การวิเคราะห์เนื้อหาข่าวภาษาไทยเพื่อติดแท็กอัตโนมัติ โดยใช้เทคนิคการสกัดคำ

The Thai News Analysis to Auto Tagging using Text Extraction

องอาจ อุ่นอนันต์[1*] สุนทรี แก่นแก้ว[1] ณพงษ์ วรรณพิรุณ[1] และ ทักศิณา คงสมลาภ[1]

Aongart Aun-a-nan[1*], Soontaree Kankaew[1], Naphong Wannapirun[1]

and Thaksina Khongsomlap[1]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ในการวิเคราะห์เนื้อหาข่าวภาษาไทยเพื่อติดแท็กอัตโนมัติ โดยใช้เทคนิคการสกัดคำ ซึ่งทำให้สามารถวิเคราะห์หาคำสำคัญ (Keyword) จากข้อมูลหัวข้อข่าวและข้อมูลรายละเอียดของข่าว โดยพิจารณาจากค่าความถี่ของคำแต่ละคำในบทความแล้วเปรียบเทียบประสิทธิภาพความถูกต้องระหว่างการหาคำสำคัญด้วยคอมพิวเตอร์แบบอัตโนมัติ กับข้อมูลที่มีการหาคำสำคัญโดยมนุษย์ ซึ่งค่าถูกต้องเฉลี่ยเท่ากับ 67.4% ซึ่งในการติดแท็กโดยอัตโนมัติสามารถช่วยให้มีความรวดเร็วในการค้นคืนสารสนเทศได้

คำสำคัญ : การสกัดคำ การตัดคำหยุด การติดแท็ก เนื้อหาข่าวสาร

Abstract

This research aims to analyze Thai news content for auto-tagging using text extraction. The analyze keyword information from the headlines of news and news detail then consider on the frequency of each word in the article. Then compare the accuracy of the key words for auto-tagging and data used for the human, which was the average of 67.4%, in which the tag can automatically allow a quick retrieval of information.

Keyword : Text Extraction, Stop-word Removal, Tagging, News Content

[1] มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ

* Corresponding author. E-mail: aidedecampc31020@hotmail.co.th

บทนำ

ในปัจจุบันอินเตอร์เน็ตเข้ามามีบทบาทมาก ทำให้สำนักข่าวระดับแนวหน้าและสำนักข่าวท้องถิ่น มีการจัดทำเว็บไซต์เพื่อใช้ในการเผยแพร่ข่าวสาร สื่อสาร และประชาสัมพันธ์ข่าว ซึ่งเป็นทรัพยากรที่มีคุณค่าของทางสารสนเทศ จึงทำให้ข่าวสารที่มีคุณภาพมีปริมาณมาก ทำให้การค้นหายากในกรณีที่ไม่มีการกำหนดคำสำคัญ (Keyword) เนื่องจากเครื่องมือค้นหาจากเว็บไซต์ค้นหาข้อมูล (Search Engine) ทั่วไป เช่น Google Yahoo และ Bing ซึ่งเว็บค้นหาเหล่านี้ใช้วิธีการค้นหาแบบการค้นคืนโดยใช้คำสำคัญ (Keyword-based Search) เป็นหลัก (K. Norvag and R. Oyri, 2015) จึงทำให้ตามเนื้อหาของข่าวสารจำเป็นต้องติดแท็กคำสำคัญ แต่อย่างไรก็ตามในการติดแท็กยังใช้มนุษย์เป็นผู้ติดแท็กคำสำคัญ โดยจะให้มีแท็กในบทความข่าวทุกบทความเป็นเรื่องที่ยาก ถ้ามีการติดแท็กคำสำคัญโดยอัตโนมัติด้วยคอมพิวเตอร์จะทำให้มีความสะดวกในการติดแท็ก และรวดเร็วในการค้นหามากยิ่งขึ้น

