจะให้ผลของการพาสเซอร์มากกว่าคอนเซอร์เอฟดีจีซึ่งให้ผลเพียงแค่หนึ่งโครงสร้างเท่านั้นในขณะที่ลิงค์แกรมมาให้ผลทั้งหมด 10,509 โครงสร้าง จาก 505 ประโยค

นอกจากความถูกต้องในการพาสเซอร์แล้วยังจำเป็นที่จะต้องพิจารณาถึงความเป็นไปได้ในการพาสเซอร์ที่ผิดพลาดซึ่งมีผลต่อความสำเร็จในการใช้โปรแกรมที่เกี่ยวกับภาษาธรรมชาติด้วย จึงได้มีการประเมินประสิทธิภาพของพาสเซอร์ทั้งสองชนิดในส่วนนี้ด้วย โดยใช้โปรแกรมในการคัดแยกวลีเพื่อตอบคำถามในภาษาธรรมชาติ (Answer Extraction) ซึ่งเป็นโปรแกรมที่ใช้ทางภาษาธรรมชาติในการประเมินผลทำการประเมินในสองส่วนด้วยกันคือ ส่วนของคำที่มีความหมายเหมือนกันและส่วนของคำใกล้เคียง โดยค่ารีคอล ค่าพรีซิสชัน และค่าเอฟ-สกอร์ (F-score) ซึ่งจะถูกคำนวณเมื่อการคิวรี(query) นั้นมีคำตอบ โดยพบว่าค่าจากการประเมินในส่วนที่สองนั้นดีกว่าในส่วนแรก ซึ่งในความเป็นจริงแล้วคำถามที่ไม่มีการตอบกลับในส่วนแรกนั้นน่าจะมีเป็นจำนวนมาก ดังนั้นการเปรียบเทียบในส่วนนี้จึงไม่น่าเชื่อถือ เพราะฉะนั้นจึงเน้นไปที่การประเมินในส่วนที่สอง ซึ่งผลที่ได้จากลิงค์แกรม-มาพาสเซอร์นั้นมีค่าที่ดีกว่าคอนเซอร์เอฟดีจี

สรุปการเปรียบเทียบพาสเซอร์ทั้งสองชนิดนี้คอนเซอร์เอฟดีจีพาสเซอร์จะใช้เวลาในการทำงานน้อยกว่าแต่ลิงค์แกรมมมาพาสเซอร์จะให้ผลลัพธ์ที่ดีกว่า

(Diego Molla and Ben Hutchinson, 2003)

ในปี ค.ศ. 2003 ได้มีการนำหลักการของลิงค์แกรมมาพาสเซอร์มาใช้ในการคัดแยกข้อมูลที่เป็นปฏิกิริยาไบโอเคมีจากแหล่งข้อมูล MEDLINE ซึ่งเป็นแหล่งข้อมูลขนาดใหญ่ที่ใช้ในการจัดกลุ่ม(mining) ปฏิกิริยาไบโอเคมี วิธีการที่นำเสนอโดยใช้หลักการนี้จะสามารถจัดการกับประโยคที่มีความซับซ้อนมากขึ้นได้ เนื่องจากบทคัดย่อที่ปรากฏในแหล่งข้อมูลนี้จะเป็นประโยคที่มีความซับซ้อนมากกว่าประโยคธรรมดาทั่วไป ลิงค์แกรมมาจะทำการเชื่อมคำแต่ละคู่ในประโยคด้วยลิงค์ที่แสดงหน้าที่ของคำแต่ละคู่ที่แตกต่างกันออกไป งานวิจัยนี้จึงใช้ประโยชน์จากชนิดของลิงค์ที่ได้จากลิงค์แกรมมาพาสเซอร์มาใช้การคัดแยกข้อมูลที่เป็นปฏิกิริยาระหว่างสารไบโอเคมีแต่ละคู่ได้

เพื่อเพิ่มประสิทธิภาพและความถูกต้องในการทำงานจึงต้องมีกระบวนการก่อนการเริ่มประมวลผลด้วยลิงค์แกรมมาก่อนตามกฎดังต่อไปนี้

