KKU Engineering Journal

http://www.en.kku.ac.th/enjournal/th/

โปรแกรมสนับสนุนการตรวจการโจรกรรมทางวิชาการด้วยใช้เทคนิค เชิงความหมายสำหรับเอกสารภาษาไทย

Semantic-based technique for thai documents plagiarism detection

สรวัตร ประภานิติเสถียร* ไกรศักดิ์ เกษร

Sorawat Prapanitisatian and Kraisak Kesorn

ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

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

Received March 2013

Accepted August 2013

บทคัดย่อ

การโจรกรรมทางวิชาการเป็นปัญหารุนแรงในวงการการศึกษา ในปัจจุบันมีเครื่องมือตรวจสอบการคัดลอกความคิดทางวิชาการอยู่มากมายบนอินเทอร์เน็ต อย่างไรก็ตามเครื่องมือที่มีอยู่เหล่านั้นยังมีประสิทธิภาพต่ำ เช่น เทคนิคที่ใช้การเปรียบเทียบอักขระ ซึ่งเทคนิคเหล่านั้นไม่สามารถตรวจสอบคำที่มีความหมายเหมือนกัน หรือการสลับตำแหน่งของคำได้ นอกจากนี้เครื่องมือเหล่านี้ไม่รองรับกับการตรวจสอบเชิงความหมาย โดยเฉพาะอย่างยิ่งสำหรับการตรวจสอบเอกสารภาษาไทย ดังนั้นงานวิจัยนี้จึงได้นำเสนอแนวคิดในการตรวจจับการโจรกรรมความคิดทางวิชาการโดยใช้เทคนิคการตรวจสอบเชิงความหมายโดยใช้เทคนิค Semantic role labeling และให้ค่าน้ำหนักคำ จากผลการทดลองพบว่า สามารถตรวจจับการโจรกรรมทางวิชาการได้มีประสิทธิภาพมากยิ่งขึ้น ถึงแม้ว่าจะมีการสลับคำในประโยคและสามารถทำตรวจสอบการคัดลอกเชิงความหมายได้มีประสิทธิภาพมากกว่าเทคนิคที่มีอยู่ เช่น Anti-kobpae, Turnit-in และ Traditional semantic role labeling

คำสำคัญ : การโจรกรรมความคิดทางวิชาการ การเปรียบเทียบเชิงความหมาย การให้ค่าน้ำหนักคำ

Abstract

Plagiarism is the act of taking another person's writing or idea without referring to the source of information. This is one of major problems in educational institutes. There is a number of plagiarism detection software available on the Internet. However, a few numbers of them works. Typically, they use a simple method for plagiarism detection e.g. string matching. The main weakness of this method is it cannot detect the plagiarism when the author replaces some words using synonyms. As such, this paper presents a new technique for a semantic-based plagiarism detection using Semantic Role Labeling (SRL) and term weighting. SRL is deployed in order to calculate the semantic-based similarity. The main different from the existing framework is terms in a sentence are weighted dynamically depending on their roles in the sentence

* Corresponding author. Tel.:
Email address: sorawat.etc@gmail.com

e.g. subject, verb or object. This technique enhances the plagiarism detection mechanism more efficiently than existing system although positions of terms in a sentence are reordered. The experimental results show that the proposed method can detect the plagiarism document more effective than the existing methods, Anti-kobpae, Turnit-in and Traditional Semantic Role Labeling.

Keywords: Plagiarism detection, Semantic role labeling, Semantic checking, Term weighting

1. บทนำ

การคัดลอกความคิดทางวิชาการกำลังเป็นปัญหาที่พบได้มากขึ้น เนื่องจากการเข้าถึงข้อมูลทำได้สะดวกทำให้การคัดลอกความคิดทำได้สะดวกขึ้นไปตามๆ กัน โดยการคัดลอกความคิดทางวิชาการนั้นพบได้มากในการศึกษา แม้ว่าการโจรกรรมทางวิชาการจะไม่ผิดกฎหมายแต่ถือเป็นความผิดทางจริยธรรมที่แอบอ้างเอาแนวคิดของผู้อื่นมาเป็นของตัวเองและอาจนำไปสู่ความล้มเหลวทางการศึกษาได้ ด้วยสาเหตุนี้สถาบันการศึกษาต่างๆ จึงให้ความสำคัญกับการตรวจสอบการคัดลอกความคิดทางวิชาการมากขึ้น โดยการคัดลอกความคิดทางวิชาการแบ่งออกเป็นสี่รูปแบบใหญ่ๆ คือ 1. การคัดลอกแบบคำต่อคำ (Word by word) 2. การคัดลอกโดยการสลับตำแหน่งของประโยคหรือคำในประโยค (Word reordering) 3. การเปลี่ยนไปใช้คำที่มีความหมายคล้ายกัน (Synonym replacement) 4. การคัดลอกโดยใส่คำขยาย (Adjective insertion)