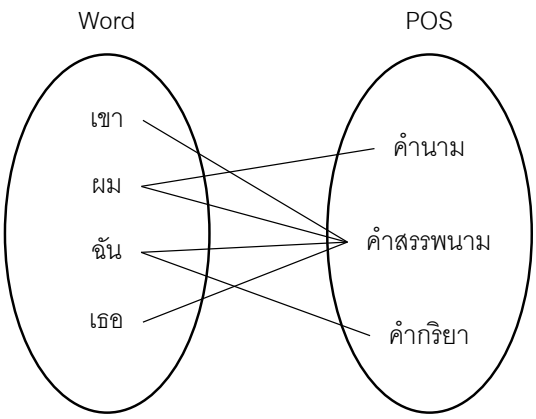
จากเหตุผลดังกล่าวงานวิจัยนี้จึงมีวัตถุประสงค์ในการศึกษาวิธีการสกัดคำสำคัญออกจากบทความข่าว เพื่อนำมาวิเคราะห์เนื้อหาข่าวและทำการติดแท็กคำสำคัญอัตโนมัติ โดยมีการประยุกต์ใช้เทคนิคการสกัดคำ (Text Extraction) ซึ่งการสกัดคำสำคัญด้วยมนุษย์เป็นเรื่องที่ง่าย แต่สำหรับการสกัดคำสำคัญด้วยคอมพิวเตอร์โดยอัตโนมัติเป็นปัญหาที่ยุ่งยาก (K. Norvag and R. Oyri, 2015) เนื่องจากภาษาธรรมชาติ มีรูปแบบของประโยคที่ซับซ้อน ซึ่งจะใช้การวจีวิภาค (Part-Of-Speech: POS) ในการช่วยวิเคราะห์คุณสมบัติของคำแต่ละคำ (C. Myint, 2011) เพื่อนำมาใช้ในการติดแท็กโดยอัตโนมัติ (Auto Tagging) และเป็นการป้องกันการเกิดข้อผิดพลาดในการติดแท็กด้วยมนุษย์ (Manual Tagging) และการเพิ่มประสิทธิภาพในการค้นคืนข้อมูลสารสนเทศ (C. Bouras and V. Tsogkas, 2010) ได้อย่างรวดเร็ว

ในงานวิจัยนี้ได้ใช้ชุดข้อมูลข่าวสารภาษาไทยที่มีการติดแท็กโดยใช้ทุกประเภทข่าว เช่น การเมือง เศรษฐกิจ กีฬา สังคม และเทคโนโลยี เพื่อนำมาสกัดคำให้ใช้ได้กับทุกประเภทของข่าว และหาคำสำคัญแล้วนำไปติดแท็กให้กับข่าวสารนั้น โดยให้เหมือนกับที่ใช้มนุษย์ในการติดแท็กของข่าวสารนั้น ในงานวิจัยได้แบ่งเนื้อหาเป็นส่วน ดังนี้ ส่วนที่ 2 ทฤษฎีที่เกี่ยวข้อง ส่วนที่ 3 วิธีดำเนินงาน ส่วนที่ 4 ผลการดำเนินงาน และส่วนที่ 5 สรุปผลและข้อเสนอแนะ

ทฤษฎีที่เกี่ยวข้อง

วจีวิภาค (Part of Speech: POS)

ในโครงสร้างของประโยคจะประกอบไปด้วยคำหลายคำ ซึ่งคำในแต่ละคำมีหน้าที่ของตนเอง ทำให้มีการแบ่งคำตามหมวดหมู่ตามการใช้งานของคำเรียกวิธีการนี้ว่าวจีวิภาค (Part of Speech: POS) โดยสามารถแบ่งเป็นหมวดหมู่ เช่น คำนาม (Noun) คำสรรพนาม (Pronoun) และคำกริยา (Verb) เป็นต้น (Javed A. Mahar and Ghulam Q. Memon, 2010) โครงสร้างของประโยคในแต่ละภาษาก็จะมีความแตกต่างกัน เช่น ในภาษาพม่า แบ่งออกเป็น 9 ส่วน (C. Myint, 2011) ภาษาอังกฤษและภาษาไทย แบ่งออกเป็น 8 ส่วน โดยปัญหาของ POS คือคำที่คลุมเครือเป็นคำที่สามารถอยู่ได้ในหลายกลุ่ม เช่น “Fly” ในภาษาอังกฤษสามารถเป็นได้ทั้งคำนามและคำกริยา “ฉัน” ในภาษาอังกฤษสามารถเป็นได้ทั้งคำนามและคำกริยา (S. Tasharofi et al., 2007) ดังภาพที่ 1



