Research Article

Received: May 4, 2019; Accepted: November 8, 2019

ระบบให้คำอธิบายเชิงความหมายแบบกึ่งอัตโนมัติสำหรับข้อมูลด้านศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่น

A Semi-Automatic Semantic Annotation System for Culture and Folk Wisdom Information

สิรยา สิทธิสาร*

สาขาวิชาคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยทักษิณ

ตำบลบ้านพร้าว อำเภอป่าพะยอม จังหวัดพัทลุง 93210

Siraya Sitthisarn*

Department of Computer and Information Technology, Faculty of Science, Thaksin University,

Ban Phraw, Pa Phayom, Phatthalung 93210

บทคัดย่อ

งานวิจัยนี้นำเสนอการพัฒนาระบบให้คำอธิบายเชิงความหมายแบบกึ่งอัตโนมัติสำหรับให้คำอธิบายและสร้างข้อมูลเชิงความหมายในรูปแบบ RDF จากข้อความด้านศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นของจังหวัดพัทลุง ระบบนี้จะผสมผสานเทคนิคแบบ (1) unsupervised และ (2) supervised สำหรับการรู้จำประเภทนิพจน์ เบื้องต้นระบบจะใช้ unsupervised technique ในการจำแนกประเภทของนิพจน์โดยอาศัยออนโทโลยีและนำเสนอให้กับผู้ใช้ หากพบความกำกวมระบบจะให้ผู้ใช้เลือกประเภทนิพจน์ด้วยตนเอง จากนั้นจะนำข้อมูลจาก log ของการเลือกไปสร้างเป็น training data set สำหรับสร้างโมเดลการจำแนก เพื่อที่ต่อไปหากพบนิพจน์ดังกล่าวในบริบทที่คล้ายคลึงกันระบบจะสามารถระบุนิพจน์ได้ถูกต้อง ผลจากการประเมินประสิทธิผลการแนะนำนามนิพจน์ของระบบ พบว่ามีค่าเฉลี่ย precision 88.07 % และค่าเฉลี่ย recall 82.10 % หมายความว่าโปรแกรมระบุนามนิพจน์ได้แม่นยำและครอบคลุมครบถ้วนในระดับดี ส่วนการสกัดความสัมพันธ์ระหว่างนิพจน์ก่อนทำความสะอาดรูปประโยคประสิทธิผลอยู่ในระดับอัตโนมัติต่ำ ทั้งนี้เนื่องจากข้อจำกัดในเรื่องการประมวลผลประโยคภาษาธรรมชาติ แต่หากมีการทำความสะอาดประสิทธิผลของการสกัดหาความสัมพันธ์จะสูงขึ้น ดังนั้นในส่วนนี้จำเป็นต้องอาศัยผู้ใช้เข้ามาให้คำอธิบาย

คำสำคัญ : การให้คำอธิบายเชิงความหมาย; ข้อมูลเชิงความหมาย; RDF; ศิลปวัฒนธรรม

Abstract

This research aimed to develop a semi-automatic annotation system for annotating textual domain content that is a part of the intangible cultural heritage of Phattalung province in southern

***ผู้รับผิดชอบบทความ :** ssitthisarn@gmail.com **doi:** 10.14456/tstj.2020.148

Thailand. A combination of unsupervised and supervised techniques for named entity recognition was adopted in the system. The unsupervised techniques were used to identify the named entity by using ontology to enable the system to provide annotating terms for users. However, if the system finds an ambiguous entity, then the user's self-annotations are allowed and the system will record the annotations in the log file. The log file will then be provided as training sets for a classification model to enable the high effectiveness of named entity recognition since the system will classify the type of entity correctly when it discovers the ambiguous entity in a similar context. The system evaluation found that the effectiveness of named entity recognition was good with an average precision of 88.07 %, with an average recall of 82.10 %, while the effectiveness of the relationship extraction component with uncleaned sentence structure was low due to a limitation of natural language processing. However, if the structure of the sentence was cleaned, then the capability of the extraction would increase. Therefore, a user's self-annotation is required for relationship annotation to increase the correctness of annotations.

Keywords: semantic annotation; semantic data; RDF; intangible cultural heritage

1. บทนำ

ศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นเป็นองค์ความรู้ที่สะท้อนถึงเอกลักษณ์และอัตลักษณ์ของคนในท้องถิ่น ปัจจุบันมีหลายองค์กรตระหนักถึงความสำคัญและพัฒนาระบบจัดการความรู้แบบดิจิตอลขึ้น [1-3] โดยเทคโนโลยีที่นิยมใช้ ได้แก่ ระบบการจัดการไฟล์และระบบฐานข้อมูล อย่างไรก็ตาม เทคโนโลยีเหล่านี้ก็มีข้อจำกัดในเรื่อง (1) ไม่มีการใช้โมเดลองค์ความรู้ (knowledge model) ที่รองรับการสร้างความสัมพันธ์กันระหว่างเนื้อหาที่เกี่ยวข้อง (2) ระบบไม่สามารถเข้าใจความหมายของเนื้อหาความรู้เหล่านี้ เนื่องจากข้อมูลส่วนมากไม่มีโครงสร้างและอยู่ในรูปแบบ text file จึงส่งผลให้ระบบสารสนเทศไม่สามารถประมวลผลอย่างชาญฉลาด ตัวอย่าง เช่น ประสิทธิภาพการสืบค้นข้อมูลที่ขึ้นกับปรากฏของคำค้นในเอกสารเป็นหลักและการไม่สามารถนำเสนอเอกสารเชื่อมโยงเพิ่มเติมที่เกี่ยวข้องกับข้อมูลที่ผู้ใช้สนใจได้ เป็นต้น

ข้อมูลเชิงความหมายเป็นข้อมูลที่มีการเพิ่มส่วนความหมายหรือ metadata แล้วแสดง metadata พร้อมข้อมูลให้อยู่ในรูปแบบที่ระบบสารสนเทศสามารถเข้าใจและประมวลผลได้อย่างอัตโนมัติ ข้อมูลเชิงความหมายจะแทนในรูปแบบมาตรฐาน เช่น RDF/XML หรือ JSON ข้อมูลเชิงความหมายมีบทบาทสนับสนุนการจัดการความรู้โดยระบบสามารถเข้าใจความหมายของข้อมูลส่งผลให้ระบบสามารถประมวลผลอย่างมีประสิทธิภาพ [4] โดยเอกสารที่ไม่มีโครงสร้างสามารถแปลงเป็นเอกสารเชิงความหมายโดยอาศัยกระบวนการให้คำอธิบายเชิงความหมาย (semantic annotation) แม้ว่าการให้คำอธิบายเชิงความโดยอาศัยความสามารถของมนุษย์เพียงอย่างเดียวจะถูกต้องและเหมาะสม แต่ก็เป็นกระบวนการที่ค่อนข้างยุ่งยาก เสียเวลา และต้องอาศัยบุคคลากรที่เข้าใจโครงสร้างของออนโทโลยีและ syntax ของภาษาเชิงความหมาย เพื่อแก้ปัญหานี้นักวิจัยได้นำเสนอเครื่องมือสำหรับการให้คำอธิบายเชิงความหมายขึ้น ซึ่งสามารถแบ่งประเภทตามระดับของความอัตโนมัติ โดย