1. ทำการเปลี่ยนแปลงตัวอักษรที่เป็นชื่อของสารไบโอเคมีที่สนใจในประโยคให้เป็นอักษรตัวพิมพ์ใหญ่
2. ทำการเชื่อมคำผสมต่าง ๆ เข้าด้วยกันด้วยเครื่องหมายขีดเส้นใต้อักขระ (underscore)
3. ทำการแทนตัวอักษรพิเศษต่าง ๆ ที่ปรากฏอยู่ในชื่อของสารไบโอเคมีด้วยเครื่องหมายขีดเส้นใต้อักขระ
4. ลดความซับซ้อนของประโยคบางประโยคด้วยการส่งบางส่วนของประโยคเข้าไปทำการประมวลผลด้วยพาสเซอร์เมื่อส่วนนั้นของประโยคมีคู่ของสารเคมีเกิดขึ้นระหว่างเครื่องหมายวรรคตอน {, ; .} ถ้ามีลิงค์ที่สมบูรณ์เกิดขึ้น หรือเส้นทางของลิงค์ที่เชื่อมระหว่างสารสองตัวนั้นสามารถทำการคัดแยกออกมาได้ก็ให้ใช้ส่วนนั้นของประโยคได้
5. ทำการเปลี่ยนแปลงหัวข้อเล็กน้อยเพื่อให้ง่ายขึ้นดังนี้
   - ทำการแปลงคำที่เป็นตัวอักษรตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็ก ยกเว้นคำที่เป็นชื่อสารเคมี
   - ทำตามกฎข้อที่ 4 ถ้าสามารถนำไปปรับใช้ได้
   - สำหรับหัวข้อที่ประกอบไปด้วยสองประโยคก็ให้ทำการเปลี่ยนแปลงตามกฎที่กำหนด

เมื่อได้ผ่านกระบวนการข้างต้นเรียบร้อยแล้ว จึงมีการนำประโยคไปประมวลผลด้วยลิงค์แกรมมาพาสเซอร์ ซึ่งจะได้ผลลัพธ์เป็นลิงค์ที่แสดงความสัมพันธ์กันของคำแต่ละคำในประโยค จากนั้นจึงทำการคัดแยกข้อมูลปฏิกิริยาเคมีจากเส้นทางของลิงค์ที่พบระหว่างสารไบโอเคมีทั้งสอง

ชนิดนั้น สำหรับความผิดพลาดที่เกิดขึ้นในการคัดแยกข้อมูลนั้นอาจเกิดขึ้นจากการที่ไม่มีคำศัพท์คำนั้นปรากฏอยู่ในดิกชันนารีของลิงค์แกรมมา จึงมีผลต่อการระบุหน้าที่ของคำทำให้เกิดความผิดพลาดได้ นอกจากนี้โครงสร้างในบางประโยคก็อาจเป็นอุปสรรคในการพาสเซอร์ เช่น รูปแบบและเครื่องหมายที่ใช้ในการเปรียบเทียบ การแสดงความน่าจะเป็น และหน่วยวัดต่าง ๆ เป็นต้น รวมทั้งบางประโยคที่มีความยาวมากเกินไปทำให้ยากต่อการพาสเซอร์

(Jing Ding et al, 2003)

งานวิจัยนี้ได้มีการนำเสนอวิธีการในการคัดแยกข้อมูลที่เกี่ยวกับเหตุการณ์ที่เกิดขึ้น เป็นการทำงานกับภาษาธรรมชาติที่ไม่ต้องมีชุดข้อมูลที่ใช้ในการเรียนรู้ และใช้ประโยชน์จากโครงสร้างในการสร้างประโยคจากลิงค์ที่ได้จากลิงค์แกรมมาพาสเซอร์มาใช้ในการคัดแยกข้อมูลที่ต้องการ โดยงานนี้จะใช้ชนิดของลิงค์ที่ได้จากลิงค์แกรมมามาใช้ในการระบุคำกริยาหลักซึ่งเป็นส่วนที่จะแสดงการกระทำต่าง ๆ ในกิริยาวลี หลังจากที่คำกริยาหลักได้ถูกกำหนดแล้วจากนั้นก็จะสามารถทำการคาดเดาคำที่เป็นประธานและกรรมของประโยคได้ สามารถแสดงส่วนขยายของคำกิริยาและคำนามได้ ผลลัพธ์ที่ได้จากวิธีนี้ทั้งประธานและกรรมของประโยคที่คัดแยกได้จะเป็นแค่คำเท่านั้น

