An Approach of Network Analysis Enhancing Knowledge Extraction in Thai Newspapers Contexts

Akkharawoot Takhom, *NECTEC,* Dhanon Leenoi, *NECTEC,* Chotanunsub Sophaken, PCSHS, Prachya Boonkwan, *NECTEC,* and Thepchai Supnithi, *NECTEC*

***Abstract*— Reading online newspapers can update us to understand the current situation, and these electronic media also support us in gaining new knowledge in various domains. However, each reader has obstacles on which influence understanding the contexts of newspapers. Different terminologies and their relations in the newspapers cause misunderstandings and ambiguous semantics. This research aims to address problems in a multidisciplinary context. The research presents a network analysis approach that extracts terminologies and represents the relationship among domain concepts in multidisciplinary contexts. A case study demonstrates different concepts in Thai newspapers contexts and visualizes them as a network. The experiment result of the research approach is discussed in identifying ambiguous concepts and their relationships across different domains.**

***Index Terms*— Network Analysis, Visualization, Natural Language Processing, Multidisciplinary Knowledge.**

I. Introduction

READING online newspapers can modernize us to understand the current situation, and these electronic media also support us in gaining new knowledge in several domains. Interpretations of newspapers require different knowledge focusing on the contexts and the relations between domain knowledge.

However, each reader may have obstacles to comprehend the point of the news, misled by words used from different domains. Different terminologies and their relations in the newspapers cause misunderstandings and ambiguous semantics.

Several multidisciplinary research studies have been published and their research approaches contributed to this problematical solving in a multidisciplinary context (Bernard & Anita, 2006). Textual analysis approaches based on a network perspective should be reviewed to clarify miscommunications through the distinct perspectives to manifest the inter-relationship among texts (Aviv et al., 2003; Chaudhry & Higgins, 2003; Daems et al., 2014; Hecking & Hoppe, 2015).

This research aims to address problems in a multidisciplinary context. The research presents the approach of network analysis for extracting terminologies and represents the relationship among domain concepts with multidisciplinary contexts existing in the online newspapers in Thai. By discerning domain-specific terms from the common ones, it is now much easier to visually identify the domains of knowledge and emphasize niche terms in the text. Furthermore, conceptual metaphors, used unconsciously and embedded in Thai language, have been unveiled.

The rest of the paper is organized as follows: Section II explains the research background of domain knowledge in multidisciplinary contexts and related works for network analysis for knowledge extraction. Section III, next, introduces knowledge extraction based on an approach of network analysis, and a network perspective for contexts analysis. An empirical study is taking question and answers contexts from the LCA domain into account in discovering cross-disciplinary concepts. Section IV elaborates our case study, in which different concepts in the newspaper context are extracted, and discusses the experiment results in identifying ambiguous concepts and their relationships across different domains. Section V concludes the main contributions and gives an outlook on further work.

II. Background And Related Works

A. Domain Knowledge in Multidisciplinary Contexts

Multidisciplinarity connects the knowledge from various domains, to explain multiple related meanings which share the same concept boundary (Bernard & Anita, 2006).

Multidisciplinary knowledge refers to a situation in which domain experts intend to explain the common terms' meaning by their own expertise. At the same time, those common terms may be interpreted in different senses but sharing the same concept boundary.

However, the different senses and understandings may lead to miscommunication among distinct domain experts, and may become an obstacle when collaborating and sharing knowledge.

Disruption in information technology has caused newspapers experiencing a significant decline. Relating to the advent of internet, not only the printed news having been challenged, but the readers' knowledge acquisition has been changed as well. Regarding multidisciplinarity, keywords from different specific domains are commonly included in the same news article, causing a cognitive road bump in knowledge acquisition. Detecting these keywords along with