ระบบแบบอัตโนมัติเป็นระบบที่ผู้ใช้ไม่จำเป็นต้องเข้ามามีส่วนร่วมใด ๆ ในการให้คำอธิบาย ซึ่งช่วยประหยัดเวลาการดำเนินการโดยมนุษย์ลงไปได้ แต่มีค่าใช้จ่ายในการประมวลผลสูง เนื่องจากความซับซ้อนในการประมวลผลภาษาธรรมชาติ (natural language processing, NLP) วิธีการให้คำอธิบายเชิงความหมายแบบกึ่งอัตโนมัติจึงเป็นทางเลือกใหม่ที่นำข้อดีของทั้งสองวิธีมารวมกัน ปัจจุบันมีนักวิจัยได้พัฒนาเครื่องมือในการให้คำอธิบายข้อมูลเชิงความหมายแบบกึ่งอัตโนมัติสำหรับภาษาอังกฤษอยู่จำนวนหนึ่ง [5-7] โดยงานวิจัยเหล่านี้ใช้เครื่องมือทาง NLP หลากหลายประเภทที่รองรับภาษาอังกฤษ สำหรับเครื่องมือทาง NLP ภาษาไทยก็มีการพัฒนาอย่างต่อเนื่อง เช่น เครื่องมือตัดคำไทย ได้แก่ LongLexto[8] และ Tlex[9] ในส่วนการวิเคราะห์รูปประโยคต่าง ๆ มักนิยมใช้การโปรแกรมผ่าน API หรือ package เช่น PyThaiNLP [10] อย่างไรก็ตาม แม้มีเครื่องมือสนับสนุนที่มากขึ้น แต่การพัฒนาเครื่องมือให้คำอธิบายเชิงความหมายแบบกึ่งอัตโนมัติสำหรับภาษาไทยยังมีน้อยทั้งนี้เพราะต้องมีการผสานทั้งเทคโนโลยี NLP, text mining และ semantic web เข้าด้วยกัน นอกจากนี้ยังมีข้อจำกัดเกี่ยวกับความจำเพาะเจาะจงของโดเมนออนโทโลยีที่เกี่ยวข้องอีกด้วย

งานวิจัยนี้จึงได้พัฒนาระบบให้คำอธิบายเชิงความหมายแบบกึ่งอัตโนมัติสำหรับข้อมูลด้านศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นด้วยการผสมผสานเทคนิคที่กล่าวมาข้างต้น โดยระบบจะมีความสามารถระบุและนำเสนอประเภทนิพจน์ที่ใช้ในการอธิบายข้อมูลให้กับผู้ใช้ โดยประเภทนิพจน์เหล่านี้จะสอดคล้องกับออนโทโลยีศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นของจังหวัดพัทลุงที่ได้ออกแบบไว้ การแนะนำประเภทของนิพจน์จะใช้ทั้ง unsupervised และ supervised technique โดยจะมีการเก็บ log ของการเลือกประเภทนิพจน์จากผู้ใช้ แล้วนำมาสร้างโมเดลการจำแนก (classification model) เพื่อช่วยจำแนกคำอธิบายที่เหมาะสมกับบริบทและลดความกำกวมในการระบุนิพจน์ในครั้งต่อไป

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ข้อมูลเชิงความหมายในรูปแบบ RDF

ภาษา RDF เป็นภาษาที่อาศัยไวยากรณ์ของภาษา XML ใน RDF ข้อมูลจะแสดงในรูปแบบของ statement แต่ละ statement ประกอบด้วย 3 ส่วน คือ ประธาน (subject) ความสัมพันธ์ (predicate) และกรรม (object) [11] RDF statement สามารถแสดงในลักษณะกราฟ ซึ่งประธานและกรรมจะแทนด้วยโหนด ส่วนความสัมพันธ์จะแทนด้วยเส้น โหนดมี 2 ชนิด คือ resource และ literal โดย literal แทนค่าที่เป็นค่าคงที่ เช่น ตัวอักษร ตัวเลข ส่วน resource แทนด้วย URI โดย resource เป็นได้ทั้งประธานและกรรมของประโยค ภาษา RDF เอื้ออำนวยให้เกิดการเชื่อมโยงข้อมูลเชิงความหมายจากหลายแหล่งและการเปิดให้บริการข้อมูลกับแอปพลิเคชันต่าง ๆ ที่ต้องการได้

2.2 การให้คำอธิบายเชิงความหมาย

การทำงานของระบบให้คำอธิบายเชิงหมายนั้นขึ้นอยู่กับระดับความอัตโนมัติในการระบุประเภทนิพจน์และความสัมพันธ์ระหว่างนิพจน์ โดยทั่วไปนิยมใช้เทคนิคแบ่งเป็น 2 กลุ่ม คือ unsupervised และ supervised technique โดยเทคนิคแบบ supervised นิยมใช้ในระบบแบบกึ่งอัตโนมัติที่มีการเก็บผลลัพธ์จากการที่ผู้ใช้ให้คำอธิบายในอดีต เพื่อมาสอนให้ระบบสามารถจำแนกและแนะนำนิพจน์ที่เหมาะสม ส่วนเทคนิคแบบ unsupervised ไม่เกี่ยวข้องกับกระบวนการ machine learning ใด ๆ นิยมใช้ในระบบให้อธิบายแบบอัตโนมัติ ตัวอย่าง เช่น การรู้จำประเภทนิพจน์โดยอาศัยคำศัพท์จาก ontology การใช้ pattern

matching

ตัวอย่างระบบที่ใช้เทคนิคแบบ unsupervised คือ Ontotext [5] ระบบนี้จะเกี่ยวข้องกับการให้คำอธิบายเอกสารที่มีการเผยแพร่ผ่านสื่อ เช่น ข่าวจากโทรทัศน์ ภาพยนต์ และอื่น ๆ metadata ที่ใช้ในการอธิบายข้อมูลจะมาจากออนโทโลยี KIMO ที่มีการออกแบบมาเฉพาะโดเมน ต่อมาได้ขยายขีดความสามารถของ KIMO ในการใช้งานร่วมกับออนโทโลยีต่าง ๆ อีกมากมาย เทคนิคด้าน NLP ได้นำไปใช้เพื่อให้ Ontotext ระบุชนิดของคำในเอกสารได้อย่างถูกต้องสอดคล้องกับองค์ประกอบในออนโทโลยี สำหรับงานด้านศิลปวัฒนธรรม Uldis Bojars [6] และคณะ ได้นำเสนอเครื่องมือให้คำอธิบายเชิงความหมายแบบกึ่งอัตโนมัติเพื่อใช้ในหอสมุดแห่งชาติของประเทศลัตเวีย โดยระบบสนับสนุนการให้คำอธิบาย 3 ประเภท คือ การให้คำอธิบายพื้นฐาน การให้คำอธิบายเชิงโครงสร้าง และการให้คำอธิบายที่เกี่ยวโยงหลายเอกสาร ผู้ใช้ได้เข้ามามีส่วนร่วมในการเลือกคำอธิบายผ่านคลาสในออนโทโลยีที่มีการกำหนดไว้ล่วงหน้า โดยเอกสารด้านประวัติศาสตร์ที่ผ่านการให้คำอธิบายอยู่ในรูปแบบ JSON ส่วน Semantic Field Book Annotator [7] เป็นระบบให้คำอธิบายเอกสารหนังสือด้านประวัติศาสตร์ในรูปแบบรูปภาพจากไฟล์สแกน ระบบได้เตรียมกลุ่มคำอธิบายจากออนโทโลยี NHC และออนโทโลยีพื้นฐานต่าง ๆ สำหรับให้นักวัฒนธรรมอธิบายข้อมูลด้วยตนเอง ข้อมูลการให้คำอธิบายอยู่ในรูปแบบ RDF ส่วนระบบให้คำอธิบายแบบกึ่งอัตโนมัติที่นำเทคนิคแบบ supervised มาประยุกต์ใช้ ได้แก่ GoNTogle [12] ที่เตรียมเฟรมเวิร์คสำหรับให้คำอธิบายและสืบค้นข้อมูลโดยอาศัยออนโทโลยี GoNTogle ได้นำข้อมูลการให้คำอธิบายที่ผ่านมาจากผู้ใช้มาสร้างเป็น training data set สำหรับ classification model ที่ใช้อัลกอริทึม k-NN ทั้งนี้เพื่อแนะนำคำอธิบายที่เหมาะสมกับผู้ใช้ต่อไป Zemanta [13] เป็นระบบให้คำอธิบายที่ใช้ในการประมวลผลคำโดยระบบสามารถแนะนำ link เพื่อเชื่อมไปยังเนื้อหาต่าง ๆ บนเว็บ ไม่ว่าจะเป็น Tag ภาพ ชื่อบทความ โดยเนื้อหานี้ได้มาจากกระบวนการประมวลผลภาษาธรรมชาติที่ผสมผสานกับเทคนิค machine learning เพื่อการแนะนำเนื้อหาที่ใช้ในการให้คำอธิบายที่ถูกต้อง