ภาพที่ 1: POS กลุ่มคำคลุมเครือ

โดยเทคนิค POS เป็นสิ่งที่สำคัญในขั้นตอนทำการประมวลภาษาธรรมชาติ (Natural Language Processing: NLP) (C. Myint, 2011) เช่น ระบบรู้จำด้วยเสียง (Speech Recognition) ระบบการอ่านออกเสียง (Text to Speech) การแก้ไขคำกำกวม (Word Sense Disambiguation) การค้นคืนสารสนเทศ (Information Retrieval) และกระบวนการเชิงความหมาย (Semantic Processing)

การสกัดคำ (Text Extraction)

การสกัดคำนั้นเป็นกระบวนการของการทำเหมืองข้อความ (Text Mining: TM) เพื่อใช้การวิเคราะห์คำออกจากเอกสาร ข่าวสาร ข้อความ และสารสนเทศต่างๆ ที่เป็นตัวอักษร โดยสามารถนำไปทำการแบ่งกลุ่ม (Clustering) จำแนกกลุ่ม (Classification) และหาความสัมพันธ์ (Associate) ซึ่งการแบ่งกลุ่มเอกสาร (Document Clustering) (W. Lam, 2004) เป็นการวัดความคลายคลึงกันของข้อความในตัวเอกสาร โดยข้อมูลตัวอักษรจะถูกแปลงเป็นเมทริกซ์ตัวเลขเพื่อให้คอมพิวเตอร์สามารถเข้าใจได้ และทำขั้นตอนการแบ่งกลุ่มโดยใช้เทคนิคต่างๆ เช่น DBScan K-mean SOM และ Hierarchical แต่ก่อนการทำเหมืองข้อความ จะต้องผ่านขั้นตอนการเตรียมข้อมูล (Preprocess) ก่อนซึ่งมีขั้นตอน (X. Dai et al., 2010; U. Gunasinghe et al., 2012; K. Norvag and R. Oyri, 2005) ดังนี้

1. การตัดคำ (Word Segmentation) เป็นการแยกแต่ละคำจากเอกสารออกจากกัน โดยยังคงมีความหมายที่ถูกต้องสมบูรณ์อยู่ โดยการตัดคำนั้นใช้ฐานข้อมูลพจนานุกรมคำศัพท์ ในการแบ่งคำออกมา
2. การกำจัดคำหยุด (Stop-word Removal) เป็นการตัดคำที่ไม่มีความหมายออกจากเอกสาร โดยการกำจัดคำหยุดนั้นใช้ฐานข้อมูลคำศัพท์ที่เป็นคำที่ไม่มีความหมาย เช่น “กับ” “แก่” “แด่” “ต่อ” “ที่” “ซึ่ง” เป็นต้น ในการกำจัดคำออก เมื่อทำการตัดคำเรียบร้อยแล้วจึงทำการเลือกคำที่ต้องการใช้ในการวิเคราะห์ (Feature Selection)

ระบบแท็ก (Tagging)

ระบบแท็กเป็นวิธีการใหม่ในการจัดการข้อมูล ใช้คำสำคัญ (Key Word) แทนข้อมูลทั้งหมด ช่วยทำให้ข้อมูลมีระเบียบ ง่ายในการค้นหาทำให้ประสิทธิภาพในการค้นหาเพิ่มมากยิ่งขึ้น โดยในการจัดการแท็กมีหลายวิธี เช่น Collaborative Tagging, Auto Tagging และ Manual Tagging (M. Dostal and K. Jezek, 2010)