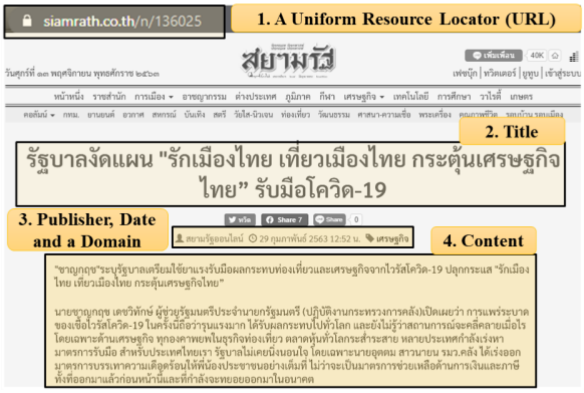
Akkharawoot Takhom, Dhanon Leenoi, Prachya Boonkwan and Thepchai Supnithi are with Language and Semantic Technology Laboratory, National Electronics and Computer Technology Center (NECTEC), Pathumthani 12120, Thailand., Contact authors e-mail: akkharawoot.tak@ncr.nstda.or.th

Chotanunsub Sophaken is with Princess Chulabhorn Science High School (PCSHS), Chiang Rai 57000, Thailand., Contact authors e-mail: chokunworking4713@gmail.com

Manuscript received February 22, 2021; accepted October 27, 2021; revised October 29, 2021; published online October 31, 2021.

their domains in advance helps the readers understand the domain context quicker, and ultimately leverage the text's understandability.

As illustrated in Fig. 1, An online newspaper has four typical components: (1) links or uniform resource locator, commonly known as URL, (2) headline, (3) publisher information consisting of publisher's name, date of issue and specific domain, and (4) news content composed of lead, body, and tail.


1. A Uniform Resource Locator (URL)
สยามรัฐ
วันศุกร์ที่ ๑๓ พฤศจิกายน พุทธศักราช ๒๕๖๓
เพิ่มเพื่อน
40K
เฟซบุ๊ก | ทวิตเตอร์ | ยูทูบ | เข้าสู่ระบบ
หน้าหนึ่ง ราชสำนัก การเมือง อาชญากรรม ต่างประเทศ ภูมิภาค กีฬา เศรษฐกิจ เทคโนโลยี การศึกษา วาไรตี้ เกษตร
คอลัมน์ กทม. ยานยนต์ อากาศ สหกรณ์ บันเทิง สตรี วัยใส-นิวเจน ท่องเที่ยว วัฒนธรรม ศาสนา-ความเชื่อ พระเครื่อง คุณภาพชีวิต รอบบ้าน รอบเมือง
2. Title
3. Publisher, Date and a Domain
ทวีต
Share 7
Share
0
4. Content

siamrath.co.th/n/136025

รัฐบาลจัดแผน "รักเมืองไทย เที่ยวเมืองไทย กระตุ้นเศรษฐกิจไทย" รับมือโควิด-19

สยามรัฐออนไลน์ 29 กุมภาพันธ์ 2563 12:52 น. เศรษฐกิจ

"ชาญกฤช"ระบุรัฐบาลเตรียมใช้ยาแรงรับมือผลกระทบท่องเที่ยวและเศรษฐกิจจากไวรัสโควิด-19 ปลุกกระแส "รักเมืองไทย เที่ยวเมืองไทย กระตุ้นเศรษฐกิจไทย"

นายชาญกฤช เดชวิทักษ์ ผู้ช่วยรัฐมนตรีประจำนายกรัฐมนตรี (ปฏิบัติงานกระทรวงการคลัง)เปิดเผยว่า การแพร่ระบาดของเชื้อไวรัสโควิด-19 ในครั้งนี้ถือว่ารุนแรงมาก ได้รับผลกระทบไปทั่วโลก และยังไม่รู้ว่าสถานการณ์จะคลี่คลายเมื่อไร โดยเฉพาะด้านเศรษฐกิจ ทุกองคาพยพในธุรกิจท่องเที่ยว ตลาดหุ้นทั่วโลกระส่ำระสาย หลายประเทศกำลังเร่งหามาตรการรับมือ สำหรับประเทศไทยเรา รัฐบาลไม่เคยนิ่งนอนใจ โดยเฉพาะนายอุตตม สาวนายน รมว.คลัง ได้เร่งออกมาตรการบรรเทาความเดือดร้อนให้พี่น้องประชาชนอย่างเต็มที่ ไม่ว่าจะเป็นมาตรการช่วยเหลือด้านการเงินและภาษีทั้งที่ออกมาแล้วก่อนหน้านี้และที่กำลังจะทยอยออกมาในอนาคต

Fig. 1 An online newspaper has four typical components: (1) links or uniform resource locator, commonly known as URL, (2) headline, (3) publisher information consisting of publisher's name, date of issue and specific domain, and (4) news content composed of lead, body and tail.