ระบบการให้คำอธิบายเชิงความหมายในงานวิจัยนี้เป็นแบบกึ่งอัตโนมัติ เนื่องจากเอกสารข้อมูลด้านศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นมีความละเอียดอ่อนซับซ้อน การอธิบายจึงควรผ่านการตรวจสอบจากนักวัฒนธรรมร่วมด้วย ระบบมีส่วนการทำงานทั้งแบบส่วนอัตโนมัติที่อาศัยเทคนิคแบบ unsupervised และส่วนกึ่งอัตโนมัติที่อาศัยผู้ใช้ในการตัดสินใจเลือกคำอธิบาย โดยระบบจะนำเสนอคำแนะนำจากเทคนิค supervised ที่เกี่ยวข้องกับกระบวนการ machine learning

3. วิธีการดำเนินงาน

3.1 สถาปัตยกรรมภาพรวมของระบบต้นแบบให้คำอธิบายแบบกึ่งอัตโนมัติ

สถาปัตยกรรมระบบให้คำอธิบายเชิงความหมายแบบกึ่งอัตโนมัติสำหรับข้อมูลด้านศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นแสดงในรูปที่ 1 โดยระบบอ่านข้อมูลจากกลุ่มประโยค แล้วนำไปประมวลผลตามคอมโพเนนท์ต่าง ๆ ดังต่อไปนี้

3.1.1 การตัดคำและการแบ่งหน่วยประโยค (tokenization and sentence segmentation) คอมโพเนนท์นี้ทำหน้าที่วิเคราะห์กลุ่มประโยคแล้วหาขอบเขตนิพจน์ (token) โดยการเปรียบเทียบกับฐานคำศัพท์ในพจนานุกรม จากนั้นแบ่งช่วงกลุ่มนิพจน์ดังกล่าวให้อยู่ในรูปแบบหน่วยประโยคในลักษณะประโยคความเดียว เพื่อนำกลุ่มนิพจน์ในแต่ละประโยค

ส่งไปยังขั้นตอนการรู้จำนิพจน์

3.1.2 การรู้จำประเภทของนิพจน์ (named entity recognition) ทำหน้าที่จำแนกประเภทของนิพจน์ว่าเป็นนิพจน์ประเภทใดตาม class ที่มีการกำหนดไว้ในออนโทโลยีด้านศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นของจังหวัดพัทลุง [14] โดยนิพจน์ที่ได้รับการระบุประเภทจะเรียกว่านิพจน์ระบุนาม (named entity) จากนั้นส่งนิพจน์ระบุนามไปสกัดโครงสร้างความสัมพันธ์ของนิพจน์ อย่างไรก็ตาม หากขั้นตอนนี้ไม่สามารถระบุประเภทของนิพจน์ เนื่องจากความกำกวมของคำและบริบท ก็จะส่งนิพจน์นั้นไปยังคอมโพเนนท์ขจัดความกำกวมของนิพจน์

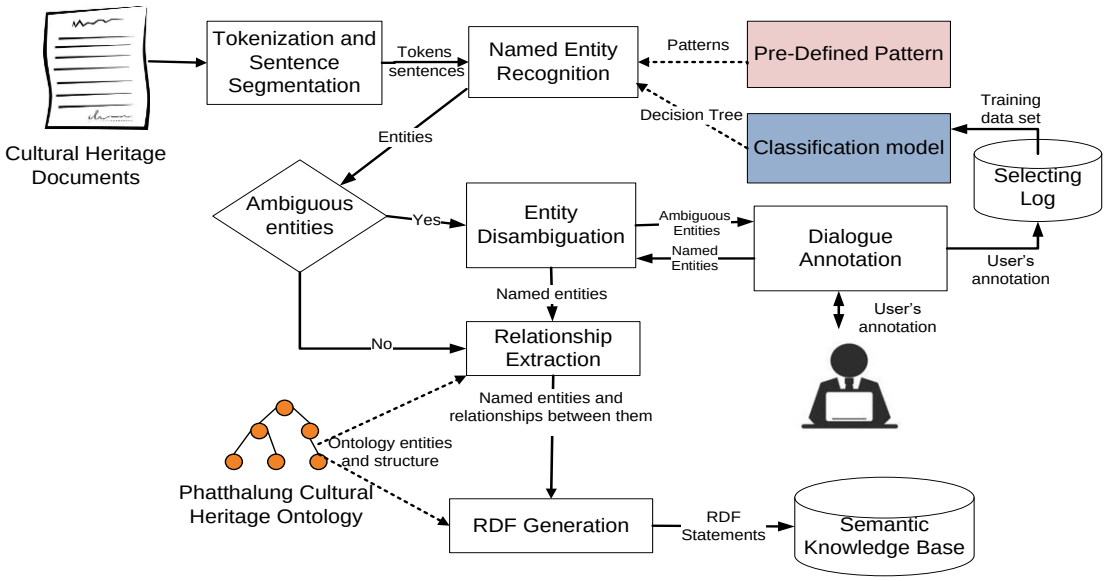


Figure 1 Architecture of the semi-automatic semantic annotation system

3.1.3 การขจัดความกำกวมของนิพจน์ (entity disambiguation) จะเกี่ยวข้องกับการให้ผู้ใช้ตัดสินใจเลือกประเภทของนิพจน์ด้วยตนเอง โดยระบบจะนำเสนอกลุ่มของ class จากออนโทโลยีที่เป็นไปได้เพื่อให้ผู้ใช้ตัดสินใจเลือกประเภทของนิพจน์ ผลลัพธ์จากการเลือกนำไปเก็บใน log ของการเลือก (selecting log) เพื่อนำไปเป็น training data set สำหรับสร้างโมเดลการจำแนกต่อไป ทั้งนี้เพื่อในครั้งหน้าระบบสามารถระบุและแนะนำ class ที่เหมาะสมและถูกต้องแก่ผู้ใช้

3.1.4 การสกัดความสัมพันธ์ระหว่างนิพจน์ (relationship extraction) คอมโพเนนท์นี้นำกลุ่มของนิพจน์ระบุนามในประโยคมาเปรียบเทียบกับโครงสร้างของออนโทโลยีแล้วหาความสัมพันธ์ระหว่างนิพจน์ที่เหมาะสมเพื่อสร้างความสัมพันธ์ให้อยู่ในรูป statement

3.1.5 การสร้างข้อมูล RDF (RDF statement generation) คอมโพเนนท์นี้จะนำ statement ที่ได้มาสร้างเป็น RDF statement ผลลัพธ์ที่ได้ คือ ชุดข้อมูลในรูปแบบ RDF และส่งไปจัดเก็บในฐานความรู้เชิงความหมาย