1. Collaborative Tagging เป็นวิธีการหนึ่งโดยที่ผู้ใช้งานสามารถเลือกติดแท็กที่มีการกำหนดมาให้อยู่ก่อนแล้ว และสามารถสร้างแท็กใหม่ขึ้นมาได้เองในกรณีที่ไม่มีแท็กนั้นอยู่ ข้อเสีย คือ ระบบไม่สามารถทำการกำหนดแท็ก

ไว้ได้ครอบคลุมข้อมูลทั้งหมดที่จะมีเข้ามาใหม่อยู่เสมอ โดยในความเป็นจริงการติดแท็กที่เหมาะสมนั้นจะต้องไม่เลือกติดแท็ก จากแท็กที่ทำกำหนดไว้ก่อนล้วงหน้า (Sanjay C. Sood et al., 2007) เนื่องจากแท็กที่กำหนดไว้นั้นอาจไม่ตรงกับความต้องของผู้ใช้งานที่จะติดแท็กอื่นที่ไม่มี เป็นการจำกัดขอบเขตของบทความที่ต้องการนำเสนอ

2. Auto Tagging เป็นการติดแท็กแบบอัตโนมัติ มีพื้นฐานบน Corpuses โดยมีการเรียนรู้จากข้อมูลด้วยตนเอง หรือกระบวนการคำนวณจากหลักสถิติ ส่วนมากใช้การนับความถี่ของคำ แต่การเรียนรู้ด้วยตนเองใช้การจัดกลุ่มข้อมูลมาใช้ในการวิเคราะห์ข้อมูลที่ต้องการจะติดแท็ก

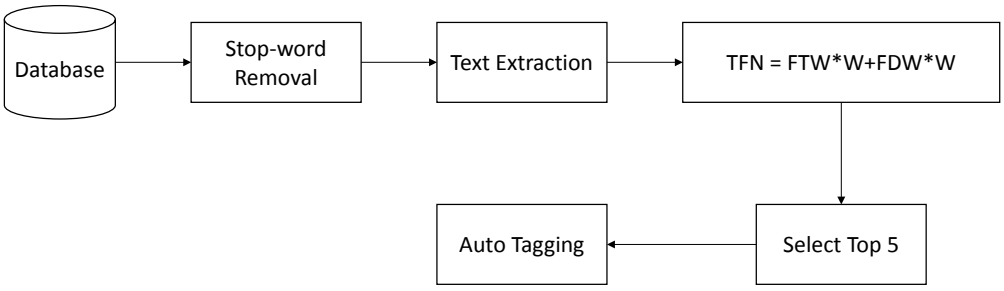
3. Manual Tagging เป็นการติดแท็กแบบอิสระโดยผู้ใช้งาน ซึ่งวิธีการนี้เป็นเรื่องยากในการตรวจสอบแท็กว่ามีความถูกต้องหรือเป็นคำสำคัญจริง ๆ ของบทความ ซึ่งมีความผิดพลาดสูง

งานวิจัยที่เกี่ยวข้อง

ในงานวิจัยส่วนใหญ่ได้ใช้เทคนิค POS ในการจัดการคำในประโยค โดยได้ใช้ POS ในการแปลงประโยคจากภาษาพม่าเป็นภาษาอังกฤษ (C. Myint, 2011) และใช้ในการแปลงภาษาเปอร์เซียเป็นภาษาอังกฤษเช่นกัน (S. Tasharofi etal., 2007) ในส่วนของการสกัดคำนั้นมีงานทำการสกัดเนื้อหาของข่าวออกจากหน้าเว็บไซต์ ซึ่งในหน้าเว็บจะมีข้อมูลที่ไม่เป็นประโยชน์อยู่ด้วย (K. Norvag and R. Oyri, 2005) เช่น โฆษณา หรือลิงก์ที่ไม่เกี่ยวข้อง โดยใช้การสกัดจากแท็ก HTML ในส่วนงานของการติดแท็กอัตโนมัติ (Sanjay C. Sood et al., 2007) ทำการติดแท็กโดยการนับความถี่จากบทความในบล็อก เพื่อกำหนดประเภทของบทความ โดยมีค่าความแม่นยำอยู่ที่ 42.10 %

