การตรวจการโจรกรรมทางวิชาการด้วยใช้เทคนิค N-gram ร่วมกับเทคนิคการตรวจสอบเชิงความหมายสำหรับเอกสารภาษาไทย

A Hybrid Approach for Thai Documents Plagiarism Detection using N-gram and Semantic Role Labeling Technique

*สรวัตร ประภานิติเสถียร[1], ไกรศักดิ์ เกษร**

*Sorawat Prapanitisatian, Kraisak Kesorn**

[1] ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand.

ABSTRACT - Plagiarism detection methods have been required by many education institutions because students in these institutions use other people works or idea without referencing the original authors. However, the existing methods obtain poor performance when applying them to Thai documents because the sentence structure of Thai is different from English (e.g. no space between words). This research presents a hybrid approach for Thai document plagiarism detection using the N-gram and Semantic Role Labeling (SRL) techniques. In addition, keywords of documents are weighted to enhance similarity computation efficiency. As a result, the proposed technique can improve the plagiarism detection performance compared to traditional SRL and other existing methods.
KEY WORDS -- Plagiarism detection, Semantic-based technique, Semantic Role Labeling, N-grams

บทคัดย่อ--การตรวจจับการคัดลอกเชิงวิชาการเป็นสิ่งที่ได้รับความสนใจอย่างมาก โดยเฉพาะในสถาบันการศึกษาเนื่องจากนักศึกษามักจะกระทำผิดโดยการนำเอาผลงานหรือแนวคิดผู้อื่นมาแอบอ้างเป็นงานของตนเอง แต่เทคนิคในการตรวจจับการคัดลอกของเอกสารที่นิยมใช้กันในปัจจุบันนั้น เมื่อนำมาใช้กับเอกสารภาษาไทยพบว่ามีประสิทธิภาพที่ต่ำเนื่องจากปัญหาด้านโครงสร้างไวยากรณ์ของภาษาไทย งานวิจัยนี้จึงนำเสนอการนำหลักไวยากรณ์และการตารางความน่าจะเป็นแบบ 5 และ 3 แกรม (N-gram) ที่สร้างจากตัวแบบทำนายต้นไม้ในการปรับปรุงโครงสร้างของประโยคและเทคนิค Semantic Role Labeling ร่วมกับการให้ค่าน้ำหนักของคำในการเปรียบเทียบเชิงความหมาย จากการทดลองพบว่าทำการปรับปรุงโครงสร้างของประโยคแล้วมีประสิทธิภาพในการตรวจจับมากยิ่งขึ้นกว่าการใช้เทคนิค Semantic Role Labeling ร่วมกับการให้ค่าน้ำหนักของคำเพียงอย่างเดียว
คำสำคัญ การโจรกรรมความคิดทางวิชาการ, การเปรียบเทียบเชิงความหมาย, Semantic Role Labeling, N-grams

1. บทนำ

การตรวจจับการคัดลอกทางวิชาการนั้นเป็นสิ่งที่สถาบันการศึกษาเริ่มให้ความสนใจมากขึ้นแต่การตรวจสอบการคัดลอกเป็นเรื่องยากและต้องใช้เวลามากในการตรวจสอบด้วยมือ มหาวิทยาลัยหลายแห่งมีการใช้โปรแกรมประยุกต์มาร่วมในการตรวจจับการคัดลอกทางวิชาการ เพื่อช่วยลดทรัพยาการบุคคลในการตรวจจับการคัดลอกจากเอกสารจำนวนมาก ซึ่งโปรแกรมประยุกต์ส่วนใหญ่นั้นถูกพัฒนาเพื่อใช้ในการตรวจจับการคัดลอกภาษาอังกฤษซึ่งมีการเว้นวรรคตอนระหว่างคำและมีการใช้เครื่องหมายเพื่อบอกถึงจุดสิ้นสุดของประโยคอย่างชัดเจน แต่เนื่องจากภาษาอังกฤษมีหลักไวยากรณ์ที่ต่างจากภาษาไทย ทำให้ซอฟต์แวร์ที่มีอยู่ในปัจจุบันมีประสิทธิภาพที่ลดลงเมื่อนำมาใช้กับเอกสารภาษาไทย

หัวใจสำคัญของการตรวจสอบการโจรกรรมทางวิชาการคือการใช้เทคนิคของการประมวลผลภาษาธรรมชาติ (Natural Language Processing) มาช่วยในการตรวจจับการคัดลอกเชิงวิชาการ โดยจะใช้ในการวิเคราะห์ประโยคเพื่อเปรียบเทียบความคล้ายคลึงกันของประโยค วิธีที่นิยมใช้ในการวิเคราะห์ประโยควิธีหนึ่งได้แก่ Semantic Role Labeling (SRL) เทคนิคนี้จะทำการแยกโครงสร้างของประโยคเป็นแต่ละหน้าที่ของคำในลักษณะโครงสร้างต้นไม้ด้วยหลักไวยากรณ์และทำการเชื่อมโยงคำศัพท์แต่ละคำกับคำคล้ายของคำศัพท์นั้นๆ เพื่อให้สามารถวิเคราะห์ประโยคได้ว่าประโยคนั้นมีใจความว่าอะไร,ถูกทำโดยใคร,ทำอะไรที่ไหน,เมื่อไร อย่างไรก็ตามผู้วิจัยได้พบปัญหาในขั้นตอนการวิเคราะห์ประโยคโดยใช้เทคนิค SRL กับประโยคภาษาไทยดังนี้

1) ปัญหาการไม่พบคำศัพท์ในฐานข้อมูล Lexitron ทำให้ไม่สามารถทราบได้ว่าคำศัพท์นั้นทำหน้าที่อะไรในประโยค จึงทำให้มีโอกาสที่จะจัดประโยคให้อยู่ในโครงสร้างของ SRL ผิดพลาดได้ได้เนื่องจาก SRL ใช้หน้าที่ของคำในพจนานุกรมเป็นหลักในการจัดโครงสร้าง