3.2 การตัดคำและการแบ่งหน่วยประโยค

คอมโพเนนท์นี้อาศัยหลักการตัดคำโดยใช้พจนานุกรม เนื่องจากเป็นวิธีที่ประมวลผลได้รวดเร็ว

มีความความถูกต้องสูง ทั้งนี้ขึ้นอยู่กับขนาดของคำศัพท์ในพจนานุกรมที่ใช้ ผู้วิจัยได้ใช้เครื่องมือตัดคำ LongLexto [8] ซึ่งอาศัยเทคนิคการเลือกคำที่ยาวที่สุดโดยอ้างอิงจากพจนานุกรม โดยผู้วิจัยได้เพิ่มและแก้ไขคำศัพท์ในกลุ่มศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นภาคใต้เข้าไปในพจนานุกรมด้วย ซึ่งจากประโยคตัวอย่าง คือ “โนราเป็นศิลปะการแสดงพื้นบ้านที่นิยมของคนในภาคใต้ โดยองค์ประกอบหลักในการแสดงโนรา คือ เครื่องแต่งกายและเครื่องดนตรี” สามารถตัดคำและแบ่งประโยคดังนี้

การตัดคำ
โนรา/เป็น/ศิลปะการแสดง/พื้นบ้าน/ที่/นิยม/ของ/คน/ใน/ภาคใต้/ /โดย/องค์ประกอบ/หลัก/ใน/การแสดง/โนรา/ /คือ/เครื่องแต่งกาย/และ/เครื่องดนตรี
การแบ่งหน่วยประโยค
ประโยคที่ 1: โนราเป็นศิลปะการแสดงพื้นบ้านที่นิยมของคนในภาคใต้
ประโยคที่ 2: โดยองค์ประกอบหลักในการแสดงโนราคือเครื่องแต่งกายและเครื่องดนตรี

การแบ่งหน่วยประโยคเป็นการแบ่งช่วงกลุ่มนิพจน์ให้อยู่ในรูปแบบประโยคความเดียวที่ประกอบด้วยนิพจน์ที่ทำหน้าที่เป็นประธาน กริยา และกรรม โดยวิธีการแบ่งหน่วยประโยคจะแบ่งตามการปรากฏของนิพจน์ระบุขอบเขต (discourse segment cue) [15] ตัวอย่างนิพจน์ระบุขอบเขตขึ้นต้นประโยค ได้แก่ ขณะที่ ทั้งนี้ เนื่องจาก โดย เป็นต้น ส่วนนิพจน์ระบุขอบเขตลงท้ายประโยค ได้แก่ ก็ตาม นัก เป็นต้น

3.3 การรู้จำประเภทของนิพจน์

การรู้จำประเภทของนิพจน์เป็นการกำหนดและระบุประเภทของนิพจน์ ซึ่งสอดคล้องกับ ontology entity ได้แก่ นิพจน์อาจเป็นประเภทบุคคล ประเพณี การแสดงพื้นบ้าน เป็นต้น สำหรับเทคนิคการรู้จำประเภท unsupervised ผู้วิจัยจะใช้เทคนิครูปแบบที่กำหนด (pre-defined pattern) ร่วมกับ vector space model ในการเทียบความคล้ายคลึงระหว่างนิพจน์กับ ontology entity และกับนิยามจากพจนานุกรม ส่วนเทคนิค supervised เกี่ยวพันกับการสร้าง classification model เพื่อจำแนกประเภทของนิพจน์ที่กำกวมได้ถูกต้องเมื่อพบในบริบทที่ใกล้เคียงกับ training data set ในการทำงานครั้งต่อไป

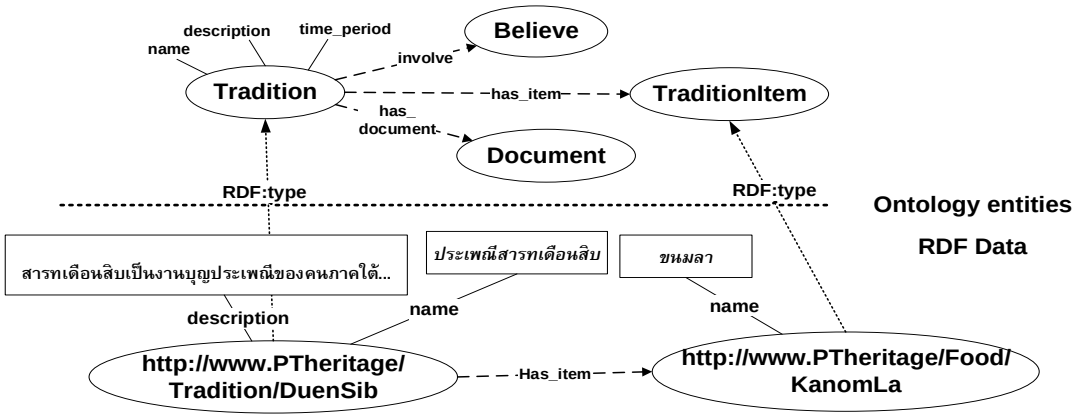
3.3.1 เทคนิครูปแบบที่กำหนด (pre-defined pattern) การรู้จำโดยอาศัยรูปแบบที่กำหนดเป็นวิธีการเพื่อให้คอมพิวเตอร์สามารถแยกแยะประเภทของนิพจน์โดยเน้นที่การกำหนดรูปแบบ (pattern) ของนิพจน์ที่ปรากฏอยู่ในประโยค เช่น บุคคลประกอบด้วยคุณสมบัติคำนำหน้า ชื่อ และนามสกุล โดยตัวอย่างการกำหนดรูปแบบเพื่อระบุประเภทนิพจน์สำหรับข้อมูลบุคคลและประเพณี ดังนี้

- การกำหนดรูปแบบการรู้จำประเภทบุคคล
  Pattern = [(‘คำนำหน้า’) (‘ชื่อใด ๆ’)( ) (‘นามสกุลใด ๆ’)]
  คำนำหน้า ::= (‘นาย’ | ‘นาง’ | ‘นางสาว’|) หมายถึง คำนำหน้าชื่อตามสำเนาทะเบียนบ้าน
  ชื่อ ::= (‘ชื่อใด ๆ’) หมายถึง ชื่อของบุคคล
  นามสกุล ::= (‘นามสกุลใด ๆ’) หมายถึง นามสกุลของบุคคล
- การกำหนดรูปแบบการรู้จำประเภทประเพณี (Tradition)
  Pattern = [(‘ประเพณี’) (‘ชื่อใด ๆ’)]
  คำนำหน้า ::= (‘ประเพณี’) หมายถึง นิพจน์ขึ้นต้นบ่งชี้ถึงประเพณี
  ชื่อประเพณี ::= (‘ชื่อใด ๆ’) หมายถึง ชื่อของประเพณี

รูปแบบการรู้จำที่ได้กล่าวมาข้างต้น เมื่อมีสายอักขระเข้ามาในระบบคอมพิวเตอร์ทำให้ระบบสามารถระบุประเภทนิพจน์ เช่น ข้อมูลที่รับจากขั้นตอนการตัดคำ คือ “| ประเพณี | สารทเดือนสิบ | จัด | ขึ้น | ใน |วัน | แรม | ๑ | ค่ำ | และ | แรม | ๑๕ | ค่ำ| เดือน | ๑๐ | ของ | ทุก |ปี |” สามารถระบุประเภทนามแก่กลุ่มนิพจน์ตามรูปแบบตามการรู้จำประเภทประเพณีได้ดังนี้ (ประเพณี) ::= เป็นการระบุว่ากลุ่มนิพจน์คือประเภทประเพณี (สารทเดือนสิบ) ::= เป็นชื่อของประเพณี