B. *Network Analysis for Knowledge Extraction*

To manifest the inter-relationship among texts, some textual analysis approaches based on a network perspective should be reviewed to clarify miscommunications through the distinct perspectives.

This work relies on a text mining approach to analyze contexts in the online news. A 'bags of words' approach of Blei (Blei, 2012) is one of the approved methods which based on probabilistic topic models. *Network-Text Analysis* (*NTA*) is an approach which extracts the concepts from texts considering the position of words and generates model of knowledge as a concepts' network (Carley et al., 2013). The meta-matrix (Diesner & Carley, 2005), distinguishing categories of synonymous words, introduces ontological categories to the analysis. The relations between concepts are introduced when corresponding words appear within certain proximity in the normalized text. The resulting networks or graphs could be specialized according to the categories of the meta-matrix. Through this way, concepts and knowledge-with-relation figure a semantic network or concept map as described by Novak et al. (Novak & Cañas, 2008). Similar approaches of using *knowledge graphs* (*KG*) have been described by Popping (Popping, 2003). In the analysis of collaborative learning interactions, such representations have been used to trace the development of knowledge in learners (Engelmann et al., 2009; Engelmann & Hesse, 2010; Schreiber & Engelmann, 2010). Since combining text mining with a networked representation of knowledge which incorporating ontological categories corresponding to multi-disciplinary perspective in distinct domains, the NTA approach has been selected in this work.

As illustrated in Table I, our approach has been determined by considering the following criteria: (1) related work, (2) usage, (3) information source, and (4) multidisciplinarity. Related works are elaborated in the following explanation to figure out the uniqueness of this paper.

TABLE I
RELATED WORKS BASED AN APPROACH OF NETWORK-TEXT ANALYSIS

Related Work	Information Source	Multidisciplinarity
Aviv et al., 2003	Discussion Forum	Education
Chaudhry & Higgins, 2003	Discussion Forum	Education
Daems et al., 2014	Discussion Forum	Education
Hecking & Hoppe, 2015	Discussion Forum	Education
Takhom et al., 2020	Discussion Forum	Sustainable Development
This research study	Newspapers	Social, Politic, Sport, Economic, Environment, Culture, Education, Finance

Firstly, a source of information, Aviv et al. analyzed the information features in academic domain. Their findings manifest the benefit of a formal structure and closed-forum reaching high efficiency of students' critical thinking.

Next, based on the utilization of a discussion forum, Chaudhry et al. (Chaudhry & Higgins, 2003) discovered the nature of multidisciplinary in sources of information from academic websites from various countries. Each curriculum having subtopics crossed to other curriculums has been found. The conclusion mentioned the issue of sharing knowledge for designing a new curriculum. The limitation of information volume causes ambiguity of topic sharing. Through the designing and development of a multidisciplinary paradigm, the contexts have been shared within those works.

Consequently, an analytical approach to the text network has been reviewed. Many works have been overcome those limitations, especially in an academic research field. The NTA approach has been employed in the following. Daem et al. (Daems et al., 2014) presented an analytical process of finding cross-relation between texts through collaborative activities of student's text editing. Afterwards, Hecking et al. (Hecking & Hoppe, 2015) granted the utilization of the NTA approach to present the contributors in collaborative writing scenarios. As the aforementioned benefits, sources of information possess many features to be extracted before developing a supporting tool, especially knowledge construction systems.

C. *Multidisciplinarity in Newspaper Articles*

In this work, our approach assumes the selected method could discover the ambiguous parts of information sharing through discovering of concepts and relation-in-between. The result of discovering concepts is exploited in a process of knowledge extraction based on a technique of Text Mining as the NTA approach. This work relies on many approaches in the following related works.