2) คำศัพท์หนึ่งคำทำหน้าที่หลายอย่างทำให้มีโอกาสที่จะถูกนำไปเปรียบเทียบกับคำศัพท์ที่อยู่ในกลุ่มหน้าที่ของคำผิดได้เช่นคำว่า ‘‘เริ่ม’สามารถทำหน้าที่ของ V และ AUX ได้ได้

3) การตัดประโยคผิดพลาดทำให้ค่าน้ำหนักของคำในขั้นตอนการเปรียบเทียบผิดซึ่งจะส่งผลให้การเปรียบเทียบความเหมือนคลาดเคลื่อน

ในงานวิจัยนี้จึงนำเสนอการนำเทคนิค n-gram และการเทคนิคการวิเคราะห์ประโยคตามหลักไวยากรณ์ เพื่อใช้ในการแก้ไขปัญหาของการไม่พบคำศัพท์ในพจนานุกรมและการเลือกหน้าที่ของคำในกรณีที่คำนั้นมีได้หลายหน้าที่ร่วมกับเทคนิค SRL

2. งานวิจัยที่เกี่ยวข้อง

จากปัญหาการคัดลอกหรือโจรกรรมความคิดหรือผลงานทางวิชาการดังกล่าว จึงได้มีการคิดค้นและพัฒนาวิธีตรวจสอบการคัดลอกผลงานวิจัยต่างๆ โดยใช้เทคนิคของการประมวลผลภาษาธรรมชาติขึ้นมา Schleimer et al.[1] ได้เสนอวิธีตรวจสอบการคัดลอกเอกสารทั้งเอกสาร โดยใช้ Winnowing algorithm [2] เพื่อช่วยให้สามารถตรวจสอบการคัดลอกเมื่อมีการสลับตำแหน่งของอักขระหรือคำในบทความได้ โดยนำเสนอแนวทางการแก้ไขปัญหาการคัดลอกบทความจากหลายๆ แหล่งมารวมกันโดยการตัดคำในเอกสารออกส่วนๆ เท่าๆ กันและทำการเข้ารหัสข้อมูลเหล่านั้นเป็นตัวเลข (document fingerprinting) ด้วยการใช้เทคนิค Winnowing algorithm และทำการเปรียบเทียบตัวเลขเหล่านั้นกับเอกสารทดสอบ ซึ่งผลที่ได้พบว่ามีความแม่นยำในการตรวจจับการคัดลอกเอกสารแบบคำต่อคำสูง อย่างไรก็ตามการทำงานของวิธีนี้มีข้อจำกัด เนื่องจากมีการทำงานบนพื้นฐานของค่าแฮช (Hash number) ดังนั้นจึงมีโอกาสที่ข้อความต่างกัน แต่คำนวณได้ค่าแฮชเท่ากัน ซึ่งจะทำให้เกิดความผิดพลาดขึ้นในการเปรียบเทียบเอกสาร

ต่อมา Du et.al et al.[3] ทำการปรับปรุงการคำนวณค่าแฮชเพื่อเพิ่มความแม่นยำในการตรวจสอบให้ดีขึ้นด้วยการปรับปรุงสมการในการคำนวณในการหาค่าแฮชเพื่อแก้ไขปัญหา กรณีที่จำนวนชุดของค่าแฮชมีค่าเท่ากัน อย่างไรก็ตามการตรวจสอบการคัดลอกวิธีนี้ยังมีพื้นฐานการตรวจสอบแบบเปรียบเทียบคำต่อคำ หมายความว่าหากมีการคัดลอกและผู้คัดลอกทำการสลับตำแหน่งของคำ วิธีการนี้อาจจะไม่สามารถตรวจจับการคัดลอกได้ ดังนั้น Gipp et al.[4] ได้นำเสนอการตรวจสอบการคัดลอกเอกสาร โดยนำเอกสารอ้างอิง (references) ร่วมพิจารณากับการใช้การเปรียบเทียบคำต่อคำ เพื่อให้สามารถขยายขอบเขตการตรวจจับการคัดลอกแบบเปลี่ยนคำได้ดียิ่งขึ้น โดยมีแนวคิดว่าการเอกสารใดๆ ที่มีการอ้างอิงเหมือนกันอาจจะมีการคัดลอกกันเกิดขึ้น อย่างไรก็ตามการนำเอกสารอ้างอิงมาใช้ในการหาความสัมพันธ์ของเอกสารนั้นมีข้อจำกัดคือหากในเอกสารไม่มีการระบุแหล่งอ้างอิงจะทำให้ประสิทธิภาพในการตรวจสอบลดลง จากวิธีการต่างๆ ที่กล่าวมาข้างต้นเป็นการพิจารณาการคัดลอกเอกสารโดยเปรียบคำต่างๆ ในเอกสาร แต่วิธีการเหล่านี้ไม่ได้พิจารณาความหมายของคำตามบริบทที่แท้จริงหรือเรียกว่า “การตรวจสอบเชิงความหมาย” ดังนั้นนักวิจัยชื่อ Osman et al.[5] จึงได้นำเสนอการนำวิธี Semantic role labeling (SRL) ซึ่งเป็นเทคนิคที่ใช้ในการประมวลผลภาษาธรรมชาติ โดยมีจุดเด่นที่สามารถวิเคราะห์ประโยคที่มีการใช้คำที่มีความหมายคล้ายกันได้ โดยนำเทคนิค SRL มาใช้ในการเปรียบเทียบความคล้ายคลึงของประโยคเชิงความหมาย แต่จำกัดขอบเขตของการหาคำที่มีความหมายเหมือนกันเฉพาะคำกริยา