ในอดีตนั้นการตรวจสอบการคัดลอกความคิดทางวิชาการจะใช้เจ้าหน้าที่เป็นผู้ตรวจสอบด้วยการอ่านเปรียบเทียบ ซึ่งทำได้ยากอีกทั้งยังเป็นการสิ้นเปลืองเวลาและทรัพยากรเป็นอย่างมาก ดังนั้นนักวิจัยจึงพัฒนาวิธีการตรวจสอบการโจรกรรมทางวิชาการขึ้น โดยใช้ระบบคอมพิวเตอร์มาช่วยในการเปรียบเทียบความคล้ายกันของงานวิจัย ซึ่งการตรวจสอบนั้นจะใช้การเปรียบเทียบสายอักขระ คือการตรวจสอบการคัดลอกแบบคำต่อคำและแบบสลับตำแหน่งของประโยค โดยมุ่งเน้นในเรื่องความแม่นยำและความเร็วในการตรวจสอบกับฐานข้อมูลงานวิจัยขนาดใหญ่เช่น ระบบตรวจสอบความคล้ายกันของเอกสาร (Anti-kobpae) ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) แต่การเปรียบเทียบสายอักขระจะไม่สามารถตรวจสอบการคัดลอก เมื่อมีการเปลี่ยนไปใช้คำที่มีความหมายเหมือนกัน ทำให้ความแม่นยำในการตรวจสอบลดลง ผู้วิจัยจึงได้มีแนวคิดที่จะพัฒนาเทคนิคการตรวจสอบการคัดลอกความคิดทางวิชาการเชิงความหมายเพื่อเพิ่มประสิทธิภาพในการตรวจจับการโจรกรรมทางวิชาการได้ดียิ่งขึ้น

2. งานวิจัยที่เกี่ยวข้อง

ปัจจุบันการเข้าถึงแหล่งข้อมูลงานวิจัยของนักวิจัยทำได้อย่างสะดวกมากเนื่องจากมีระบบเครือข่ายอินเทอร์เน็ตที่ทั่วถึง ทำให้เกิดปัญหาการคัดลอกงานวิจัยของขึ้นตามมา จึงได้มีการคิดค้นวิธีตรวจสอบการคัดลอกผลงานวิจัยต่างๆ ขึ้นมา Schleimer et al. [1] ได้เสนอวิธีตรวจสอบการคัดลอกเอกสารทั้งเอกสาร โดยใช้ Winnowing algorithm เพื่อช่วยให้สามารถตรวจสอบการคัดลอกแบบสลับตำแหน่งของอักขระ โดยนำเสนอแนวทางการแก้ไขปัญหาการคัดลอกบทความจากหลายๆ แหล่งมารวมกันโดยการตัดคำในเอกสารออกส่วนๆ เท่าๆ กันและทำการเข้ารหัสข้อมูลเหล่านั้นเป็นตัวเลข (Document fingerprinting) ด้วยการใช้เทคนิค Winnowing algorithm และทำการเปรียบเทียบตัวเลขเหล่านั้นกับเอกสารทดสอบ ซึ่งผลที่ได้พบว่ามีความแม่นยำในการตรวจจับการคัดลอกเอกสารแบบคำต่อคำสูง อย่างไรก็ตามการทำงานของวิธีนี้อาจมีข้อผิดพลาดได้เนื่องจากมีการทำงานบนพื้นฐานของ Hash number ดังนั้นจึงมีโอกาสที่ข้อความต่างกัน แต่คำนวณค่า Hash number ได้ค่าตัวเลขเดียวกัน ซึ่งจะทำให้เกิดความผิดพลาดขึ้นในการเปรียบเทียบเอกสารได้

ต่อมา Du et.al et al.[2] ทำการปรับปรุงสมการ (hash function) เพื่อเพิ่มความแม่นยำในการตรวจสอบให้ดีขึ้นด้วยการในกรณีที่จำนวนชุดของ hash number ไม่เท่ากัน อย่างไรก็ตามการตรวจสอบการคัดลอกวิธีนี้ยัง