First, Rajman et al (Rajman & Besançon, 1998) utilized Text Mining techniques to discover two examples of information has been automatically extracted from text collections: probabilistic associations of key-words and prototypical document instances.

Next, Alani et al (Alani et al., 2003) used the knowledge extraction tool to search the online documents and to extracts knowledge matching the given classification structure. Afterwards, they provided knowledge in a machine-readable format which automatically being maintained in a *knowledge base* (*KB*). Knowledge extraction is further enhanced using a lexicon-based term expansion mechanism that provides extended ontology terminology.

Recently, Savova et al (Savova et al., 2010) built and evaluated an open-source NLP system for information extraction from electronic medical record clinical free-text, the *clinical Text Analysis and Knowledge Extraction System* (*cTAKES*). Its components, specifically trained for the clinical domain, create rich linguistic and semantic annotations. The cTAKES annotations are the foundation for methods and modules for higher-level semantic processing of clinical free.

Therefore, the research approach in this paper intends to discover the multidisciplinarity of newspaper contexts through the NTA approach. The research challenge is to extract knowledge from Thai online-newspapers, which may contain information leading the possibility of misunderstandings. Moreover, the essential dimension of this research is the ambiguity of information sharing, where different domains manifest a certain degree of unpredictability and some perspectives of ambiguity. This problem has been considered as a kind of communication, which is a significant obstacle to comprehend the multi-domain key-phrases existing in newspaper contexts.

D. *Natural Language Processing for Thai*

For the *preprocessing* on Thai language, common state of technology and research progress on the Thai language processing (Sornlertlamvanich, Potipiti, Wutiwiwatchai, et al., 2000; Tapsai et al., 2019, 2020) narrates the Thai characteristics and the approaches to deal with the difficulties in each processing task.

Unlike English including other languages with word boundaries, neither word nor sentence boundary have been clearly indicated in Thai language. In many cases, Thai words are implicitly recognized and fluctuated depending on the individual judgement. This causes difficulties on the Thai language processing.

*Word and Sentence Segmentation***,** the problem of word identification and segmentation has to be performed. For the most part, the Thai language processing relies on manually handcrafted dictionaries (Sornlertlamvanich, Potipiti, & Charoenporn, 2000), which encounter the inconsistencies in defining word-units and limitation in the quantity. For word segmentation, the longest matching, maximal matching and probabilistic segmentation had been applied in the early research. Conversely, these approaches have some limitations dealing with unknown words. Some word segmentation with advanced-techniques occupied in many language features: context words, parts of speech, collocations and semantics (Kawtrakul, 1997; Meknavin et al., 1997). At the same time, the trigram model was adopted to increase accuracy on sentence segmentation.

As the complexity of Thai language, the Thai NLP had been slowly developed until the advent of AI FOR THAI in 2019 (National Electronics and Computer Technology, 2021), a hub of the Thai NLP and AI research dissemination for developers and general users to create any application for the real-world business usage. In the hub, it consists of basic NLP, such as Thai text and speech processing, including image processing. Also, a Thai-English machine-readable dictionary and corpus, a Thai OCR and plagiarism, a Thai Q/A and speech database, are available (Tapsai et al., 2019, 2020).

III. Knowledge Extraction Based on An Approach of Network Analysis

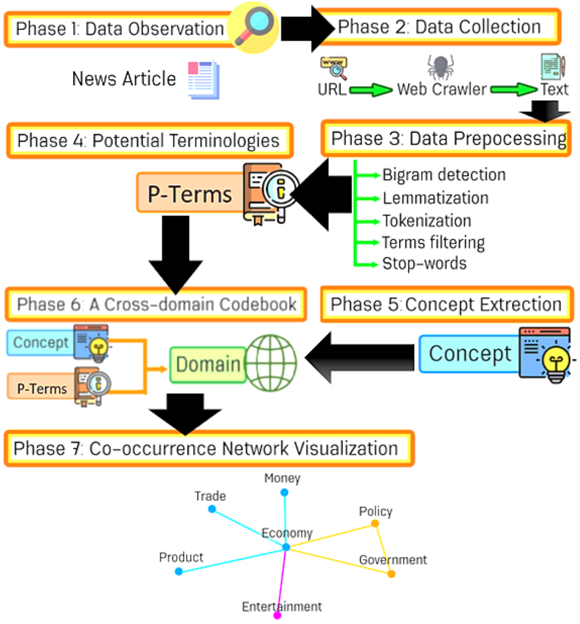


Fig. 2 Seven phases in a procedure workflow based on Network-Text Analysis (NTA) (Daems et al., 2014; Diesner & Carley, 2004; Takhom et al., 2017).