เท่านั้น ทำให้สามารถตรวจสอบการคัดลอกเชิงความหมายในส่วนของคำกริยาได้แม่นยำยิ่งขึ้น แต่วิธีนี้มีข้อจำกัดในภาษาไทย เนื่องจากโครางสร้างของภาษาไทยมีความแตกต่างจากภาษาอังกฤษตามที่กล่าวในบทนำ จึงได้มีการนำเทคนิค N-gram มาร่วมในการตัดคำภาษาไทย โดยเทคนิคนี้จะใช้วิธีการคำนวนความน่าจะเป็นที่จะเกิดรูปแบบคำศัพท์จากจำนวนคำ N คำก่อนหน้าสำหรับภาษาไทยนั้นนิยมใช้ค่า N เป็น 3 และ 4 เช่น [6]ได้นำเสนอการใช้ N-gram มาช่วยในการตัดคำเพื่อการค้นคืนระบบสารสนเทศ จากผลการทดลอง พบว่าการทำ index โดยใช้ 4-grams ในการตัดคำมีความแม่นยำมากที่สุด หรือ [7] ใช้ 3-grams ในขั้นตอนการตัดคำของการย่อความให้เหลือแต่ใจความสำคัญเป็นต้น [8] ได้ทำการทดลองเปรียบเทียบการตัดคำกับภาษาที่ไม่มีจุดสิ้นสุดของประโยคเช่นภาษาไทย,จีน,ญี่ปุ่น พบว่าขนาดของ N-grams ส่งผลกับความถูกต้องโดยพบว่า 3 grams เหมาะสมกับการตัดคำไทยมากที่สุด[9] ได้ทำการทดลองโดยใช้ N-gram เพื่อวัดประสิทธิภาพของการตัดคำพบว่าควรใช้หลักไวยากรณ์ร่วมในการตัดคำด้วยเพื่อเพิ่มประสิทธิภาพในการตัดคำ

จากปัญหาดังกล่าวดังนั้นผู้ทำวิจัยจึงได้มีแนวคิดที่จะพัฒนาเทคนิคการคัดลอกเอกสารในลักษณะของการเปรียบเทียบเชิงความหมายสำหรับเอกสารภาษาไทยให้มีประสิทธิภาพมากยิ่งขึ้น โดยการปรับปรุงขั้นตอนของการวิเคราะห์ประโยคโดยมุ่งเน้นไปที่การใช้โมเดล N-gram ในการคาดเดาความน่าจะเป็นของหน้าที่คำศัพท์เมื่อไม่พบคำศัพท์ในพจนานุกรมหรือพบคำศัพท์ที่มีได้หลายหน้าที่ร่วมกับการใช้ไวยากรณ์โครงสร้างวลี (phase structure grammars)

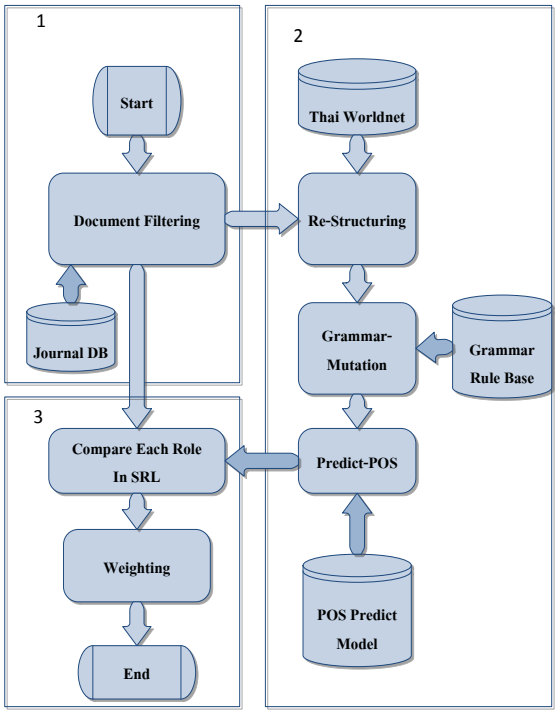
3. ระเบียบวิธีวิจัย

งานวิจัยนี้ใช้วิธี SRL ซึ่งประกอบด้วยขั้นตอน 3 ส่วน (รูปที่ 1) ดังนี้

1) การคัดกรองเอกสารและการเตรียมข้อมูล (Document filtering and data preparing) เป็นการคัดเลือกเอกสารที่ต้องการตรวจสอบซึ่งจะเรียกว่า “เอกสารกลุ่มเสี่ยง” โดยขั้นตอนนี้จะคัดเลือกจากการเปรียบเทียบชื่องานวิจัยที่ต้องการตรวจสอบกับคำสำคัญของงานวิจัยที่อยู่ในฐานข้อมูล และทำการตัดคำโดยใช้ Swath API[1] ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ซึ่งเป็นเครื่องมือในการตัดคำภาษาไทยที่มีประสิทธิภาพสูงโดยจะทำการตัดคำโดยเปรียบเทียบกับพจนานุกรม เมื่อผ่านกระบวนการตัดคำแล้ว จึงจะทำการตัดคำที่ไม่มีความสำคัญออก (Remove stop word)

2) การเปลี่ยนโครงสร้างประโยค (Sentence Re-structuring) คือการจัดประโยคให้อยู่ในโครงสร้างของ SRL ด้วยใช้ไวยากรณ์โครงสร้างวลี และการใช้โมเดลต้นไม้ทางเลือกที่จัดเก็บรูปแบบหน้าที่คำศัพท์แบบ 5-gramและ 3-gram ในการทำนายหน้าที่ของคำศัพท์ที่ไม่ปรากฏในพจนานุกรม Lexitron หรือคำศัพท์ที่มีหลายหน้าที่

3) การเปรียบเทียบความคล้ายคลึง (Similarity comparison) โดยจะทำการเปรียบเทียบความเหมือนด้วยการเปรียบเทียบคำศัพท์ตามแต่ละหน้าที่ของคำ เข้าด้วยกันดังรูปที่ 1