มีพื้นฐานการตรวจสอบแบบเปรียบเทียบคำต่อคำ หมายความว่า หากมีการคัดลอกและผู้คัดลอกทำการสลับตำแหน่งของคำ วิธีการนี้อาจจะไม่สามารถตรวจจับการคัดลอกได้ Gipp et al.[3] ได้นำเสนอการตรวจสอบการคัดลอกเอกสารจากเอกสารอ้างอิง (References) ร่วมกับการใช้การเปรียบเทียบคำต่อคำ เพื่อให้สามารถขยายขอบเขตการตรวจจับการคัดลอกแบบเปลี่ยนคำได้ดียิ่งขึ้น โดยมีแนวคิดว่าการเอกสารใดๆ ที่มีการอ้างอิงเหมือนกันอาจจะมีการคัดลอกกันเกิดขึ้น อย่างไรก็ตาม การนำเอกสารอ้างอิงมาใช้ในการหาความสัมพันธ์ของเอกสารนั้นมีข้อจำกัดคือหากไม่มีการระบุแหล่งอ้างอิงจะทำให้ประสิทธิภาพในการตรวจสอบลดลง จากวิธีการต่างๆ ที่กล่าวมาข้างต้นเป็นการพิจารณาการคัดลอกเอกสารโดยเปรียบคำต่างๆ ในเอกสาร แต่วิธีการเหล่านี้ไม่ได้พิจารณาความหมายของคำตามบริบทที่แท้จริงหรือเรียกว่า “การตรวจสอบเชิงความหมาย” ดังนั้นนักวิจัยชื่อ Osman et al.[4] จึงได้นำเสนอการนำ Semantic role labeling (SRL) ซึ่งเป็นเทคนิคที่ใช้ในการประมวลผลภาษาธรรมชาติ โดยมีจุดเด่นที่สามารถวิเคราะห์ประโยคที่มีการใช้คำที่มีความหมายคล้ายกันได้ โดยนำเทคนิค SRL มาใช้ในการเปรียบเทียบความคล้ายคลึงของประโยคเชิงความหมาย แต่จำกัดขอบเขตของการหาคำที่มีความหมายเหมือนกันเฉพาะคำกริยาเท่านั้น ทำให้สามารถตรวจสอบการคัดลอกเชิงความหมายในส่วนของคำกริยาได้แม่นยำยิ่งขึ้น แต่วิธีนี้มีข้อจำกัดคือไม่สามารถตรวจสอบการคัดลอกได้หากในประโยคมีคำที่ไม่ทราบหน้าที่ของคำ

จากปัญหาดังกล่าวดังนั้นผู้ทำวิจัยจึงได้มีแนวคิดที่จะพัฒนาเทคนิคการคัดลอกเอกสารในลักษณะของการเปรียบเทียบเชิงความหมายสำหรับเอกสารภาษาไทยให้มีประสิทธิภาพมากยิ่งขึ้น

3. ระเบียบวิธีวิจัย

งานวิจัยนี้ใช้วิธี SRL ซึ่งประกอบด้วยขั้นตอน 3 ส่วนหลักๆ ดังนี้ ส่วนแรกคือ การคัดกรองเอกสารและการเตรียมข้อมูล (Document filtering and data preparing) เป็นการคัดเลือกงานวิจัยกลุ่มเสี่ยง โดยขั้นตอนนี้จะคัดเลือกจากการเปรียบเทียบชื่องานวิจัยที่ต้องการตรวจสอบกับคำสำคัญของงานวิจัยที่อยู่ในฐานข้อมูล และทำการตัดคำโดยใช้ Swath API ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ซึ่งเป็นเครื่องมือในการตัดคำภาษาไทยที่มีประสิทธิภาพสูงโดยจะทำการตัดคำโดยเปรียบเทียบกับพจนานุกรม เมื่อผ่านกระบวนการตัดคำแล้ว จึงจะทำการตัดคำที่ไม่มีความสำคัญออก (Remove stop word) ส่วนที่สองการเปลี่ยนโครงสร้างประโยค (Re-structuring) คือการจัดประโยคให้อยู่ในโครงสร้างของ SRL ส่วนที่สามคือการเปรียบเทียบความคล้ายคลึง (Similarity comparison) โดยจะทำการเปรียบเทียบความเหมือนด้วยการเปรียบเทียบคำศัพท์ตามแต่ละหน้าที่ของคำ เข้าด้วยกัน ดังรูปที่ 1

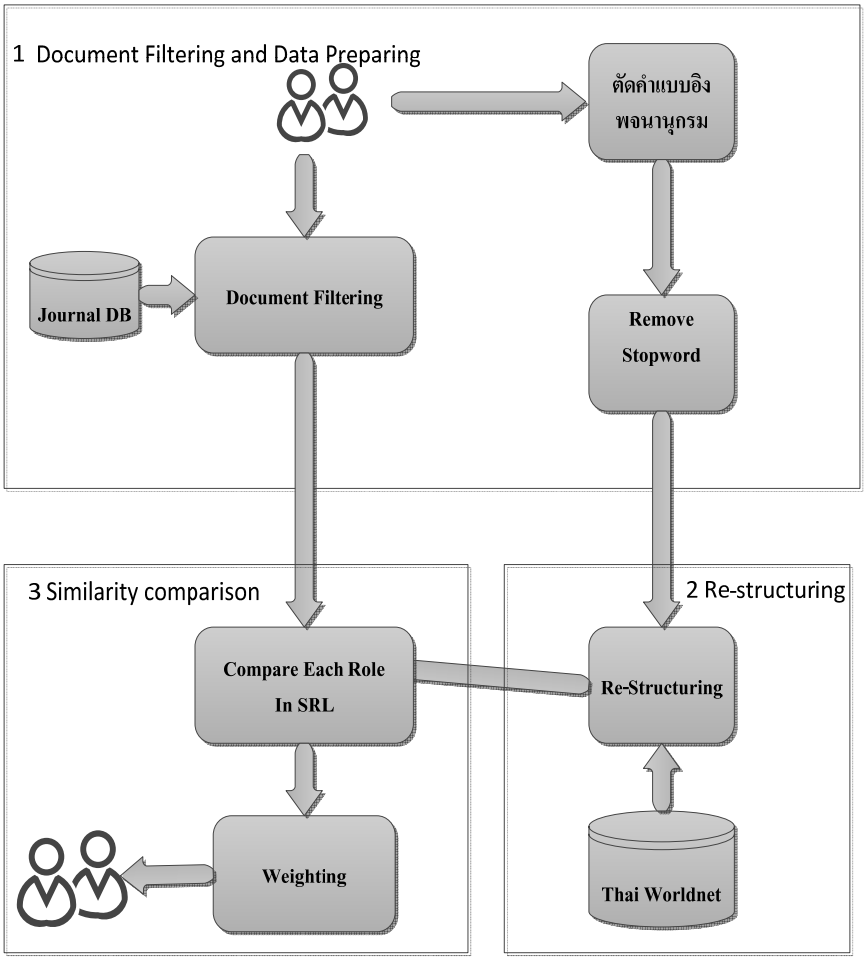
3.1 การคัดกรองเอกสารและการเตรียมข้อมูล (Document filtering and data preparing)