วิธีดำเนินงาน

วิธีดำเนินงานวิจัยสำหรับวิเคราะห์เนื้อหาข่าวเพื่อการติดแท็กอัตโนมัติ โดยใช้เทคนิคการสกัดคำ ได้เตรียมชุดข้อมูล การตัดคำหยุด การสกัดคำ การคำนวณหาค่าคำสำคัญ และการวัดประสิทธิภาพ โดยมีการออกแบบขั้นตอนการดำเนินงาน แสดงดังภาพที่ 2



ภาพที่ 2: ขั้นตอนการดำเนินงาน

ข้อมูลที่ใช้ในการวิจัย

ข้อมูลที่ใช้เพื่อการวิเคราะห์ และวัดประสิทธิภาพ เป็นข้อมูลข่าวที่มีการติดแท็กบนเว็บไซต์ข่าวภาษาไทย เก็บในรูปแบบ HTML มีข้อมูลหัวข้อข่าว ข้อมูลแหล่งที่มาของข่าว ข้อมูลลิงก์ข่าว ข้อมูลวันที่เวลาเก็บข้อมูล ข้อมูลวันที่ลงข่าว ข้อมูลหัวข้อข่าว ข้อมูลรายละเอียดข่าว จำนวน 50 ข่าว โดยมีรายละเอียดของขอบเขตข้อมูล แสดงดังตารางที่ 1

ตารางที่ 1: ตารางรายละเอียดชุดข้อมูล

No	Name	Detail	Type
1	Title	หัวข้อข่าว Webpage	Char
2	Source	แหล่งข่าว	Char
3	Link	ลิงก์ข่าว	Text
4	Querytime	ข้อมูลวันที่เวลาเก็บข้อมูล	Datetime
5	Articledate	วันที่ลงข่าว	Char
6	Heading	หัวข้อข่าว	Char
7	Article	รายละเอียดข่าว	Text

การตัดคำหยุด (Stop Word Removal)

ในขั้นตอนนี้เป็นการทำ Preprocessing เพื่อกำจัดคำที่ไม่มีความหมาย ออกจากเนื้อหาและหัวข้อข่าว โดยคำเหล่านี้ถูกเก็บไว้ในพจนานุกรมข้อมูล (*Data Dictionary*: DD) จะเป็นคำที่มี POS Tagger อยู่ 7 ประเภท ได้แก่ คำสรรพนาม (Pronoun) คำคุณศัพท์ (Adjective) คำวิเศษณ์ (Adverb) คำบุพบท (Preposition) คำสันธาน (Conjunction) คำอุทาน (Interjection) และสัญลักษณ์ (Symbol) ดังตารางที่ 2

ตารางที่ 2: ตารางตัวอย่างคำหยุด

No	POS Tagger	Sample Word
1	Pronoun	ฉัน คุณ เรา เขา เธอ มัน
2	Adjective	ที่ ใหม่ เก่า แก่ ดี แค่ เพียง
3	Adverb	ตอนนี้ วันนี้ เร็วๆนี้
4	Preposition	บน ใน ของ ก่อนหน้า หลังจาก ล่าง
5	Conjunction	และ หรือ ถ้า อย่างไรก็ตาม ก็ต่อเมื่อ แล้ว
6	Interjection	โอ้ ว้าว ไซโย
7	Symbol	(,), ?, !, "...", :, -,

การสกัดคำ (Text Extractions)

ในขั้นตอนนี้เป็นการสกัดคำ ที่ผ่านขั้นตอนการตัดคำหยุดมาแล้ว ทำให้เหลือคำที่มี POS Tagger ประเภทคำนาม (Noun) และคำกริยา (Verb) โดยเอาคำที่ผู้วิจัยสนใจมาใช้เป็นคำสำคัญ ซึ่งกำหนดให้ใช้คำที่มี POS Tagger เป็นคำนาม จึงทำการตัดคำกริยาออก โดยคำกริยาเหล่านี้ถูกเก็บไว้ในพจนานุกรมข้อมูล