รูปที่ 1. แนวคิดของระบบ

3.1 แนวการคัดกรองเอกสารและการเตรียมข้อมูล (document filtering and data preparing)

ในการตรวจสอบการคัดลอกงานวิจัยนั้นขั้นตอนที่ใช้เวลามากที่สุดคือขั้นตอนการเปรียบเทียบเชิงความหมายซึ่งต้องเปรียบเทียบเอกสารกลุ่มเสี่ยงทีละคู่ ทำให้จำเป็นต้องมีการจำกัดขอบเขตของเอกสารกลุ่มเสี่ยง ผู้ทำวิจัยได้นำเสนอวิธีในการจัดกัดของเขตโดยการเปรียบเทียบชื่องานวิจัย (Title) ที่ต้องการตรวจสอบกับคำสำคัญ (Keyword) ของเอกสารกลุ่มเสี่ยงจากวิธีนี้สามารถลดเวลาในการตรวจสอบได้ เมื่อทำการคัดกรองเอกสารแล้วจึงทำการเตรียมข้อมูลงานวิจัยที่ต้องการตรวจสอบจะถูกตัดคำออกโดยใช้ Swath API ซึ่งเป็นเครื่องมือในการตัดคำที่มีความแม่นยำสูง โดยเปรียบเทียบกับคำศัพท์ในพจนานุกรม และทำการตัดคำที่ไม่สำคัญ

[1]http://www.thaisemantics.org/service/swath/index

ทิ้ง (stop word) โดยคำที่ถูกตัดออกไปได้แก่คำที่เป็นสัญลักษณ์ที่ไม่มีความหมายและคำที่มีความถี่การถูกใช้บ่อยเกินกว่า 80% เนื่องจากเป็นคำที่ไม่สามารถใช้เป็นตัวแทนเอกสารใดๆ ได้ เมื่อทำขั้นตอนนี้แล้วพบว่าการเปลี่ยนประโยคให้อยู่ในโครงสร้างของ SRL ใช้เวลาลดลง

3.2 การเปลี่ยนโครงสร้างประโยค (Sentence re-structuring)

การเปลี่ยนโครงสร้างของประโยคใหม่ให้อยู่ในโครงสร้างของ SRL เป็นขั้นตอนที่สำคัญและพบปัญหามากที่สุด โดยระบบจะทำการวิเคราะห์ประโยคตามโรงสร้างของไวยากรณ์วลี โดยใช้เทคนิคการวิเคราะห์ส่วนประชิด (Immediate Constituent) [10] ซึ่งจะจัดประโยคเป็นสองส่วนคือนามวลี (noun phrase) และกริยาวลี (verb phrase) โดยจะเริ่มค้นหาคำนามและคำกริยาที่อยู่ติดกันเพื่อแบ่งประโยคเป็นประโยคย่อย และทำแบบนี้ไปเรื่อยๆ หากประโยคย่อยใดมีสมาชิกเพียงคำเดียวก็จะถูกระบุหน้าที่ของคำทันที โดยผลลัพธ์ที่ได้จากขั้นตอนนี้คือประโยคที่อยู่ในโครงสร้าง SRL ดังตัวอย่างในรูปที่ 4 ในขั้นตอนนี้ผู้วิจัยพบว่าปัญหาเกิดขึ้นในการจัดโครงสร้าง 2 ประการคือ 1 การไม่พบคำศัพท์ในพจนานุกรมและคำศัพท์บางคำมีได้หลายความหมายดังตารางที่ 1 จะเห็นว่าคำว่า ” เริ่ม” มี 3 หน้าที่ของคำหากใช้เพียงการจัดคำแบบพจนานุกรมแล้วจะพบว่าประโยค “โดยเริ่มวิเคราะห์นามวลี” มีความน่าจะเป็นที่จะเกิดโครงสร้างของประโยคถึง 6 รูปแบบ ดังตารางที่ 3 ผู้วิจัยจึงได้นำหลักไวยากรณ์วลีที่ถูกเก็บไว้ในลักษณะของกฎดังตัวอย่างในรูปที่ 2 มาร่วมในการรวมหรือเปลี่ยนหน้าที่ของคำศัพท์ตามหลักไวยากรณ์โดยจะเรียกขั้นตอนย่อยนี้ว่า Grammar- Mutation ทำให้พบว่าคำว่าเริ่มจะทำหน้าที่เป็นกริยาร่วม (AUX) เมื่ออยู่หน้าคำกริยาอื่นทำให้สามารถลดรูปแบบโครงสร้างของประโยคเหลือเพียง 2 รูปแบบ คือ ID ที่ 2 และ 5 หลังจากนั้นจึงใช้ฐานข้อมูลความน่าจะเป็นของคำศัพท์ดังตัวอย่างในตารางที่ 3 โดยจะเลือกค้นหาคำที่ติดกัน 5 คำก่อนหากไม่พบจะทำการค้นหาจาก ค่าเฉลี่ยความน่าจะเป็นของของการเลื่อน 3-grams ดังรูปที่ 3 จากตัวอย่างพบว่าไม่พบทั้ง 2 รูปแบบในฐานข้อมูลความน่าจะเป็นของคำคำที่ติดกัน 5 คำจึงต้องค้นหาจาก ฐานข้อมูลความน่าจะเป็นของคำคำที่ติดกัน 3 คำ พบว่า ID ที่ 2 มีความน่าจะเป็นที่จะเกิดมากที่สุด

ตารางที่ 1. แสดงตารางคำศัพท์และหน้าที่ของคำ ของประโยค “โดยเริ่มวิเคราะห์นามวลี”