ในการตรวจสอบการคัดลอกงานวิจัยนั้นขั้นตอนที่ใช้เวลามากที่สุดคือขั้นตอนการเปรียบเทียบเชิงความหมายซึ่งต้องเปรียบเทียบเอกสารกลุ่มเสี่ยงทีละคู่ ทำให้จำเป็นต้องมีการจำกัดขอบเขตของเอกสารกลุ่มเสี่ยง ผู้ทำวิจัยได้นำเสนอวิธีในการจัดกัดของเขตโดยการเปรียบเทียบชื่องานวิจัยที่ต้องการตรวจสอบกับคำสำคัญของเอกสารกลุ่มเสี่ยงจากวิธีนี้สามารถลดเวลาในการตรวจสอบได้ เมื่อทำการคัดกรองเอกสารแล้วจึงทำการเตรียมข้อมูลงานวิจัยที่ต้องการตรวจสอบจะถูกตัดคำออกโดยใช้ Swath API ซึ่งเป็นเครื่องมือในการตัดคำที่มีความแม่นยำสูง โดยเปรียบเทียบกับคำศัพท์ในพจนานุกรม และทำการตัดคำที่ไม่สำคัญทิ้ง (Stop word) โดยคำที่ถูกตัดออกไปได้แก่คำที่เป็นสัญลักษณ์ที่ไม่มีความหมายและคำที่มีความถี่การถูกใช้บ่อยเกินกว่า 80% โดยใช้ข้อมูลจากงานวิจัยกลุ่มเสี่ยง เมื่อทำขั้นตอนนี้แล้วพบว่าการเปลี่ยนประโยคให้อยู่ในโครงสร้างของ SRL ใช้เวลาลดลง

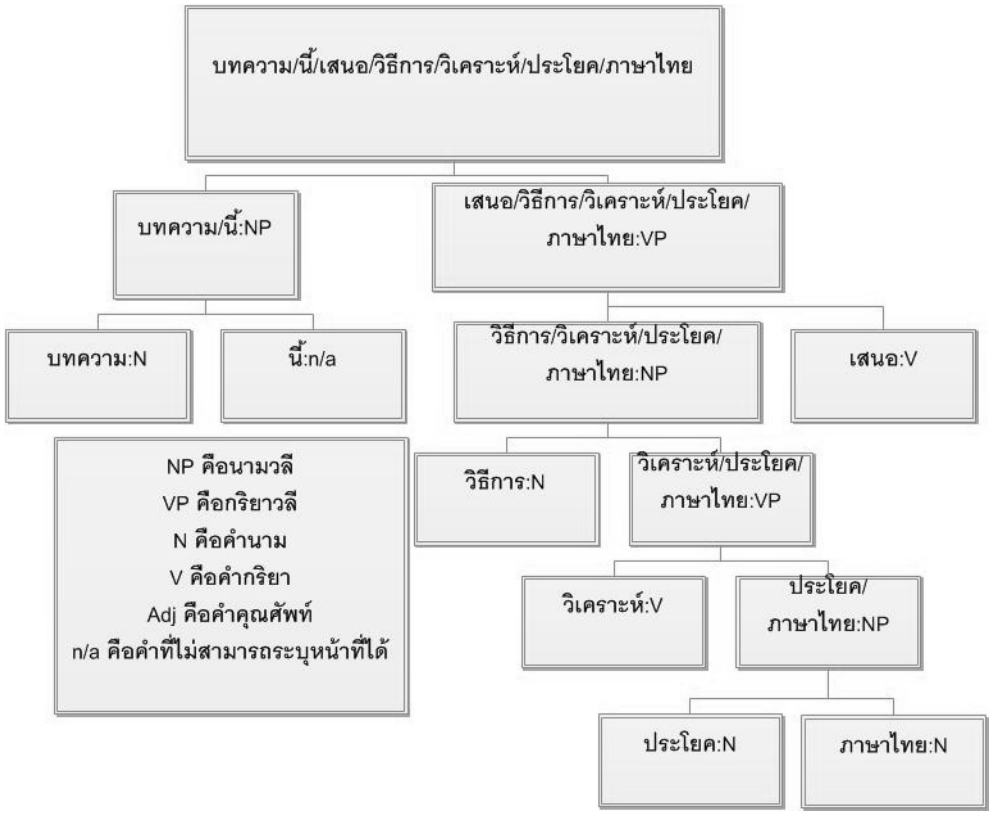
3.2 การเปลี่ยนโครงสร้างประโยค (Re-structuring)