สำหรับแรงบันดาลใจของงานวิจัยนี้เริ่มต้นจากวิธีการของมนุษย์ที่ต้องการคัดแยกข้อความจากบทความ เมื่อมีความต้องการในการค้นหาเหตุการณ์ใดเหตุการณ์หนึ่ง สิ่งแรกที่ต้องทำคือการค้นหาคำสำคัญบางคำที่จะแสดงถึงเหตุการณ์ที่ต้องการได้ เมื่อพบแล้วจึงมุ่งจุดสนใจไปที่ประโยคนั้นแล้วจึงทำปฏิบัติตามกระบวนการต่อไป งานวิจัยที่นำเสนอนี้ก็ใช้แนวคิดดังกล่าวมาใช้แต่จะมุ่งสนใจไปที่การกระทำบางอย่างที่แสดงลักษณะพิเศษเท่านั้น ซึ่งเป็นเพียงแค่ส่วนหนึ่งของกรณีที่จะเกิดขึ้นจริงเท่านั้นและมีกระบวนการในการทำงานดังนี้

1. ผู้ใช้จะต้องป้อนคำกิริยาที่เป็นคำสำคัญที่คิดว่าเป็นคำที่ดีที่สุดที่สามารถแสดงการกระทำที่เป็นลักษณะพิเศษสำหรับเหตุการณ์ที่ต้องการ
2. ทำการตีความคำที่มีความหมายพ้องกันและคำที่อยู่ในระดับต่ำกว่า (hyponyms) ของคำสำคัญนั้น ซึ่งคำที่อยู่ในระดับต่ำกว่านั้นจะมีพื้นฐานคล้ายกันทางด้านของความหมายแต่ว่าเฉพาะเจาะจงมากกว่า
3. เลือกเอกสารที่ต้องการแล้วป้อนเข้าสู่ลิงค์แกรมมาพาสเซอร์เพื่อทำการระบุชนิดของคำ (part of speech) และโครงสร้างของประโยค คำแต่ละคำจะถูกแสดงชนิดของคำ เช่น คำนามแสดงด้วยตัวอักษร “n” คำกิริยาแสดงด้วยตัวอักษร “v” และคำบุพบทแสดงด้วยตัวอักษร “p” เป็นต้น

4. ทำการค้นหาเหตุการณ์ที่เกิดขึ้นทั้งหมดจากคำกิริยาจากขั้นตอนที่สอง และเลือกเฉพาะกรณีที่เกิดขึ้นจากคำกิริยาหลักเท่านั้น
5. พิจารณาประโยคทั้งหมดที่ได้จากขั้นตอนที่สี่ แล้วใช้กฎในการพิจารณาประธานและกรรมของประโยค

ส่วนที่สำคัญสำหรับงานวิจัยนี้ก็คือกฎต่าง ๆ ที่ใช้ในการทำนายคำที่มีหน้าที่เป็นประธานและกรรมของประโยค ในการระบุกิริยาหลักของประโยคนั้น ถึงแม้ว่าลิงค์พาสเซอร์จะมีการแสดงชนิดของคำที่เป็นกิริยาด้วยตัวอักษร “v” แล้วก็ตาม แต่ไม่ได้หมายความว่ากิริยาทุกตัวจะเป็นกิริยาหลักและจะต้องมีประธานเสมอไป คำที่จะถูกพิจารณาเป็นคำกิริยาหลักจะต้องเป็นคำที่อยู่ในกิริยาวลีที่แสดงการกระทำในประโยคนั้น นอกจากนี้คำที่เป็นคำกิริยาบางคำก็ทำหน้าที่เป็นเหมือนกับเป็นคำคุณศัพท์ดังนั้นจึงไม่มีประธาน การระบุคำกิริยาหลักจึงต้องเป็นไปตามกฎ

สำหรับการทดลองนั้นได้ทำการทดสอบกับบทความทางหนังสือพิมพ์ออนไลน์จากประเทศต่าง ๆ เพื่อให้มีรูปแบบการเขียนที่แตกต่างกันออกไป และทำการประเมินผลด้วยค่ารีคอลและค่าพรีซิสชันซึ่งค่าที่ได้ก็มีผลเป็นที่น่าพอใจ