โดย	เริ่ม	วิเคราะห์	นาม	วลี
CONJ	ADJ	V	N	N
PREP	AUX			
	V			

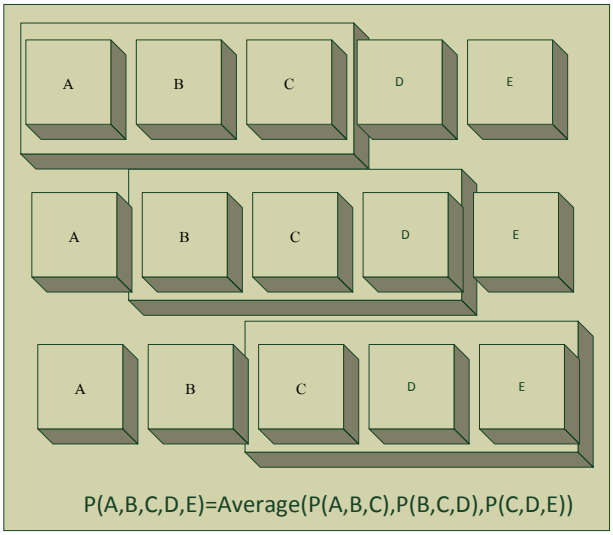
S = NP + VP

NP = Mod + N + PREP

Mod = ART + ADJ

VP = V + ADV

รูปที่ 2. แสดงตัวอย่างกฎของโครงสร้างไวยากรณ์วลี



รูปที่ 3. แสดงตัวอย่างการคำนวณความน่าจะเป็นของประโยค ABCDE จากการเลื่อน 3-gram 3ครั้ง

ตารางที่ 2. แสดงตารางคำศัพท์และหน้าที่ของคำ ของประโยค "โดยเริ่มวิเคราะห์นามวลี"

ID	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5
1	CONJ	ADJ	V	N	N
2	CONJ	AUX	V	N	N
3	CONJ	V	V	N	N
4	PREP	ADJ	V	N	N
5	PREP	AUX	V	N	N
6	PREP	V	V	N	N

ตารางที่ 3. แสดงตารางความน่าจะเป็นที่สร้างจาก microsoft decision trees

1st	2nd	3rd	4th	5th	Prob
N	DET	ADJ	ADV	endl	0.648828
N	DET	ADJ	N	space	0.648828
N	DET	ADV	N	space	0.648828
N	DET	AUX	ADJ	space	0.648828
N	DET	AUX	V	space	0.648828
N	DET	CLAS	V	space	0.648828
N	DET	N	ADJ	endl	0.648828

3.3 การเปรียบเทียบความคล้ายคลึง (similarity comparison) เป็นขั้นตอนที่ทำการเปรียบเทียบประโยคถูกจัดอยู่ในโครงสร้างของ SRL จากขั้นตอนที่โดยผลที่ได้จะเป็นรูปแบบประโยคดังรูปที่ 4 แล้วกับประโยคที่อยู่ในฐานข้อมูลซึ่งถูกจัดอยู่ในโครงสร้าง SRL อยู่แล้วโดยระบบจะจัดคำและคำที่มีความหมายคล้ายกันจากทั้งสองประโยคเข้าเป็นเซ็ตหน้าที่ของคำและทำการนับคำที่ซ้ำกันดังรูปที่ 5 ที่ทำการเปรียบเทียบประโยคในเซ็จหน้าที่คำนาม โดยจะมีการให้ค่าน้ำหนักแต่ละเซ็ตหน้าที่ของคำตามสมการที่ 1

$$\text{Weight}(\text{Role}_i)=\frac{C(\text{Args}_j)+C(\text{Args}_k)}{\sum_{i=0}^{k} C(\text{Args}_j)+\sum_{i=0}^{l} C(\text{Args}_k)} \quad (1)$$

โดยให้ $\text{Weight}(\text{Role}_i)$ คือค่าน้ำหนักของหน้าที่ของคำ i

$C(\text{Args}_j)$ คือ กลุ่มคำ j ในเซ็ตหน้าที่ของคำ i

$C(\text{Args}_k)$ คือ กลุ่มคำ k ในเซ็ตหน้าที่ของคำ i

$\sum_{i=0}^{k} C(\text{Args}_j)$ คือผลรวมของจำนวนคำทั้งหมดในประโยค j ซึ่งมีจำนวน k คำ

$\sum_{i=0}^{l} C(\text{Args}_k)$ คือผลรวมของจำนวนคำทั้งหมดในประโยค k ซึ่งมีจำนวน l คำ

จากสมการที่ 1 สามารถคำนวนค่าน้ำหนักของกลุ่มคำนาม ได้ดังนี้ (4+4)/(8+8)=0.5 และเมื่อนำไปคำนวณหาค่าน้ำหนักของคำของประโยคของประโยค "โดยเริ่มวิเคราะห์นามวลีและกริยาวลี"และ "โดยริเริ่มวิเคราะห์หากลุ่มคำกริยาและกลุ่มคำนาม" *ได้ผลออกมาดังตารางที่ 4*