การจัดโครงสร้างของประโยคใหม่ให้อยู่ในโครงสร้างของ SRL โดยเป็นขั้นตอนที่ใชเวลามากที่สุดรองจากขั้นตอน

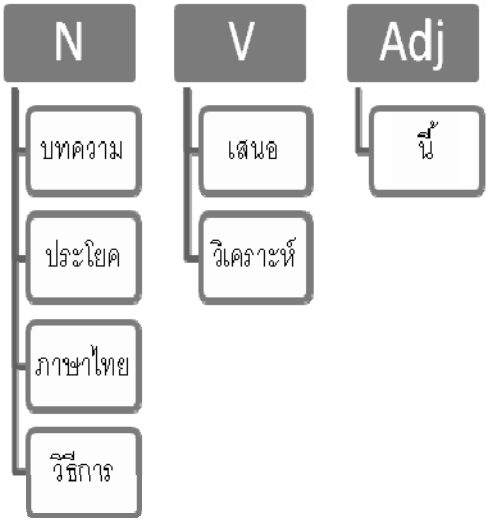
การเปรียบเทียบโดยจะต้องทำการค้นหาข้อมูลคำคล้ายและหน้าที่ของคำแต่ละคำในประโยคออกมาก่อนเมื่อได้ข้อมูลครบแล้วจะทำการวิเคราะห์โครงสร้างโดยใช้เทคนิคการวิเคราะห์ส่วนประชิด(5) ซึ่งจะจัดประโยคเป็นสองส่วนคือนามวลี (noun phrase)และกริยาวลี (verb phrase) โดยจะเริ่มค้นหาคำนามและคำกริยาที่อยู่ติดกันเพื่อแบ่งประโยคเป็นประโยคย่อยทำแบบนี้ไปเรื่อยๆ หากประโยคย่อยใดมีสมาชิกเพียงคำเดียวก็จะถูกระบุหน้าที่ของคำทันทีโดยผลลัพธ์ที่ได้จากขั้นตอนนี้คือประโยคที่อยู่ในโครงสร้าง SRL ดังรูปที่ 2 ในขั้นตอนนี้ผู้วิจัยพบว่ามีคำศัพท์จำนวนมากที่ไม่ถูกเก็บในฐานข้อมูล Thai WordNet ทำให้ไม่สามารถหาหน้าที่ของคำหรือคำที่มีความหมายเหมือนกันของคำศัพท์นั้นได้ ดังนั้นผู้วิจัยได้เสนอวิธีแก้ไขปัญหาโดยการเก็บข้อมูลสถิติโครงสร้างของประโยคและทำการเลือกโครงสร้างที่ใกล้เคียงที่สุดมาเปรียบเทียบเพื่อหาหน้าที่ของคำศัพท์ที่หายไปนั้นเช่นคำว่า "นี้" ไม่มีในฐานข้อมูลคำที่มีความหมายเหมือนกัน แต่จากการใช้เทคนิคการแบ่งกลุ่มพบว่าคำว่า "นี้" เป็นคำคุณศัพท์สำหรับคำและคำที่มีความหมายคล้ายจะถูกนำไปแยกประเภทไว้ตามแต่ละหน้าที่ของคำดังรูปที่ 3 เพื่อรอการเปรียบเทียบในขั้นตอนต่อไป