การคำนวณหาค่าคำสำคัญ

ในการคำนวณหาค่าคำสำคัญ ใช้ผลรวมของความถี่ (Total of Frequency News: TFN) ซึ่งเป็นค่าความถี่ของคำแต่ละคำร่วมกับค่าถ่วงน้ำหนัก ในหัวข้อข่าวและรายละเอียดข่าว โดยมีสมการดังนี้

$$TFN_i = FTW_i * W_1 + FDW_i * W_2 \quad (1)$$

โดยให้ *FTW* แทนค่าความถี่ของคำที่ *i* ในหัวข้อข่าว ให้ *FDW* แทนค่าความถี่ของคำที่ *i* ในรายละเอียดข่าว ให้ W_1 แทนค่าถ่วงน้ำหนักของคำในหัวข้อข่าวโดยมีค่าเท่ากับ 0.80 เนื่องหัวข้อข่าวจะพูดถึงภาพรวมของข่าวและมีความยาวของข้อความน้อย มีโอกาสที่คำจะเกิดขึ้นซ้ำกันน้อยจึงทำให้มีค่าน้ำหนักที่มากกว่า และให้ W_2 แทนค่าถ่วงน้ำหนักของคำในรายละเอียดข่าวโดยมีค่าเท่ากับ 0.20 เนื่องจากเนื้อหาข่าวมีความยาวของข้อความมากทำให้การซ้ำกันของข้อความมีโอกาสมาก

การวัดประสิทธิภาพ

ในขั้นตอนการวัดประสิทธิภาพนี้ ผู้วิจัยทำการเปรียบเทียบผลการทดลองติดแท็กอัตโนมัติ กับข้อมูลแท็กของข่าวจริงที่มีการแท็กจำนวน 50 ข่าว แล้วหาค่าความแม่นยำ

ผลการดำเนินงาน

คณะผู้วิจัยได้ดำเนินงานสำหรับการตัดคำหยุด การสกัดคำ การคำนวณหาค่าคำสำคัญ และผลการวัดประสิทธิภาพ

ผลการตัดคำหยุด

ในการดำเนินการตัดคำหยุด 7 ประเภท ตามข้อมูลที่อยู่ภายในพจนานุกรมข้อมูล มีการตัดคำหยุดได้อย่างถูกต้อง โดยคำที่แสดงออกมามีเพียงคำประเภทคำนาม และคำกริยาเท่านั้น ดังภาพที่ 3

กรมอุตุนิยมวิทยา | พยากรณ์ | ลักษณะ | อากาศ | ทั่วไป | วันที่ | 16 | มีนาคม | 2559 | ความกดอากาศ | ต่ำ | เนื่องจาก | ความร้อน | ปกคลุม | ประเทศไทย | ตอน | บน | ลักษณะ | เช่นนี้ | ทำให้ | บริเวณ | ดังกล่าว | มี | อากาศ | ร้อน | โดยทั่วไป | กับ | มี | อากาศ | ร้อน | จัด | หลาย | พื้นที่ | และ | มี | ฟ้าหลัว | ใน | ตอนกลางวัน | ส่วน | ลม | ตะวันออกเฉียงใต้ | พัด | นำ | ความชื้น | จาก | ทะเล | จีนใต้ | และ | อ่าวไทย | เข้า | ปกคลุม | ภาค | ตะวันออกเฉียงเหนือ | ตอน | ล่าง | ภาค | กลาง | และ | ภาค | ตะวันออก | ทำให้ | บริเวณ | ดังกล่าว | มี | ฝน | เกิดขึ้น | ได้ | ใน | ระยะนี้ | พยากรณ์อากาศ | สำหรับ | กรุงเทพฯ | และ | ปริมณฑล | เวลา | 17.00 | น. | วันนี้ | 17.00 | น. | วันพรุ่งนี้ | อากาศ | ร้อน | กับ | มี | ฟ้าหลัว | ใน | ตอนกลางวัน | และ | มี | ฝน | เล็กน้อย | อุณหภูมิ | ต่ำสุด | 27 | 28 | องศาเซลเซียส | อุณหภูมิ | สูงสุด | 32 | 38 | องศาเซลเซียส | ลม | ใต้ | ความเร็ว | 10 | 30 | กม. | ชม.