3.3.2 เทคนิคการเทียบความคล้ายคลึงระหว่างนิพจน์กับ Ontology Entity และนิยามจากพจนานุกรม เนื่องจากออนโทโลยีประกอบด้วย entity ต่าง ๆ เช่น class, property, instance ที่อยู่ในรูปแบบลำดับโครงสร้างที่ชัดเจน โดย literal จะเป็นค่าสมบัติ (data property) ของ instance ใด ๆ โดยหาก literal จาก instance ใดมีความคล้ายคลึงกับนิพจน์สูง

ก็จะตั้งสมมุติฐานว่านิพจน์นั้นน่าจะอยู่ใน class เดียวกับ instance ดังกล่าว การคำนวณความคล้ายคลึงจะอาศัย vector space model โดยหาค่าน้ำหนักของคำใน literal ของ instance เพื่อนำมาสร้าง index ของเอกสารสำหรับแต่ละ instance โดย index ของแต่ละเอกสารถูกพิจารณาเป็น vector เพื่อนำมาหาความคล้ายคลึงกับ vector ของนิพจน์โดยใช้ cosine similarity ในสมการที่ 2 ส่วนรูปที่ 2 แสดงตัวอย่าง literal ของ instance ต่าง ๆ ในออนโทโลยี และการสกัด literal ผ่านการใช้ SPARQL query เพื่อนำมาสร้าง index




```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
SELECT DISTINCT ?x ?p ?y ?c WHERE
        {?p rdf:type owl:DatatypeProperty.
    ?x ?p ?y.
    ?x rdf:type ?c. }
```

Figure 2 Literal extraction using SPARQL query

ตัวอย่างในรูปที่ 2 เอกสารสำหรับ instance ที่ชื่อ สารทเดือนสิบ ประกอบด้วยฟิลด์ต่อไปนี้

(1) Entity type field = “instance”

(2) ตัวแปร ?x แทน Entity URI = http://www.PTheritage/Tradition/DuenSib

(3) ตัวแปร ?c แทน Class URI = http://www.PTheritage/Tradition

(4) ตัวแปร ?p แทน Property URI = http://www.PTheritage/description

(5) ตัวแปร ?y แทน Literal = “สารทเดือนสิบเป็นงานบุญประเพณีของคนภาคใต้ของประเทศไทยโดยมีจุดมุ่งหมายสำคัญเพื่อเป็นการอุทิศส่วนกุศลให้แก่ดวงวิญญาณของบรรพชนและญาติที่ล่วงลับ”

vector space model มีการกำหนดให้ $w_{i,j}$ เป็นน้ำหนักของคำในเอกสารที่มีความสัมพันธ์กันเป็นคู่อันดับ (k_i,d_j) โดย k คือ คำลำดับที่ i ที่ปรากฏในเอกสารของ instance (d) ที่ j โดยค่าน้ำหนักต้องมีค่ามากกว่าหรือเท่ากับ 0 ($w_{i,j} >= 0$) เวกเตอร์สำหรับ

เอกสาร d_j เขียนแทนด้วย $\overrightarrow{dj}$ = ($w_{1,j}$, $w_{2,j}$, ..., $w_{i,j}$) โดยค่าน้ำหนักของคำหาได้จาก

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_i} \quad (1)$$

โดยที่ w คือ น้ำหนักของคำ (term weight); $tf_{i,j}$ คือ ความถี่ของคำ (term frequency) ลำดับที่ *i* ที่ปรากฏในเอกสารของ instance (d) ที่ *j*; df_i คือ จำนวนเอกสาร instance ที่มีคำที่ i ปรากฏ

การหาความคล้ายคลึงกันของเอกสาร instance (d_j) และนิพจน์ (d_v) หาโดยใช้ค่าความคล้ายคลึงแบบโคไซน์ (cosine similarity) ของเวกเตอร์สองเวกเตอร์ คือ $\overrightarrow{d_j}$ (เวกเตอร์ของเอกสาร Instance d_j) และ $\overrightarrow{d_v}$ (เวกเตอร์ของนิพจน์) คำนวณได้ดังต่อไปนี้

$$Sim\left(d_j, d_v\right) = \frac{\overrightarrow{d_j} \cdot \overrightarrow{d_v}}{|\overrightarrow{d_j}| \times |\overrightarrow{d_v}|} \quad (2)$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,v}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,v}^2}}$$

โดย t คือ จำนวนคำทั้งหมดในระบบ

ผลที่ได้จากค้นคือเอกสารที่เกี่ยวข้องกับนิพจน์และคะแนนความคล้ายของแต่ละเอกสารของ instance โดย class ของเอกสารจาก instance ที่มีคะแนนมากที่สุดจะกำหนดให้เป็นประเภทของนิพจน์ ซึ่งหากปรากฏชุดของ instance entity ที่เป็นไปได้หลายค่า เราจะเปรียบเทียบคะแนน เพื่อเลือกเฉพาะกลุ่ม entity ที่มีค่ามากกว่า threshold ที่กำหนดเพื่อนำเสนอ class ที่เป็นไปได้ให้ผู้ใช้เลือกผ่าน user tags และจะนำไปเก็บใน log สำหรับสร้าง training set ต่อไป

อย่างไรก็ตาม หากไม่สามารถเทียบนิพจน์ระบุนามกับ instance entity ใด ก็อาศัยการคำนวณความคล้ายคลึงระหว่างเวกเตอร์นิยามของนิพจน์ที่สืบค้นจากพจนานุกรมกับเวกเตอร์ instance entity หรือเวกเตอร์ของ class ใด ๆ ในออนโทโลยีอีกครั้ง ตัวอย่าง เช่น นิพจน์ “สงกรานต์” จากพจนานุกรม Lexitron ของ NECTEC ได้ให้นิยามว่า คือ “เทศกาลเนื่องในการขึ้นปีใหม่อย่างเก่า ซึ่งกำหนดตามสุริยคติตกวันที่ 13 - 14 - 15 เมษายน” ซึ่งนิยามนี้ได้นำมาสร้างเป็นนำมาสร้างเวกเตอร์แล้วนำมาเปรียบเทียบความคล้ายกับเวกเตอร์ของ instance และ class ใด ๆ ซึ่งผลจากการคำนวณพบว่านิยามของนิพจน์ “สงกรานต์” มีคล้ายคลึงกับ class:Festival = 0.258 และไม่พบความคล้ายคลึงกับ entity อื่นจึงสามารถระบุ named entity ของนิพจน์ “สงกรานต์” ว่าเป็นเทศกาล

3.4 การจำแนกประเภทนิพจน์ที่มีความกำกวมด้วยเทคนิคเหมืองข้อมูล

กรณีที่เกิดความกำกวมขึ้นเนื่องจากระบบระบุประเภทของนิพจน์ได้มากกว่า 1 ประเภท ระบบจะเตรียม annotation combo box เพื่อนำเสนอ class ที่เป็นไปได้ เพื่อให้ผู้ใช้ตัดสินใจเลือกประเภทของนิพจน์ได้ หรือถ้า class ที่ระบบเสนอให้ไม่ถูกต้อง ผู้ใช้สามารถให้คำอธิบายได้ด้วยตนเอง โดย log ของการเลือกถูกนำไปเป็น training data set สำหรับสร้างโมเดลในส่วนเครื่องจักรการเรียนรู้ การดำเนินการสร้างโมเดลประกอบด้วยขั้นตอน 3 ส่วนหลัก คือ (1) การคัดเลือกคุณลักษณะของคำ (2) การสร้างโมเดลการจำแนก และ (3) การทดสอบโมเดล