รูปที่ 1 กรอบแนวคิดการวิจัย



รูปที่ 2 โครงสร้าง SRL ของประโยค บทความนี้เสนอวิธีการวิเคราะห์ประโยคภาษาไทย



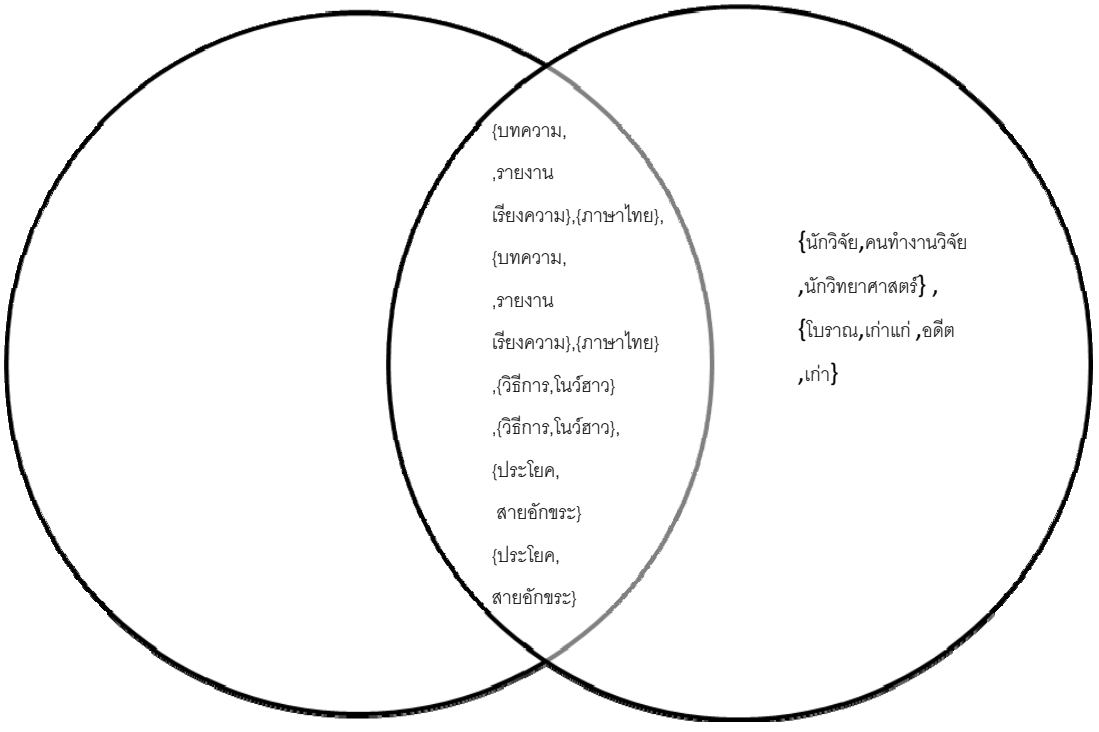
รูปที่ 3 ตัวอย่างการแบ่งกลุ่มคำตามหน้าทีของประโยค "บทความนี้เสนอวิธีการวิเคราะห์ประโยคภาษาไทย"

3.3 การเปรียบเทียบความคล้ายคลึง (Similarity comparison)

เป็นขั้นตอนที่ทำการเปรียบเทียบประโยคถูกจัดอยู่ในโครงสร้างของ SRL จากขั้นตอนที่แล้วกับประโยคที่อยู่ในฐานข้อมูลซึ่งถูกจัดอยู่ในโครงสร้าง SRL อยู่แล้วโดยระบบจะจัดคำและคำที่มีความหมายคล้ายกันจากทั้งสองประโยคเข้าเป็นเซ็ตหน้าที่ของคำและทำการนับคำที่ซ้ำกันโดยใช้สมการที่ 1

$$\text{similarity}(\text{Role}_i) = \frac{C(\text{Args}_j) \cap C(\text{Args}_k)}{C(\text{Args}_j) \cup C(\text{Args}_k)} \quad (1)$$

โดยให้ $\text{similarity}(\text{Role}_i)$ คือค่าความเหมือนของคำที่มีหน้าที่ของคำ i



รูปที่ 4 จำลองการเปรียบความคล้ายของหน้าที่ของคำ N(คำนาม) ของประโยค บทความนี้เสนอวิธีการวิเคราะห์ประโยคภาษาไทยและรายงานนี้นักวิจัยเสนอวิธีการวิเคราะห์สายอักขระของคำภาษาไทยโบราณ

$C(Args_j)$ คือ กลุ่มคำ j ในเซ็ตหน้าที่ของคำ i

$C(Args_k)$ คือ กลุ่มคำ k ในเซ็ตหน้าที่ของคำ i

จากสมการที่ (1) ให้ค่าความคล้ายของแต่ละหน้าที่ของคำเท่ากับจำนวนคำที่ซ้ำกันของแต่ละหน้าที่ของกลุ่มคำในหน้าที่ i และ j หารด้วยจำนวนคำทั้งหมดของสองของหน้าที่นั้นรวมกัน จากรูปที่ 4 พบว่าค่าความเหมือนของหน้าที่ของคำ N = 8/10=0.8 และได้ค่าความเหมือนของแต่ละหน้าที่ของคำทั้งหมดออกมาดังตารางที่ 1

เมื่อได้ค่าความเหมือนของแต่ละหน้าที่ของคำแล้วจึงสามารถหาค่าความเหมือนของประโยคหาได้จากสมการที่ 2

$$\text{similarity}\left(S_j, S_k\right) = \frac{\sum_{i=1}^{n} \text{similarity}(\text{Role}_i)}{n} \qquad (2)$$

โดยให้ $\text{similarity}\left(S_j, S_k\right)$ คือค่าความเหมือนของประโยคที่ j และ k

n คือจำนวนหน้าที่ของคำทั้งหมดในประโยคที่ถูกนำไปเปรียบเทียบ

ตารางที่ 1 ตัวอย่างตารางค่าความเหมือนที่ได้จากสมการที่ 1

หน้าที่ของคำ	ค่าความเหมือน
N	0.8
V	1
Adj	0

จากสมการที่ 2 ให้ค่าผลรวมของค่าความเหมือนในแต่ละหน้าที่หารด้วยจำนวนหน้าที่ในประโยค เมื่อนำไปคำนวณในการเปรียบเทียบ ประโยค “ในบทความเสนอวิธีการวิเคราะห์ประโยคภาษาไทย”และ”รายงานนี้นักวิจัยเสนอวิธีการวิเคราะห์สายอักขระของคำภาษาไทยโบราณ” พบว่าค่าความเหมือนของประโยคจากสมการที่ 2 คือ