To discover cross-disciplinary concepts, we follow the NTA experiment practice (Diesner & Carley, 2004) as described in Fig. 2.

A. *Phase 1: Data observation*

We identify available sources of knowledge and observe various sources of knowledge along with suitable tools for data manipulation (Bird et al., 2009).

B. *Phase 2: Data collection*

We collect data from selected sources of knowledge.

C. *Phase 3: Data* preprocessing

To prepare data in textual form acquiring lexical analysis. Collected data in this paper are the online news, Thai language, in a domain of interest that contains a large number of words.

As illustrated in TABLE II, we divide this phase into five steps. We will use the following Thai text as an example for running through.

Thai: [สำนักข่าวซีเอ็นเอ็นรายงานในวันที่ 2 ก.ค. 2563 ว่างานแต่งงานดังกล่าวจัดขึ้นที่เมืองปาลีกันชะ]
/sǎmnák kʰà:o si: ʔen ʔen rɑ:i-ŋɑ:n nɑi vɑn tʰî: sɔ̌:ŋ kɔ:-kʰɔ: sɔ̌:ŋ-hâ:-hòk-sǎ:m vâ: ŋɑ:n tæ̀ŋ-ŋɑ:n ɕɑt kʰɯ̂: tʰî: mɯɑŋ pa:-li:-kɑn-ɕʰɑʔ/

English: ‘CNN News Agency reported on July 2, 2020 that the wedding was held in Palikancha.’

Step 3.1) Thai Word Segmentation:

This step is to tokenize a stream of text into words. *Thai language tokenization* utilized the Bidirectional deep learning of context representation for joint word segmentation and POS tagging (Boonkwan & Supnithi, 2018).

As shown in the TABLE II shown, source data in Thai was translated in English as follows.

Thai: สำนัก|ข่าว|ซี|เอ็น|เอ็น| |รายงาน|ใน|วัน|ที่| |2| |ก.ค.| |2563| | ว่า| |งาน| |แต่งงาน|ดัง|กล่าว|จัด|ขึ้น|ที่|เมืองปาลี|กันชะ| /sǎmnák//kʰà:o//si://ʔen//ʔen//rɑ:i ŋɑ:n//nɑi//vɑn//tʰî://sɔ̌:ŋ//kɔ: kʰɔ://sɔ̌:ŋ hâ: hòk sǎ:m//vâ:/ /ŋɑ:n//tæ̀ŋ ŋɑ:n//ɕɑt//kʰɯ̂://tʰî://mɯɑŋ//pa: li://kɑn//ɕʰɑʔ/

Segmentation: agency | news | C | N | N | to report | in | date | of | 2 | | Jul | | B.E.2563 | that | the ceremony | wedding | as mentioned | to be held | at | Pali | kancha

English: ‘CNN News Agency reported on July 2, 2020 that the wedding was held in Palikancha.’

Step 3.2) Stopwords Removal

This step is to remove stopwords to enhance the language processing.

Step 3.3) Detecting bigram and frequency distribution

This step is to detect a sequence of two words in juxtaposition, so-called two-gram words, and each key-phrase is counted by cumulative frequency. Bigram Detection is a sequence of two adjacent elements from a string of tokens, as words. As shown in the TABLE II, the example of detecting bigram with frequency distribution.

Our primary objective is to observe the network of concept terms in the dataset. Here we focus on extracting cross-disciplinary technical terms, where most of them are a composition of common words. Based on our data observation, most cross-disciplinary technical terms are short and consist of only one or two words. Longer technical terms are used only when domain disambiguation is needed. *Named entity recognition* (NER) (Ritter et al., 2011; Tirasaroj & Aroonmanakun, 2011) would come in handy when a cross-disciplinary technical term contains a proper name.