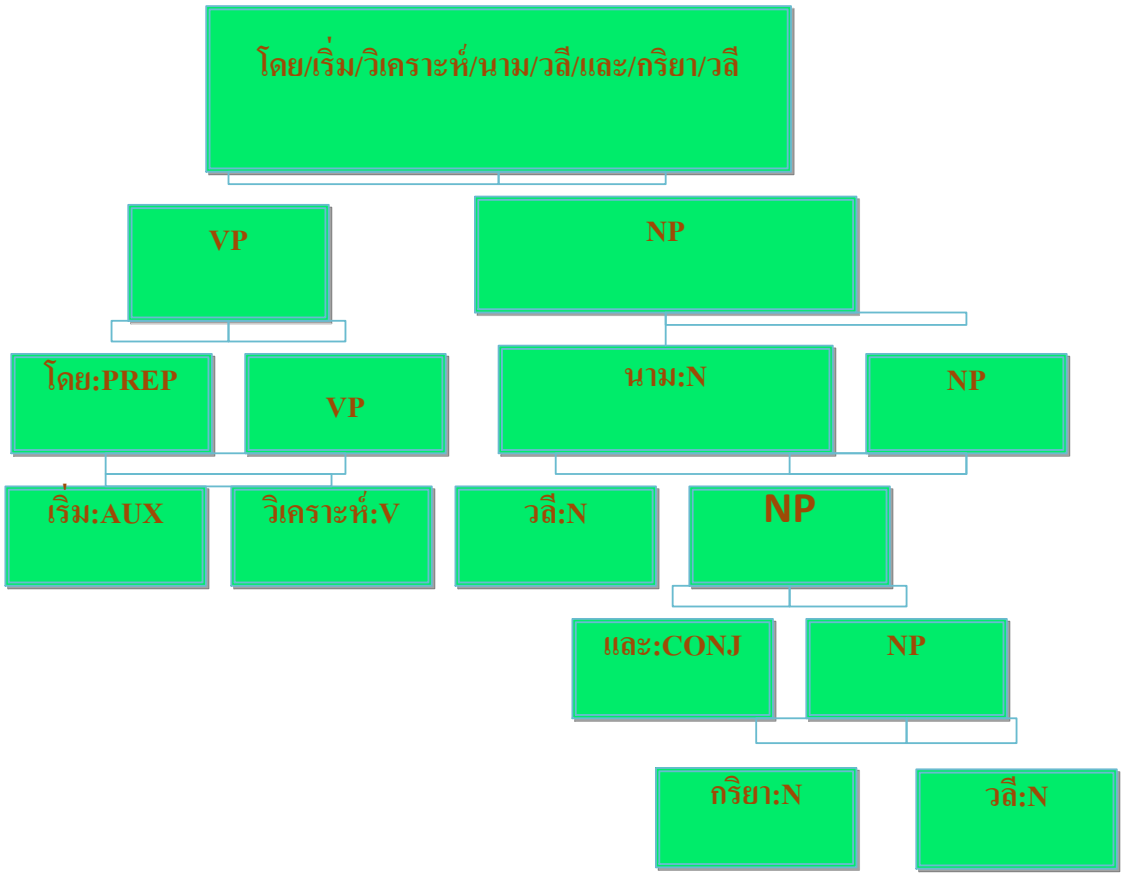
ซึ่งพบว่าค่าน้ำหนักของคำจะมีค่าแปลผันตรงกับจำนวนคำในแต่ละหน้าที่นั้นๆ เมื่อได้ค่าน้ำหนักของคำแล้วจะได้ค่าสมการความเหมือนดังสมการที่ 2

$$\text{similarity}\left(\text{Role}_i\right)=\text{Weight}\left(\text{Role}_i\right)\times\frac{C(\text{Args}_j)\cap C(\text{Args}_k)}{C(\text{Args}_j)\cup C(\text{Args}_k)} \quad (2)$$

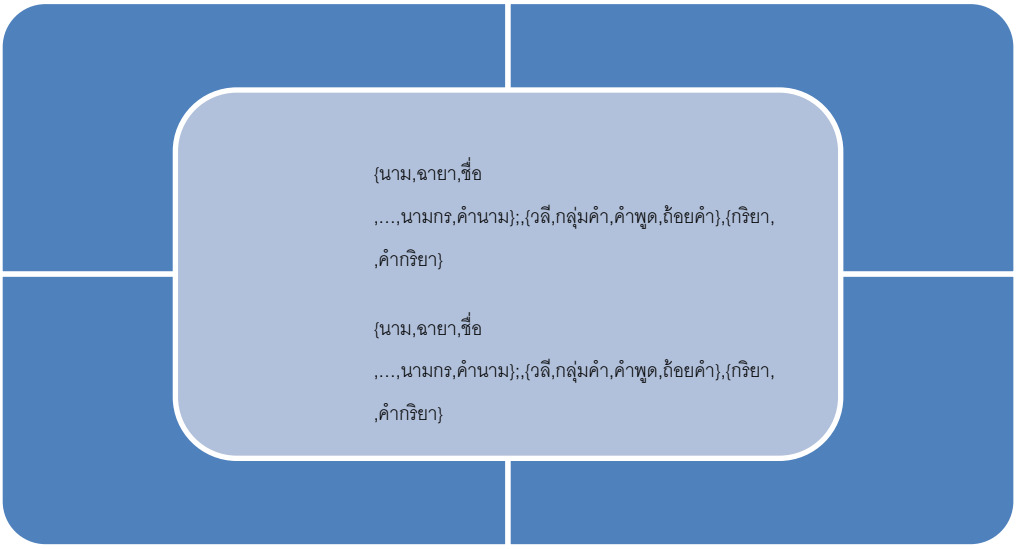
เมื่อนำสมการที่ (2) ไปใช้ในการเปรียบเทียบกับตัวอย่าง ได้ผลลัพธ์ดังตารางที่ 5

จากสมการที่ 2 จะเห็นว่าค่าความเหมือนของแต่ละหน้าที่ของคำนั้นได้ผ่านการให้ค่าน้ำหนักมาแล้วทำให้สามารถหาค่าความเหมือนของประโยคได้จากสมการที่ 3

$$\text{similarity}\left(S_j,S_k\right)=\sum_{i=1}^{n}\text{similarity}(\text{Role}_i) \quad (3)$$



รูปที่ 4. โครงสร้าง SRL ของประโยค "โดยเริ่มวิเคราะห์นามวลี"



รูปที 5. จำลองการเปรียบความคล้ายของหน้าที่ของคำ N(คำนาม) ของประโยค "โดยเริ่มวิเคราะห์นามวลีและกริยาวลี"และ"โดยริเริ่มวิเคราะห์หากลุ่มคำกริยาและกลุ่มคำนาม"