ภาพที่ 3: การตัดคำหยุด

ผลการสกัดคำ

ในการดำเนินการสกัดคำ โดยใช้คำนามเป็นคำสำคัญ จึงดำเนินการตัดคำกริยาตามพจนานุกรมข้อมูล มีการตัดคำได้อย่างถูกต้อง โดยคำที่แสดงออกมามีเฉพาะคำนามเท่านั้น ดังภาพที่ 4

กรมอุตุนิยมวิทยา | พยากรณ์ | อากาศ | ความกดอากาศ | ความร้อน | ปกคลุม | ประเทศไทย | บริเวณ | อากาศ | ร้อน | อากาศ | ร้อน | พื้นที่ | ฟ้าหลัว | ตอนกลางวัน | ลม | ตะวันออกเฉียงใต้ | พัด | นำ | ความชื้น | ทะเล | จีนใต้ | อ่าวไทย | ปกคลุม | ภาค | ตะวันออกเฉียงเหนือ | ภาค | กลาง | ภาค | ตะวันออก | ทำให้ | บริเวณ | ฝน | เกิดขึ้น | พยากรณ์อากาศ | กรุงเทพฯ | ปริมณฑล | เวลา | อากาศ | ร้อน | ฟ้าหลัว | ตอนกลางวัน | ฝน | เล็กน้อย | อุณหภูมิ | ต่ำสุด | องศาเซลเซียส | อุณหภูมิ | สูงสุด | องศาเซลเซียส | ลม | ใต้ | ความเร็ว |

ภาพที่ 4: การสกัดคำนาม

ผลการคำนวณหาค่าคำสำคัญ

ในการคำนวณหาค่าความถี่ของคำสำคัญ โดยใช้การคำนวณจากสมการที่ (1) โดยเมื่อได้ค่าผลรวมของความถี่ในแต่ละคำเก็บไว้ในตัวแปรอาเรย์ (Array) แล้วทำการเรียงลำดับจากผลรวมของค่าความถี่มากไปหาน้อย แล้วทำการเลือก 5 ลำดับแรกที่มากที่สุด และต้องค่า TFN มากกว่า 0.5 มาใช้เป็นคำสำคัญแล้วติดแท็กโดยอัตโนมัติ เช่น ร้อน (TFN = 2.2) อากาศ (TFN = 1.8) พยากรณ์ (TFN = 1.2) ภาค (TFN = 0.6) ดังภาพที่ 5

```
Title
Array ([ร้อน] => 2 [อุตุฯ] => 1 [พยากรณ์] => 1 [เย็น] => 1 [ไทย] => 1 [อากาศ] => 1)
Detail
Array ([อากาศ] => 5 [ร้อน] => 3 [ภาค] => 3 [พยากรณ์ ] => 2 [ปกคลุม] => 2 [บริเวณ] => 2 [ฟ้าหลัว] => 2
[ฝน] => 2 [ตอนกลางวัน] => 2 [ลม] => 2 [อุณหภูมิ] => 2 [องศาเซลเซียส] => 2 [กรมอุตุนิยมวิทยา] => 1
[ความกดอากาศ] => 1 [ความร้อน] => 1 [ประเทศไทย] => 1 [พื้นที่] => 1 [ตะวันออกเฉียงใต้] => 1 [พัด] => 1
[นำ] => 1 [ความชื้น] => 1 [ทะเล] => 1 [จีนใต้] => 1 [อ่าวไทย] => 1 [ตะวันออกเฉียงเหนือ] => 1 [กลาง] => 1
[ตะวันออก] => 1 [เกิดขึ้น] => 1 [กรุงเทพฯ] => 1 [ปริมณฑล] => 1 [เวลา] => 1 [เล็กน้อย] => 1 [ต่ำสุด] => 1
[สูงสุด] => 1 [ใต้] => 1 [ความเร็ว] => 1)
```