3.4.1 การคัดเลือกคุณลักษณะคำ เริ่มจากการนำข้อมูลประโยคและผลการเลือกโดยผู้ใช้จาก log ไฟล์ มาเตรียมข้อมูลให้อยู่ในรูปแบบพร้อมประมวลผล กล่าวคือ ตัดคำ จากนั้นลบคำที่ไม่เกี่ยวข้อง หรือการลบคำที่ไม่มีประโยชน์ในการนำไปวิเคราะห์ออก ซึ่งพบว่าแอตทริบิวต์ (attribute) หรือคำในประโยคมีจำนวนมาก คำเหล่านี้บางคำก็ไม่ได้มีความสำคัญในการแบ่งแยกประเภทหรือ class ดังนั้นจึงจำเป็นต้องคัดเลือกคุณลักษณะเฉพาะคำที่สำคัญมาใช้ในการ

จำแนก ในที่นี้ใช้ TF-IDF ในการหาค่าความถี่และถ่วงน้ำหนักคำ โดยการถ่วงน้ำหนักคำนี้เป็นการวัดเชิงสถิติที่ใช้เพื่อประเมินความสำคัญของคำต่าง ๆ ที่ปรากฏอยู่บนเอกสาร การคัดเลือกคุณลักษณะ (feature selection) จะเลือกคำที่มีค่า TF-IDF ที่มีค่ามากกว่าค่า tread hold ที่กำหนด ทั้งนี้เพื่อตัดคุณลักษณะคำที่ใช้น้อย ซึ่งเป็นการลดจำนวนโครงสร้าง (Schema reduction) ของ training set งานวิจัยนี้แบ่ง class ข้อมูลที่พบความกำกวมจำนวน 6 class ได้แก่ อาหาร ภูมิปัญญา พิธีกรรม การแสดง ความเชื่อ และศาสนา โดยแบ่งเป็น class ละ 25 ข้อความ

3.4.2 การสร้างโมเดล classification แบ่งข้อมูลทดสอบโดยใช้วิธีการประเมินผล K-fold cross validation โดย k=5 สำหรับการทดสอบจะใช้โปรแกรม Weka ในที่นี้ใช้เทคนิค Naïve Bayes และ decision tree [16] ในการจำแนก

3.4.3 การทดสอบ หลังจากที่ได้สร้าง classification model ต่อมาเราก็จะหาค่าเฉลี่ยการวัดประสิทธิภาพของแต่ละเทคนิค การประเมินจะวัดจากค่า precision, recall และค่า accuracy ผลการทดสอบพบว่าโมเดลที่ใช้อัลกอริทึม decision tree ให้ค่าเฉลี่ย accuracy 85.7 % ค่า precision 0.9 และค่า recall 0.85 ซึ่งสูงกว่าค่าประสิทธิภาพของเทคนิค Naïve Bayes ดังนั้นจึงเลือกใช้อัลกอริทึม decision tree ในการสร้างโมเดลจำแนกประเภทนิพจน์ที่มีความกำกวม และนำโมเดลมาพัฒนาเป็นส่วนหนึ่งของระบบในการแนะนำประเภทนิพจน์แก่ผู้ใช้

3.5 การสกัดความสัมพันธ์ระหว่างนิพจน์และการสร้างข้อมูล RDF

ออนโทโลยีสามารถสกัดโครงสร้างเป็น statement ซึ่งประกอบด้วย 3 ส่วนหลัก คือ domain, property และrange โดยเปรียบเทียบโครงสร้างดังกล่าวกับนิพจน์และคำต่าง ๆ ที่ปรากฏในประโยค การสกัดความสัมพันธ์เบื้องต้นจะอาศัยการค้นหาคำกริยาในหน่วยประโยคที่สอดคล้องกับ object property ตัวอย่าง เช่น ประโยค “ประเพณีสารทเดือนสิบได้รับอิทธิพลด้านความเชื่อจากทางศาสนาพราหมณ์” อาจสามารถระบุนามนิพจน์จากองค์ประกอบการรู้จำนิพจน์ดังตัวอย่าง เช่น นิพจน์ “สารทเดือนสิบ” ได้รับการระบุนามเป็นข้อมูลประเภทชื่อประเพณี (Tradition) นิพจน์ “พราหมณ์” ได้รับการระบุนามเป็นชื่อศาสนา (Religion) ส่วนนิพจน์ “ได้รับอิทธิพล” ได้ระบุเป็น object property:Involve ที่เชื่อมระหว่าง instance ของคลาส “#Tradition” และ instance ของคลาส “#Religion” ซึ่ง object property ดังกล่าวจะถูกนำเสนอให้ผู้ใช้ตรวจสอบความถูกต้องอีกครั้งก่อน หลังจากผู้ใช้ยืนยันจะนำผลลัพธ์ดังแสดงในตารางที่ 1 มาดำเนินการสร้างให้อยู่ในรูปแบบ RDF statement โดยใช้การโปรแกรมด้วยเครื่องมือหลัก คือ Jena RDF API ผลลัพธ์สุดท้ายจะอยู่ใน RDF File เพื่อจัดเก็บเป็นฐานความรู้เชิงความหมายที่สามารถนำไปใช้ประโยชน์ได้ตามวัตถุประสงค์ต่อไป ดังตัวอย่างข้อมูล RDF ในรูปที่ 3

3.6 ส่วนการติดต่อกับผู้ใช้

ระบบให้คำอธิบายเชิงความหมายแบบกึ่งอัตโนมัติสำหรับข้อมูลด้านศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นมีรูปแบบหน้าจอ ดังแสดงในรูปที่ 3

โดยหน้าจอแบ่งเป็น 3 ส่วน คือ (1) ส่วนรับข้อมูลประโยค (sentence) เป็นส่วน input ให้ผู้ใช้พิมพ์ประโยคที่ต้องการให้คำอธิบายสู่ระบบ (2) ส่วนการให้คำอธิบายนิพจน์ (entity annotation) จะสกัดกลุ่มนิพจน์จากประโยค และระบบแนะนำ system annotation tag โดยอัตโนมัติ ซึ่งหากผู้ใช้ไม่เห็นด้วยกับนิพจน์ที่ระบบแนะนำ ผู้ใช้สามารถเลือก user tag ด้วยตนเองจาก combo box จากนั้นกดปุ่ม edit ผลจากการเลือกของผู้ใช้จะเก็บไว้ใน log (3) property

Table 1 Named entities corresponded to ontology structure

Subjects	Predicate	Objects
#Tradition/DuenSib	#name	“Sarth duensib”
#Tradition/DuenSib	rdf:type	#Tradition
#Religion/Hinduism	#name	“Brahmin”
#Religion/Hinduism	rdf:type	#Religion
#Tradition/DuenSib	#involve	#Religion/Hinduism