(0.67+1+0)/3 =0.55 จากสมการนี้ผู้ใช้พบว่าหากประโยคที่ถูกคัดลอกมีการใส่คำขยายแล้วจะทำให้ค่าความเหมือนถูกลดไปมากเนื่องจากค่า n ถูกทำให้เพิ่มขึ้น ผู้ทำวิจัยจึงได้มีการให้ค่าลำดับความสำคัญของแต่ละหน้าที่ของคำ (role) ใหม่ด้วยสมการที่ 3

$$\text{Weight}(\text{Role}_i)=\frac{C(\text{Args}_j)+C(\text{Args}_k)}{\sum_{i=0}^{k} C(\text{Args}_j)+\sum_{i=0}^{l} C(\text{Args}_k)} \quad (3)$$

โดยให้ $\text{Weight}(\text{Role}_i)$ คือค่าน้ำหนักของหน้าที่ของคำ 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 3 เมื่อนำไปคำนวณหาค่าน้ำหนักของคำของประโยคของประโยค บทความนี้เสนอวิธีการวิเคราะห์ประโยคภาษาไทยและรายงานนี้นักวิจัยเสนอวิธีการวิเคราะห์สายอักขระของคำภาษาไทยโบราณ ได้ผลออกมาดังตารางที่ 2

ตารางที่ 2 ตัวอย่างตารางค่าน้ำหนักของแต่ละหน้าที่ของคำที่ได้จากสมการที่ 3

หน้าที่ของคำ	ค่าน้ำหนัก
N	0.72
V	0.29
Adj	0.14

ซึ่งพบว่าค่าน้ำหนักของคำจะมีค่าแปลผันตรงกับจำนวนคำในแต่ละหน้าที่นั้นๆ จากการหาค่าน้ำหนักทำให้สมการค่าความเหมือนจะถูกเปลี่ยนแปลงเป็นสมการที่ 4

$$\text{similarity}\left(\text{Role}_i\right)= \text{Weight}\left(\text{Role}_i\right)\times\frac{C(\text{Args}_j)\cap C(\text{Args}_k)}{C(\text{Args}_j)\cup C(\text{Args}_k)} \quad (4)$$

เมื่อนำสมการที่ (4) ไปใช้ในการเปรียบเทียบกับตัวอย่าง ได้ผลลัพธ์ดังตารางที่ 3

ตารางที่ 3 ตัวอย่างตารางค่าความเหมือนที่ได้จากสมการที่ 4

หน้าที่ของคำ	ค่าความเหมือน
N	0.54
V	0.29
Adj	0

จากสมการที่ 4 จะเห็นว่าค่าความเหมือนของแต่ละหน้าที่ของคำนั้นได้ผ่านการให้ค่าน้ำหนักมาแล้วทำให้ค่าความเหมือนของประโยคในสมการที่ 2 ที่มีการให้ค่าน้ำหนักของแต่ละคำเท่าๆกัน จะถูกเปลี่ยนแปลงเป็นการให้ค่าน้ำหนักของคำตามจำนวนคำจึงทำให้สมการที่2ถูกปรับปรุงกลายเป็น สมการที่ 5

$$\text{similarity}\left(S_j,S_k\right)=\sum_{i=1}^{n}\text{similarity}(\text{Role}_i) \quad (5)$$

เมื่อนำสมการที่ 5 ไปใช้ในการคำนวณค่าความเหมือนของประโยค "บทความนี้เสนอวิธีการวิเคราะห์ประโยคภาษาไทย"และ"รายงานนี้นักวิจัยเสนอวิธีการวิเคราะห์สายอักขระของคำภาษาไทยโบราณ" ได้ค่าความเหมือนเป็น 0.83 ซึ่งสูงกว่าค่าความเหมือนที่ได้จากสมการที่ 2 จึงจะทำให้สามารถแก้ไขปัญหาของจากการเปรียบเทียบประโยคสองประโยคที่มีจำนวนคำแตกต่างกันหรือปัญหาการคัดลอกแบบใส่คำขยายได้เพราะค่าน้ำหนักของคำที่ถูกใส่คำขยายจะมีค่าน้อยทำให้ไม่ส่งผลกับค่าความเหมือนของประโยค