Step 3.4) Filtering frequency distribution

This step is to select high-frequency terms from bigram words.

D. Phase 4: Potential terminologies selection

This phase handcrafts a list of terms identifying concepts of expected relevance.

E. Phase 5: Concept extraction

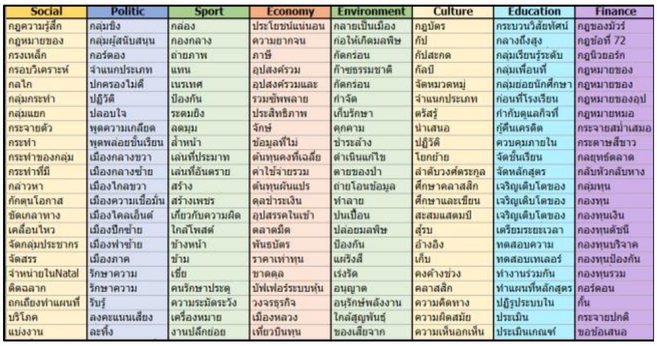
This phase is to utilize a semantic network for recognizing the cross-disciplinary concepts in specific domains’ contexts. The semantic network represents concepts and its relation, commonly known as *‘Ontology’* (Horrocks, 2008), a language of semantic expression to specify data fields, concepts, and relations between concepts. In this paper, extracting specific-domain concepts is prepared to analyze the meaning of ambiguous terms, i.e., particular meanings of multiple-domain terms. Besides, if a new concept has been discovered, it can be inserted during the process. Additional concepts could be added through extracting domain concepts, lexicons, and glossaries (e.g., economic or environment) from the specific dictionaries.

TABLE II
EXAMPLES OF SOURCE DATA AND PREPROCESSED RESULTS.

Pre-processing Type	Preprocessed Contents
Source Data	สำนักข่าว ซีเอ็นเอ็น รายงานในวันที่ 2 ก.ค. 2563 ว่า งานแต่งงานดังกล่าวจัดขึ้นที่เมืองปาลีกันชะ
Step 3.1: Word Segmentation	สำนัก|ข่าว|ซี|เอ็น|เอ็น| |รายงาน|ใน|วัน|ที่ |2| |ก.ค.| | |2563| ว่า| |งาน|แต่งงาน|ดัง|กล่าว|จัด|ขึ้น|ที่|เมืองปาลี|กัน|ชะ|
Step 3.2: Stopwords Removal	สำนัก ข่าว ซี เอ็น เอ็น รายงาน วัน ก.ค. งาน แต่งงาน กล่าว จัด ขึ้น เมืองปาลี ชะ
Step 3.3: Detecting bigram and frequency distribution	สำนัก, ข่าว, 2 ข่าว, ซี, 2 ซี, เอ็น, 1 เอ็น, เอ็น, 1 เอ็น, รายงาน, 1 รายงาน, วัน, 1 วัน, ก.ค, 1 ก.ค, ., 1 ., งาน, 1 งาน, แต่งงาน, 3 แต่งงาน, กล่าว, 1 กล่าว, จัด, 1 จัด, ขึ้น, 1 ขึ้น, เมืองปาลี, 1 เมืองปาลี, ชะ, 1
Step 3.4: Filtering frequency distribution	สำนัก, ข่าว, 2 งาน, แต่งงาน, 3 สำนัก, ข่าว, 2

F. Phase 6: Building of cross-domain codebooks

In this phase, a process of associating the terminologies with the extracted concepts to categorize under relevant domains, is achieved through the meta-matrix feature of NTA. In this phase, a cross-domain codebook containing a set of rational triplets associating with two different domains has been generated semi-automatically. Table III, the generated codebook represents domain categories in a way which integrates potential terms (Phase 4) with domain concepts (Phase 5). The English-translation-into-Thai, it is possible that potential term and concept are difference. As shown in the fourth row of Table III, an economic term ‘deflationary’ translated to ‘คลังหดตัว’ /kʰlɑŋ hòt tu:a/ but ‘contractionary’ with the same translation in economic-domain concept.