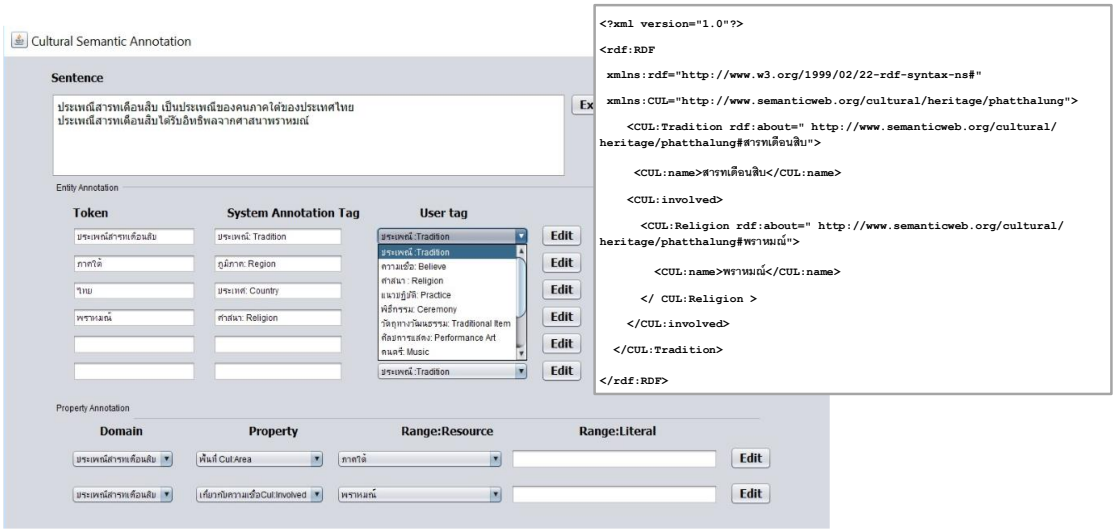


Figure 3 User interface and output in RDF format

annotation จะเป็นส่วนให้คำอธิบายสมบัติของนิพจน์ผ่าน data property และอธิบายความสัมพันธ์ระหว่างนิพจน์ผ่าน object property ผู้ใช้สามารถเลือกตามที่ระบบแนะนำ หรือสามารถเลือกด้วยตนเองเช่นกัน หลังจากผู้ใช้ยืนยันการตรวจสอบการคำอธิบายเรียบร้อย ระบบจะสร้างไฟล์ RDF แล้วจัดเก็บไว้ใน folder ตามที่ระบุ

4. การประเมินประสิทธิผลการให้คำอธิบายเชิงความหมาย

วัตถุประสงค์ของการประเมินเพื่อวัดประสิทธิผลการแนะนำนิพจน์ระบุนามและความสัมพันธ์ระหว่างนิพจน์ให้กับผู้ใช้ ซึ่งคำแนะนำนี้จะช่วยอำนวยความสะดวกแก่ผู้ใช้ในการให้คำอธิบายและเพิ่มดีกรีความอัตโนมัติให้กับระบบ โดยเกณฑ์ที่ใช้ในการวัดประสิทธิผลการให้คำอธิบาย คือ ค่า precision และค่า recall โดย precision = $\frac{TP}{TP+FP}$ และ recall = $\frac{TP}{TP+FN}$ ในที่นี้ TP คือ จำนวนนิพจน์ที่ถูกระบุนามถูกต้อง FP คือ จำนวนนิพจน์ที่ถูกระบุนามไม่ถูกต้อง และ FN คือ จำนวนนิพจน์ที่ไม่ถูกระบุนาม

Table 2 Named entity recognition results

Classes	Precision	Recall
Date	100	100
Religion	100	100
ceremony	80.9	65.25
Location	100	100
Festival	100	100
Costume	75.33	68.50
Traditional Item	82.75	78.52
Musical Instrument	78.60	78.60
Performance Art	82.23	70.66
Person	87.50	72.34
Practice	70.00	63.25
Tradition	100	100
Production	80.35	67.50
Wisdom	80.25	73.50
Food	93.50	93.50
Average	88.07	82.10

การทดลองได้นำเข้าข้อมูลเนื้อหาบทความที่เกี่ยวข้องกับศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่นของจังหวัดพัทลุง จำนวน 100 paragraph โดยมีเนื้อหาสอดคล้องกับ class ที่ใช้ทดสอบในตารางที่ 2 การทดลองอาศัยข้อมูล 2 กรณี คือ (1) ชุดข้อมูลจริง : กลุ่มข้อมูลที่ปรากฏจริงในเอกสารที่เป็นภาษาธรรมชาติที่มีโครงสร้างทางภาษาที่หลากหลาย (2) ชุดข้อมูลผ่านการทำความสะอาด : ข้อมูลจริงที่ผู้วิจัยเพิ่มส่วนโครงสร้างทางภาษาที่สูญหาย เช่น เพิ่มส่วนประธานหรือกรรมของประโยค ซึ่งจะทดสอบโดยนำเข้าข้อมูลทั้ง 2 ชุด โดยในแต่ละชุดข้อมูลระบบจะบันทึกการให้คำอธิบายที่ระบบแนะนำ นอกจากนี้ระบบจะบันทึกการให้คำอธิบายของผู้ใช้ด้วยตนเองในกรณีที่ระบบให้คำแนะนำนิพจน์หรือความสัมพันธ์ไม่สอดคล้องกับเนื้อหาหรือไม่สามารถให้คำอธิบาย จากนั้นคำนวณค่า precision และค่า recall ของนิพจน์และความสัมพันธ์แต่ละประเภทพร้อมหาค่าเฉลี่ย ซึ่งผลการทดลองแสดงในตารางที่ 2 และ 3

ตารางที่ 2 ผลการระบุนามนิพจน์โดยระบบพบว่ามีค่า precision เฉลี่ย 88.07 % หมายความว่าโปรแกรมสามารถระบุนามนิพจน์แม่นยำในระดับดี เนื่องจากการทำงานผสมผสานกันของเทคนิคต่าง ๆ ดังกล่าวข้างต้น ส่วนค่า recall 82.10 % กล่าวคือโปรแกรมสามารถระบุนามนิพจน์ครอบคลุมครบถ้วนในระดับดีเช่นกัน แต่ก็มีนิพจน์บางส่วนที่ไม่สามารถระบุนิพจน์ในการตัดคำ เนื่องจากนิพจน์ของคำศัพท์ทางวัฒนธรรมนั้นส่วนใหญ่จะเป็นคำศัพท์เฉพาะที่ไม่ปรากฏในพจนานุกรม การอาศัยฐานข้อมูลวัฒนธรรมที่ผู้วิจัยรวบรวมได้เพียงอย่างเดียว อาจไม่ครอบคลุมในทุกด้าน ซึ่งอาจแก้ข้อจำกัดดังกล่าวโดยการเข้าถึงฐานข้อมูลหรือคำศัพท์ทางวัฒนธรรมในด้านต่าง ๆ ที่หลากหลายมากขึ้น

ผลการทดลองในตารางที่ 3 นั้น พบว่าค่า precision ก่อนการทำความสะอาดข้อมูลมีค่า 60.32 % และค่า recall 46.96 % หมายความว่าโปรแกรมสามารถระบุความสัมพันธ์ได้ถูกต้องและครบถ้วนในระดับต่ำ ทั้งนี้เนื่องจากวิธีการในการระบุความสัมพันธ์ระหว่างนิพจน์ใช้การค้นหาคำกริยาที่สัมพันธ์กับนิพจน์ที่พบในประโยค จึงพบข้อจำกัดที่เกิดขึ้นจากคุณลักษณะทางภาษาธรรมชาติ ได้แก่ การอ้างอิงร่วมแบบสูญรูป การละข้อความ การใช้ประโยคความรวมและการใช้ประโยคความซ้อน ทำให้การระบุความสัมพันธ์ระหว่างนิพจน์ในการทดลองนี้อยู่ในระดับต่ำ ซึ่งเมื่อทำความสะอาดข้อมูลด้วยการเติมนิพจน์ที่สูญหายจากรูปประโยคดังกล่าว พบว่าประสิทธิผลของการแนะนำความสัมพันธ์ระหว่างนิพจน์สูงขึ้น โดยมีค่า