ตารางที่ 4 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

4. ผลการวิจัยและอภิปราย

ผู้วิจัยสร้างข้อมูลทดสอบโดยคัดลอกประโยคของงานวิจัยที่ถูกเก็บในระบบตามประเภทการคัดลอกทั้ง 4 รูปแบบ รูปแบบละ 100 ประโยค เมื่อได้ข้อมูลทดสอบแล้วจึงนำมาทดสอบประสิทธิภาพของงานวิจัย โดยมีวิธีการเปรียบเทียบจากวิธีการเดิมที่มีอยู่ 3 วิธีคือ Anti-kobpae, Turnit-in และ Traditional SRL โดยกำหนดให้เอกสารที่ถูกคัดลอกจะต้องมีค่าความเหมือนมากกว่าหรือเท่ากับ 0.6 และใช้ Precision และ Recall เป็นมาตรวัดความถูกต้อง โดยมีสมมติฐานว่า Swath API สามารถตัดคำได้ถูกต้องในระดับที่ยอมรับได้และได้ผลการทดลองดังตารางที่ 4

จากตารางที่ 4 พบว่าทุกเทคนิคสามารถตรวจจับการคัดลอกแบบคำต่อคำมีค่า Precision และ Recall เป็น 1.0 เหมือนกันหมด สำหรับการคัดลอกแบบสลับตำแหน่งพบว่า เครื่องมือ Anti-kobpae ไม่สามารถตรวจสอบการคัดลอกประเภทนี้ได้และเครื่องมือ Turnit-in มีค่า Precision และ Recall ที่ลดต่ำลงในขณะที่ เทคนิค T-SRL และ W-SRL ยังสามารถมีค่า Precision และ Recall ที่ 1.0 เช่นเดิมเนื่องจากใช้การเปรียบเทียบคำศัพท์ตามหมวดหมู่ของคำเมื่อมีการสลับตำแหน่งของคำจึงไม่ส่งผลกับค่าความเหมือน สำหรับการคัดลอกแบบเปลี่ยนไปใช้คำคล้ายพบว่า Anti-kobpae และ Turnit-in มีค่า Precision และ Recall เท่ากับ 0 เนื่องจากไม่มีการเปรียบเทียบเชิงความหมาย ในขณะที่ W-SRL มีค่า มีค่า Precision และ Recall เท่ากับ 0.69 และ 0.83 ซึ่งสูงกว่า T-SRL เนื่องจาก W-SRL ใช้เทคนิคการแบ่งกลุ่มในการทำนายหน้าที่ของคำศัพท์ในกรณีที่ไม่พบคำศัพท์ในฐานข้อมูล Thai WordNet [6] 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 ได้มีการนำค่าน้ำหนักมาช่วยในการคำนวณค่าความเหมือนทำให้สามารถป้องกันการคัดลอกโดยใส่คำขยายได้ดีมากขึ้น

5. บทสรุปและงานในอนาคต

ผู้ทำวิจัยได้นำเสนอแนวคิดของการของการตรวจจับการคัดลอกด้วยใช้เทคนิค SRL ร่วมกับการให้น้ำหนักของคำ เทคนิคการหาจุดสิ้นสุดของประโยคร่วมกับเทคนิคการวิเคราะห์ความหมายของคำโดยใช้ฐานข้อมูลของ Thai WordNet ผลการทดลองพบว่าเพิ่มประสิทธิภาพการตรวจสอบการคัดลอกความคิดทางวิชาการได้อย่างมีประสิทธิภาพมากขึ้น โดยประเมินจากค่า Precision และ Recall เมื่อเปรียบเทียบกับ Anti-kobpae และ Turnit-in

แนวทางการพัฒนาในอนาคตคือปรับปรุงวิธีการหาจุดสิ้นสุดของประโยคในภาษาไทย จากผลการทดลองพบว่าหากแบ่งประโยคออกจากบทความผิดจะส่งผลกับการเปรียบเทียบค่าความเหมือน เนื่องจาก Weighted SRL ได้ใช้หลักการในการเปรียบเทียบประโยคต่อประโยค หากแบ่งประโยคผิดจะทำให้การเปรียบเทียบเกิดความผิดพลาดขึ้นได้ ยกตัวอย่างเช่น การใช้ต้นไม้ตัดสินใจ[7] ในการแบ่งประโยคภาษาไทย

6. เอกสารอ้างอิง

[1] Schleimer S, Wilkerson DS, Aiken A. Winnowing: local algorithms for document fingerprinting. Proceedings of the 2003 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM; 2003. p. 76–85.

[2] Du Z, Wei-jiang L, Zhang L. A Cluster-Based Plagiarism Detection Method CLEF. 2010.

[3] Gipp B, Meuschke N. Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. Proceedings of the 11th ACM symposium on Document engineering. New York, NY, USA: ACM; 2011. p. 249–58.

[4] Osman AH, Salim N, Binwahlan MS, Alteeb R, Abuobieda A. An improved plagiarism detection scheme based on semantic role labeling.Appl Soft Comput. 2012;12(5):1493–502.

[5] Ruengdet P.Thailand sentence analysis2007.

[6] Kergrit Robkop ST, Thatsanee Charoenporn VS, Hitoshi I. WNMS: Connecting the Distributed WordNet in the Case of Asian WordNet. The 5th International Conference of the Global WordNet Association (GWC-2010). 2010.

[7] Charoenpornsawat P,Sornlertlamvanich V. Automatic Sentence Break Disambiguation for Thai 2001.