(Harsha V.Madhyastha and N.Balakrishman, 2003)

ระบบการคัดแยกข้อมูลมีแนวโน้มที่จะเกิดความผิดพลาดได้จากหลายเหตุผลและจากที่ได้มีการศึกษาพบว่าความผิดพลาดจำนวนมากนั้นมักเกิดขึ้นในประโยคที่มีภาษาเป็นการแสดงความคิดเห็นส่วนตัว (Subjective language) เช่น แสดงความเห็น แสดงอารมณ์ และแสดงทัศนคติ ดังนั้นงานวิจัยนี้จึงได้นำเสนอแนวความคิดในการวิเคราะห์ภาษาที่มีลักษณะดังกล่าว เพื่อปรับปรุงค่าพรีซิสชันของระบบการคัดแยกข้อมูลให้ดีขึ้น โดยการใช้การจัดประเภทของประโยคที่มีการแสดงความคิดส่วนตัวดังกล่าวให้เป็นหมวดหมู่เพื่อกลั่นกรองการคัดแยกนั้น ในการทดลองจะใช้วิธีการในการจัดประเภทของประโยคหลากหลายวิธีแตกต่างกันออกไป โดยการทดลองกับคลังข้อมูล MUC-4 ซึ่งเป็นแหล่งข้อมูลที่เป็นบทความเกี่ยวกับขบวนการก่อการร้าย และพบว่าเมื่อมีการเลือกกระบวนการในการกลั่นกรองก่อนทำการคัดแยกนั้นจะทำให้ค่าพรีซิสชันที่ได้มีค่าดีขึ้น

งานวิจัยนี้จะทำการสร้างระบบการคัดแยกข้อมูลสำหรับคลังข้อมูล MUC-4 เป็นเรื่องราวเกี่ยวกับการก่อการร้าย และจะสามารถสร้างแพตเทิร์นที่ใช้ในการคัดแยกได้โดยนำวิธีการของ AutoSlog-TS มาใช้ ซึ่งวิธีดังกล่าวจะใช้ชุดข้อมูลในการเรียนรู้แบ่งเป็น 2 กลุ่มด้วยกันคือ ชุดข้อมูลที่มีเนื้อหาเกี่ยวข้องกับการก่อการร้าย และชุดข้อมูลที่ไม่เกี่ยวข้องกับการก่อการร้าย กระบวนการใน

การเรียนรู้จะประกอบไปด้วย 2 ขั้นตอนด้วยกัน ในขั้นตอนแรกนั้นแพตเทิร์นของโครงสร้างภาษาจะถูกสร้างขึ้นเพื่อนำไปใช้ในกระบวนการเรียนรู้ ตัวอย่างเช่น แพตเทิร์น “<subj> passive verb” ถูกสร้างขึ้นเพื่อคำกิริยาทุกตัวในคลังข้อมูลที่มีรูปแบบเป็น “passive voice” มีผลให้ประธานของคำกิริยานั้นถูกคัดแยก ซึ่งในเนื้อหาของการก่อการร้ายนั้นแพตเทิร์นที่ใช้ในการคัดแยกอาจจะเป็น “<subj> was killed” “<subj> was bombed” และ “<subj> was attacked” สำหรับขั้นตอนที่สองนั้นเมื่อมีการนำเอาแพตเทิร์นเหล่านั้นไปประยุกต์ใช้กับชุดข้อมูลในการเรียนรู้แล้วจะต้องมีการเก็บสถิติของแพตเทิร์นที่ได้ว่ามีความเกี่ยวข้องหรือไม่เกี่ยวข้องกับบทความ ซึ่งการจัดเรียงลำดับของแพตเทิร์นนั้นจะต้องมีพื้นฐานอยู่บนสัมพันธ์เกี่ยวข้องกับเนื้อหาที่ต้องการ โดยมนุษย์จะเป็นผู้ตัดสินใจเลือกแพตเทิร์นที่ตรงกับจุดประสงค์ในการคัดแยกข้อมูล ซึ่งในครั้งนี้มีจุดมุ่งหมายในการคัดแยกข้อมูลเกี่ยวกับผู้กระทำ ผู้ถูกกระทำ อาวุธและจุดประสงค์ในการก่อการร้าย และจะมีการเพิ่มแต่ละแพตเทิร์นที่เกี่ยวข้องเข้าไปหลังจากที่ได้มีการกำหนดจุดมุ่งหมายในการคัดแยกแล้วโดยทำการเลือกจากค่าทางสถิติที่เก็บไว้ เมื่อทำการทดลองแล้วได้ผลลัพธ์คือ ค่าพรีซิสชัน 46 เปอร์เซนต์ และค่ารีคอล 51 เปอร์เซ็นต์