เมื่อนำสมการที่ 3 ไปใช้ในการคำนวณค่าความเหมือนของผระโยค “โดยเริ่มวิเคราะห์นามวลีและกริยาวลี”และ“โดยริเริ่มวิเคราะห์หากลุ่มคำกริยาและกลุ่มคำนาม” ที่ถูกคัดลอกโดยใช้เทคนิคการสลับคำ, การใส่คำขยาย, และการเปลี่ยนไปใช้คำคล้ายได้ค่าความเหมือนเป็น 0.99 ซึ่งสูงกว่าค่าความเหมือนที่ได้จากการเปรียบเทียบแบบคำต่อคำตำแหน่งต่อตำแหน่งซึ่งจะได้ค่าความเหมือนเป็น 2/17=0.12 (เหมือนแค่คำว่าโดยและคำว่าวิเคราะห์)จึงทำให้สามารถแก้ไขปัญหาของจากการเปรียบเทียบประโยคสองประโยคที่มีจำนวนคำแตกต่างกันหรือปัญหาการคัดลอกแบบใส่คำขยายได้เพราะค่าน้ำหนักของคำที่ถูกใส่คำขยายจะมีค่าน้อยทำให้ไม่ส่งผลกับค่าความเหมือนของประโยคได้

ตารางที่ 4 ตัวอย่างตารางค่าน้ำหนักของแต่ละหน้าที่ของคำที่ได้

หน้าที่ของคำ	ค่าน้ำหนัก
N	0.5
V	0.125
PREP	0.125
AUX	0.18
CONJ	0.125

3. การทดลองและอภิปรายผล

ผู้วิจัยได้สร้างแบบจำลองการทำนายผลที่จัดเก็บความน่าจะเป็นของหน้าที่ของคำแบบ 3-5 grams โดยใช้ฐานข้อมูลงานวิจัยจำนวน 11636 ประโยค ด้วยเทคนิคเทคนิคต้นไม้ทางเลือก (*Microsoft decision trees*) ดังตารางที่ 3และทำการเปรียบเทียบค่าความแม่นยำได้ดังตารางที่ 5

ตารางที่ 5. แสดงตารางค่าความแม่นยำของ N-grams ที่สร้างจาก Microsoft decision trees

N-grams	scores
3-grams	0.72
4-grams	0.69
5-grams	0.74

ผู้วิจัยได้เลือกใช้แบบจำลอง 5-grams และ 3-grams แบบเลื่อน 3 ครั้งในขั้นตอนการเปลี่ยนโครงสร้างประโยคจากนั้นจึงทำแบบจำลองข้อมูลทดสอบโดยคัดลอกประโยคของงานวิจัยที่ถูกเก็บในระบบตามประเภทการคัดลอกทั้ง 4 รูปแบบ รูปแบบละ 100 ประโยค เมื่อได้ข้อมูลทดสอบแล้วจึงนำมาทดสอบประสิทธิภาพของงานวิจัยโดยใช้วิธีการให้น้ำหนักของคำหรือเรียกว่า Weighted-SRL (W-SRL) โดยมีวิธีการเปรียบเทียบจากวิธีการเดิมที่มีอยู่ 3 วิธีคือ Anti-kobpae, Turnit-in[2] และ Traditional SRL (T-SRL) โดยกำหนดให้เอกสารที่ถูกคัดลอกจะต้องมีค่าความเหมือนมากกว่าหรือเท่ากับ 0.6 และใช้ Precision และ Recall เป็นมาตรวัดความถูกต้อง โดยมีสมมติฐานว่า Swath API สามารถตัดคำได้ถูกต้องในระดับที่ยอมรับได้และได้ผลการทดลองดังตารางที่ 6

จากตารางที่ 5 พบว่าทุกเทคนิคสามารถตรวจจับการคัดลอกแบบคำต่อคำ (Word by word) มีค่า Precision และ Recall เป็น 1.0 เหมือนกันหมด สำหรับการคัดลอกแบบสลับตำแหน่ง (Word reordering) พบว่า เครื่องมือ Anti-kobpae ไม่สามารถตรวจสอบการคัดลอกประเภทนี้ได้และ Turnit-in มีค่า Precision และ Recall ที่ลดต่ำลงในขณะที่เทคนิค T-SRL และ W-SRL ยังสามารถมีค่า Precision และ Recall ที่ 1.0 เช่นเดิม เนื่องจากใช้การเปรียบเทียบคำศัพท์ตามหมวดหมู่ของคำ ดังนั้นเมื่อมีการสลับตำแหน่งของคำจึงไม่ส่งผลกับค่าความเหมือน สำหรับการคัดลอกแบบเปลี่ยนไปใช้คำคล้าย (Synonym replacement) พบว่า Anti-kobpae และ Turnit-in มีค่า Precision และ Recall เท่ากับ 0 เนื่องจากไม่สามารถตรวจสอบการคัดลอกโดยใช้คำคล้ายได้ ในขณะที่ W-SRL มีค่า มีค่า Precision และ Recall เท่ากับ 0.69 และ 0.83 ซึ่งสูงกว่า T-SRL เนื่องจาก W-SRL ใช้เทคนิคการแบ่งกลุ่มในการทำนายหน้าที่ของคำศัพท์ อีกทั้งยังมีการใช้โมเดลในการทำนายผลในกรณีที่ไม่พบคำศัพท์ในฐานข้อมูลทำให้มีค่า Precision และ Recall ที่สูงกว่า อีกทั้งเมื่อมีการปรับปรุงรูปแบบของโมเดลการทำนายผลเป็น 5-grams และใช้หลักไวยากรณ์เข้าร่วมด้วย จะพบว่าสามารถเพิ่มค่า Precision และ Recall เท่ากับ 0.74 และ 0.86 ซึ่งสูงกว่าเทคนิค W-SRL เดิม สำหรับ การคัดลอกแบบใส่คำขยายก็เช่นเดียวกันพบว่า Anti-kobpae และ Turnit-in มีค่า Precision และ Recall เท่ากับ 0 ในขณะที่ W-SRL มีค่า Precision และ Recall เท่ากับ 0.74 และ 0.85 ซึ่งสูงกว่า T-SRL เนื่องจาก W-SRL ได้มีการนำค่าน้ำหนักมาช่วยในการคำนวณค่าความเหมือนทำให้สามารถป้องกันการคัดลอกโดยใส่คำ