Table 3 Relationship extraction results

Object properties	Domains	Ranges	Raw Data		Cleaned Data	
			Precision	Recall	Precision	Recall
has_document	PerformanceArt	Document	80.50	44.00	85.25	72.00
has_legend	PerformanceArt	Legend	32.50	35.71	50.42	71.14
event_in	Event	Tradition	50.50	50.00	70.25	70.00
involve	Tradition	Believe	72.25	66.66	80.5	66.66
has_ceremony	PerformanceArt	Ceremony	75.25	25.00	78.25	70.83
consist_of	PerformanceArt	Person	35.50	41.37	77.25	65.51
has_item	PerformanceArt	Item	45.50	65.50	60.23	70.00
has_item	Tradition	TraditionItem	75.5	50.00	78.5	100
has_practice	Tradition	Practice	80.5	42.85	90.5	80.00
event_in	Event	Tradition	70.25	60.00	94.11	83.33
process	Production	Item	45.33	35.53	60.25	72.00
Average			60.32	46.96	75.04	74.68

precision 75.04 % และค่า recall 74.68 % หมายความว่าโปรแกรมสามารถระบุนามนิพจน์ได้แม่นยำและครบถ้วนในระดับปานกลาง จะเห็นว่าประสิทธิผลการสกัดความสัมพันธ์ระหว่างนิพจน์ในรูปประโยคดีขึ้นทั้งนี้เพราะหน่วยประโยค ต่าง ๆ มีความครบถ้วนสมบูรณ์มากขึ้น ดังนั้นการสกัดโครงสร้างประโยคที่ถูกต้องเพื่อตรวจสอบ องค์ ประกอบและความสัมพันธ์กันขององค์ประกอบต่าง ๆ ในประโยคจากองค์ประกอบระดับล่างไปสู่องค์ ประกอบในระดับสูงของไวยากรณ์ภาษาไทยในรูปแบบ tree จึงเป็นสิ่งจำเป็นสำหรับการจัดการกับประโยคความซ้อน เมื่อเราสามารถเข้าใจโครงสร้างก็สามารถเพิ่มนิพจน์ที่ละไว้ในประโยคประเภทนี้ให้สมบูรณ์ ซึ่งน่าจะส่งผลให้การระบุความสัมพันธ์ระหว่างนิพจน์ในประโยคมีความถูกต้องเพิ่มมากขึ้น

5. สรุป

งานวิจัยนี้ได้พัฒนาระบบให้คำอธิบายเชิงความหมายแบบกึ่งอัตโนมัติสำหรับข้อมูลด้านศิลปวัฒนธรรมและภูมิปัญญาท้องถิ่น ซึ่งระบบสามารถแนะนำ metadata ที่สอดคล้องกับออนโทโลยีให้กับผู้ใช้ และยังเปิดโอกาสให้ผู้ใช้เข้ามามีส่วนร่วมในการให้คำอธิบายกรณีเกิดความกำกวมของการระบุประเภทของนิพจน์ งานวิจัยนี้ใช้เทคนิคการรู้จำประเภทนิพจน์ร่วมแบบ unsupervised techniques และ supervised technique กรณีที่นิพจน์มีความกำกวมระบบจะมีการเรียนรู้โดยนำ log ของการเลือกโดยผู้ใช้ไปเป็น training data set สำหรับสร้างโมเดลในส่วนเครื่องจักรการเรียนรู้ เพื่อที่หากพบนิพจน์ดังกล่าวในบริบทที่คล้ายคลึงกัน ระบบจะสามารถระบุนิพจน์ได้อย่างถูกต้อง การประเมินพบว่าระบบระบุนามนิพจน์ได้แม่นยำและครบถ้วนในระดับดี ส่วนการระบุความสัมพันธ์ระหว่างนิพจน์ในประโยคพบว่ามีประสิทธิภาพต่ำ ทั้งนี้เนื่องจากข้อจำกัดของการ

ประมวลผลรูปประโยคภาษาธรรมชาติ จึงเป็นจุดที่ต้องอาศัยผู้ใช้ในการตรวจสอบและระบุความสัมพันธ์ระหว่างนิพจน์ ดังนั้นจึงมีแนวคิดในการศึกษาเทคนิคและเครื่องมือด้านการประมวลผลภาษาธรรมชาติโดยเฉพาะการวิเคราะห์รูปประโยคและโครงสร้างทางภาษาเพื่อการสกัดหาความสัมพันธ์ระหว่างนิพจน์ให้มีประสิทธิภาพมากยิ่งขึ้นในอนาคต

6. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการสนับสนุนทุนวิจัยจากงบประมาณแผ่นดิน มหาวิทยาลัยทักษิณ ปีงบประมาณ 2560 สัญญาเลขที่ R01-2560A10501071

7. References

[1] Kulgun, H., 2013, Knowledge management on local culture of Tambol Aomkred, Pakkred district, Nontabury province, J. Cult. Approach 14(25): 18-30. (in Thai).

[2] Naco, A., 2017, Knowledge-based system of art and culture in the context of folk wisdom case study: Phatthalung province, Thaksin J. 20(3): 292-299. (in Thai)

[3] Fine Art Department, Ministry of Culture, Cultural Digital Data Warehouse, Available Source: http://www.digitalcenter.finearts.go.th/home, January 3, 2019. (in Thai)

[4] Antoniou, G., Groth, P., Harmelen, F. and Hoekstra, R., 2012, A Semantic Web Primer, 3rd Ed., Massachusetts Institue of Technology, Cambridge, MA., 287 p.

[5] Ontotext, Available Source: http://www.ontotext.com, May 12,2014.

[6] Uldis, B., Rasmane, A., Zogla, A., Balina, S. and Salna, E., 2018, Semantic annotation tool for cultural heritage content, Baltic J. Mod. Comput. 6: 449-463.

[7] Stork, L., Weber, A., Miracle, E.G., Verbeek, F., Plaat, A., Herik, J. and Wolstencroft, K., 2018, Semantic annotation of natural history collections, Web Semantics: Science, Services and Agents on the World Wide Web. (in press)

[8] Haruechaiyasak, C., Lexto Thai Lexeme Tokenizer, Available Source: http://www.sansarn.com /lexto, April 7, 2018.

[9] Haruechaiyasak, C. and Kongyoung, S., 2009, TLex: Thai lexeme analyzer based on the conditional random fields, In International Symposium on Natural Language Processing.

[10] PyThaiNLP, Thai Natural Language Processing in Python, Available Source: https://pythainlp.readthedocs.io/en/latest, January 18, 2019.

[11] Schreiber, G. and Raimond, Y., 2014, RDF 1.1 Primer, Available Source: https://www.w3.org/TR/rdf11-primer, January 10, 2018.

[12] Giannopoulos, G., Bikakis, N., Dalamagas, T. and Sellis T., 2010, GoNTogle: A Tool for Semantic Annotation and Search, In Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L. and Tudorache, T. (Eds.), The Semantic Web: Research and Applications, ESWC 2010, Lecture Notes in Computer Science, Vol. 6089, Springer, Berlin.

[13] Bontcheva, K., Cunningham, H., 2011, Semantic Annotations and Retrieval: Manual, Semiautomatic, and Automatic Generation, In Domingue, J., Fensel, D. and Hendler, J.A. (Eds.), Handbook of Semantic Web Technologies, Springer, Berlin.

[14] Sitthisarn, S., 2018, Ontology develop ment for intangible cultural heritage and folk wisdom of Phatthalung province, Thaksin J. 21(3): 259-266. (in Thai)

[15] Ketui, N., Theeramunkong, T. and Onsuwan, C., 2012, Rule-Based Method for Thai Elementary Discourse Unit Segmentation (TED-Seg), In 2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems.

[16] Han, J., Kamber, M. and Pei, J., 2012, Data Mining: Concept and Techniques, 3rd Ed., Morgan Kaufmann Publishers, Burlington, MA., 744 p.