จากแพตเทิร์นที่เลือกโดยมนุษย์น่าจะทำให้มีประสิทธิภาพในการคัดแยก แต่ความคิดและความเชื่อของมนุษย์ที่ใช้ในการเลือกแพตเทิร์นนั้นจะมีผลทำให้เกิดความไม่น่าเชื่อถือขึ้น เช่น แพตเทิร์น “was aimed at <np>” ซึ่งเป็นแพตเทิร์นที่สามารถแสดงได้หลายความหมาย จึงยากต่อการพิจารณาของมนุษย์ว่าควรใช้เป็นแพตเทิร์นในกาคัดแยกหรือไม่ ดังนั้นจึงได้มีการนำเทคนิคในการจัดกลุ่มเข้ามาใช้ แพตเทิร์นในการเรียนรู้จะถูกนับจำนวนครั้งที่แต่ละแพตเทิร์นปรากฏขึ้นในประโยคทั้งในประโยคที่เป็นความจริงและประโยคที่เป็นการแสดงความคิดเห็นส่วนตัว แล้วนำความถี่ที่เกิดขึ้นมาคำนวณความน่าจะเป็นที่แพตเทิร์นนั้นจะเป็นแพตเทิร์นของประโยคที่เป็นการแสดงความคิดเห็นส่วนตัว ถ้า $P(subj \mid pattern_i) > .50$ และค่าความถี่ $> 10^7$ แสดงว่าแพตเทิร์นนั้นเป็นแพตเทิร์นที่อยู่ในประโยคที่เป็นการแสดงความคิดเห็นส่วนตัว

เมื่อนำวิธีการในการจัดกลุ่มมาใช้ผลการทดลองที่ได้พบว่ามีค่าพรีซิสชันที่เพิ่มขึ้นจาก 46 เปอร์เซ็นต์ เป็น 56 เปอร์เซ็นต์ และค่ารีคอล 51 เปอร์เซ็นต์ ถึงแม้การทดลองจะยังไม่ครอบคลุมสำหรับทุกประโยคแต่เชื่อว่าน่าจะเป็นประโยชน์ต่อระบบการคัดแยกข้อมูลที่จะพัฒนาต่อไปในอนาคต

(Ellen Riloff et al, 2005)

จากเอกสารที่ได้ศึกษานั้นพบว่าลิงค์แกรมมาพาสเซอร์ได้ถูกนำมาใช้ในการคัดแยกข้อมูลอยู่พอสมควร โดยเป็นการนำเอาชนิดของลิงค์ที่ได้จากลิงค์แกรมมาพาสเซอร์มาใช้ในการคัดแยก ดังนั้นข้อมูลที่ได้จากการคัดแยกจึงได้ผลเป็นคำเท่านั้น ทำให้เกิดแนวความคิดในการคัดแยกข้อมูลจากภาษาธรรมชาติเพื่อให้ได้ข้อมูลที่มีความกระชับและมีความหมายมากขึ้น และพบว่าลิงค์แกรม-มาพาสเซอร์นั้นมีความสามารถในการวิเคราะห์โครงสร้างของประโยคโดยแบ่งเป็นวลีต่าง ๆ ได้ จึงได้นำลิงค์แกรมมาพาสเซอร์มาใช้ในการคัดแยกข้อมูลเพื่อให้ได้ข้อมูลที่มีความหมายมากขึ้น

ในบทต่อไปจะนำเสนอหลักการลิงค์แกรมมาพาสเซอร์ในการวิเคราะห์โครงสร้างของประโยคเพื่อใช้ในการคัดแยกข้อมูลที่ต้องการต่อไป