[2] http://www.turnitin.com/

ขยายได้ดีมากขึ้น และพบว่า 5-grams W-SRL มีค่า Precision และ Recall เพิ่มขึ้นเป็น 0.79 และ 0.91 เนื่องจากสามารถทำนายผลหน้าที่ของคำได้แม่นยำขึ้นด้วยเช่นเดียวกัน

4. บทสรุป

งานวิจัยนี้เสนอเทคนิคในการตรวจสอบการโจรกรรมทางวิชาการโดยใช้เทคนิคผสมระหว่างการตรวจสอบเชิงความหมายและ N-gram โดยวิธีการที่นำเสนอสามารถแก้ไขปัญหาการโจรกรรมทางวิชาการทั้ง 4 รูปแบบได้แก่ การคัดลอกแบบคำต่อคำ การสลับตำแหน่งของคำ การใช้คำคล้าย และการคัดลอกแบบใส่คำขยาย ซึ่งซอฟต์แวร์ที่มีอยู่ในขณะนี้ยังมีประสิทธิภาพต่ำในการตรวจสอบการคัดลอกใน 2 แบบสุดท้าย อีกทั้งวิธีการให้น้ำหนักของคำที่นำเสนอ ยังสามารถเพิ่มประสิทธิภาพการตรวจจับการคัดลอกเนื่องจากเมื่อมีการคัดลอกโดยใส่คำขยายจะทำให้จำนวนคำศัพท์ในเซ็ตหน้าที่ของคำมากยิ่งขึ้นส่งผลให่ค่าน้ำหนักของเซ็ตหน้าที่ของคำมีค่าลดลง ทำให้การใส่คำขยายส่งผลกับค่าความเหมือนของประโยคน้อยลงและการพัฒนาโมเดลการทำนายผลเมื่อไม่พบคำศัพท์และการใช้หลักไวยากรณ์มาร่วมในการแปลงโครงสร้างของประโยคทำให้การจัดคำเข้าเซ็ตหน้าที่ของคำเพื่อเปรียบเทียบความคล้ายมีความถูกต้องมากยิ่งขึ้นส่งผลให้ประสิทธิภาพในการเปรียบเทียบดีขึ้นกว่าวิธีดั้งเดิม

ในขณะนี้ผู้วิจัยได้ทำการศึกษาหลักไวยากรณ์ภาษาไทยเพื่อหาวิธีที่ในการแยกประโยคออกจากกันโดยมีแนวคิดที่จะนำหลักไวยากรณ์ในการแบ่งประโยคมาใช้ร่วมกับโมเดลทำนายผล เพื่อให้สามารถแยกประโยคความซ้อนออกจากประโยคหลักและนำปรับปรุงค่าน้ำหนักตามระดับความสำคัญของประโยคและระดับความใกล้ชิดของคำคล้ายเพื่อพัฒนาวิธีการเปรียบเทียบความคล้ายเชิงความหมายของภาษาไทยในอนาคต

ตารางที่ 6 ผลการทดลอง

PG Tool	Word by word		Word reordering		Synonym replacement		Adjective insertion	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Anti-kobpae	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Turnit-in	1.0	1.0	0.8	0.73	0.0	0.0	0.0	0.0
T-SRL	1.0	1.0	1.0	1.0	0.64	0.75	0. 70	0.61
W-SRL	1.0	1.0	1.0	1.0	0.69	0.83	0.74	0.85
5 grams W-SRL	1.0	1.0	1.0	1.0	0.74	0.86	0.79	0.91

เอกสารอ้างอิง

[1] S. Schleimer, et al., "Winnowing: local algorithms for document fingerprinting," in Proceedings of the 2003 ACM SIGMOD international conference on Management of data, ed. New York, NY, USA: ACM, 2003, pp. 76–85

[2] L. Nick, "Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm," vol. 1988, pp. 285–318(2).

[3] Z. Du, et al., "A Cluster-Based Plagiarism Detection Method " in Conference and Labs of the Evaluation Forum, 2010.

[4] B. Gipp and N. Meuschke, "Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence," in Proceedings of the 11th ACM symposium on Document engineering, ed. New York, NY, USA: ACM, 2011, pp. 249–258

[5] A. H. Osman, et al., "An improved plagiarism detection scheme based on semantic role labeling," Appl. Soft Comput., vol. 12, pp. 1493–1502 2012.

[6] ศ. น. สิทธิโชค ปัญญาฤกษ์ชัย, "Information Retrieval System Using N-Gram Technique," presented at the The 5th National Conference on Computing and Information Technology, 2009.

[7] ช. จ. อัษฎางค์ แตงไทย "การย่อความเอกสารภาษาไทยโดยกรรมวิธีการแยกค่าแบบเดี่ยว," presented at the National Computer Science and Engineering Conference, 2004.

[8] อ. เอกวงศ์อนันต์, "การระบุคำไทยและคำทับศัพท์ด้วยแบบจำลองเอ็นแกรม," วิทยานิพนธ์มหาบัณฑิต ภาควิชาภาษาศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย, 2548

[9] T. Chumwatana, "Using N-gram and Frequent Max Substring Techniques for Index-Term Extraction from Non-Segmented Texts: A Comparison of Two Techniques," Journal of Information Science and Technology, vol. 3, pp. 8-15, JANUARY - JUNE 2012 2012.

[10]R. Pankhuenkhat, "การวิเคราะห์ประโยคภาษาไทย (Immediate Constituents)," สารภาษาไทยและวัฒนธรรมไทย vol. June - November 2007, 2007.