ภาพที่ 5: การเรียงลำดับค่าความถี่ของคำ

ผลการวัดประสิทธิภาพ

ในการวัดประสิทธิภาพได้วัดการแท็กของข่าวที่มีการติดมาก่อนแล้ว จากนั้นนำมาเปรียบเทียบหาค่าความถูกต้อง (Accuracy) กับคำสำคัญที่ได้จากคอมพิวเตอร์เลือก โดยได้ค่าความถูกต้องเมื่อเทียบกับแท็กของข่าวสารที่ติดมาจำนวน 223 คำ มีการติดแท็กถูกต้องจำนวน 150 คำ ทำให้มีค่าความถูกต้องเฉลี่ย มีค่าเท่ากับ 67.40 %

สรุปผลและข้อเสนอแนะ

การวิเคราะห์เนื้อหาข่าวภาษาไทยเพื่อการติดแท็กอัตโนมัติ ได้ใช้เทคนิคการสกัดคำ และการคำนวณหาค่าคำสำคัญ ทำให้หาคำสำคัญที่อยู่ภายในบทความข่าวได้ โดยมีค่าความถูกต้องเฉลี่ยเท่ากับ 67.40% สามารถติดแท็กโดยอัตโนมัติได้จากการหาคำที่มีค่าความถี่มากที่สุด 5 ลำดับแรก และมีค่า TFN มากกว่า 0.5

จากผลการทดลองมีค่าความถูกต้อง ยังไม่คงที่เนื่องจากวิธีการเลือกคำสำคัญของมนุษย์มีหลักในการเลือกที่แตกต่างกัน งานในอนาคตจึงควรทำการศึกษารูปแบบ วิธีการในการเลือกคำสำคัญ แล้วนำมาพัฒนาเป็นอัลกอริทึมสำหรับการเลือกคำสำคัญ ร่วมทั้งควรทดลองปรับค่าถ่วงน้ำหนักของหัวข้อข่าวและรายละเอียดข่าวที่มีความเหมาะสม เพื่อเพิ่มประสิทธิภาพความแม่นยำในการเลือกคำสำคัญให้มากขึ้น

เอกสารอ้างอิง

Bouras, C. and Tsogkas, V. 2010. Assigning Web News to Clusters. The 5th International Conference on Internet and Web Applications and Services (ICIW), 2010, 1-6.

Norvag, K. and Oyri, R. 2005. News Item Extraction for Text Mining in Web Newspapers. WIRI'05 International Workshop on Challenges in, 2005, 195-204.

Myint, C. 2011. A Hybrid Approach for Part-of-Speech Tagging of Burmese Texts. International Conference on Computer and Management (CAMAN), 2011, 1-4.

Mahar, J. A. and Memon, G. Q. 2010. Sindhi Part of Speech Tagging System Using Wordnet. Int. J. Comput. Theory Eng., vol. 2010, no. 4: 538-545.

Tasharofi, S., Raja, F., Oroumchian, F. and Rahgozar, M. 2007. Evaluation of statistical part of speech tagging of Persian text. The 9th International Symposium on Signal Processing and Its Applications (ISSPA), 2007. 1-4.

Lam, W., Cheung, P.-S. and Huang, R. 2004. Mining events and new name translations from online daily news. ACM/IEEE Conference on Digital Libraries, 2004, 287-295.

Dai, X., He, Y. and Sun, Y. 2010. A Two-layer Text Clustering Approach for Retrospective News Event Detection. International Conference on Artificial Intelligence and Computational Intelligence (AICI), 2010, vol. 1: 364-368.

Gunasinghe, U., Matharage, S. and Alahakoon, D. 2012. A sequence based dynamic SOM model for text clustering. International Joint Conference on Neural Networks (IJCNN), 2012, 1-8.

Dostal, M. and Ježek, K. 2010. Automatic tagging based on linked data: Unsupervised methods for the extraction of hidden information. IEEE International Conference on Service-Oriented Computing and Applications (SOCA), 2010, 1-4.

Sood, S.C., Owsley, S. H., Hammond K. J. and Birnbaum, L. 2007. TagAssist: Automatic Tag Suggestion for Blog Posts. ICWSM'2007, vol. 2007.