Social	Politic	Sport	Economy	Environment	Culture	Education	Finance
กฎความรู้สึก	กลุ่มชิง	กล่อง	ประโยชน์แน่นอน	กลายเป็นเมือง	กฎบัตร	กระบวนวิสัยทัศน์	กฎของมัวร์
กฎหมายของ	กลุ่มผู้สนับสนุน	กองกลาง	ความยากจน	ก่อให้เกิดมลพิษ	กัป	กลางถึงสูง	กฎข้อที่ 72
กรงเหล็ก	กอร์ดอง	ถ่ายภาพ	ภาษี	กัดกร่อน	กัปสะกด	กลุ่มเรียนรู้ระดับ	กฎนิวยอร์ก
กรอบวิเคราะห์	จำแนกประเภท	แทน	อุปสงค์รวม	ก๊าซธรรมชาติ	กัลป์	กลุ่มเพื่อนที่	กฎหมายของ
กลไก	ปกครองไม่ดี	เนรเทศ	อุปสงค์รวมและ	กัดกร่อน	จัดหมวดหมู่	กลุ่มย่อยนักศึกษา	กฎหมายของ
กลุ่มกระทำ	ปฏิวัติ	ป้องกัน	รวมซัพพลาย	กำจัด	จำแนกประเภท	ก่อนที่โรงเรียน	กฎหมายของอุป
กลุ่มแยก	ปลอบใจ	ระดมยิง	ประสิทธิภาพ	เก็บรักษา	ตรัสรู้	กำกับดูแลกิจที่	กฎหมายหมอ
กระจายตัว	พูดความเกลียด	ลดมุม	จักษ์	คุกคาม	นำเสนอ	กู้คืนเครดิต	กระจายสม่ำเสมอ
กระทำ	พูดพล่อยชั้นเรียน	ล้ำหน้า	ข้อมูลที่ไม่	ชำระล้าง	ปฏิวัติ	ควบคุมภายใน	กระดาษสีขาว
กระทำของกลุ่ม	เมืองกลางขวา	เล่นที่ประมาท	ต้นทุนคงที่เฉลี่ย	ดำเนินแก้ไข	โยกย้าย	จัดชั้นเรียน	กลยุทธ์ตลาด
กระทำที่มี	เมืองกลางซ้าย	เล่นที่อันตราย	ค่าใช้จ่ายรวม	ตายของป่า	ลำดับวงศ์ตระกูล	จัดหลักสูตร	กลับหัวกลับหาง
กล่าวหา	เมืองไกลขวา	สร้าง	ต้นทุนผันแปร	ถ่ายโอนข้อมูล	ศึกษาคลาสสิก	เจริญเติบโตของ	กลุ่มทุน
กักตุนโอกาส	เมืองความเชื่อมั่น	สร้างเพชร	ดุลชำระเงิน	ทำลาย	ศึกษาและเขียน	เจริญเติบโตของ	กองทุน
ขัดเกลาทาง	เมืองไคลเอ็นต์	เกี่ยวกับความผิด	อุปสรรคในเข้า	ปนเปื้อน	สะสมแสตมป์	เจริญเติบโตของ	กองทุนเงิน
เคลื่อนไหว	เมืองปีกซ้าย	ไกล์โพสต์	ตลาดมืด	ปล่อยมลพิษ	สู้รบ	เตรียมระยะเวลา	กองทุนดัชนี
จัดกลุ่มประชากร	เมืองฟาซ้าย	ข้างหน้า	พันธบัตร	ป้องกัน	อ้างอิง	ทดสอบความ	กองทุนบริจาค
จัดสรร	เมืองภาค	ข้าม	ราคาเท่าทุน	แผ่รังสี	เก็บ	ทดสอบเทเลอร์	กองทุนป้องกัน
จำหน่ายในNatal	รักษาความ	เชี่ย	ขาดดุล	เร่งรัด	คงค้างช่วง	ทำงานร่วมกัน	กองทุนรวม
ติดฉลาก	รักษาความ	คนรักษาประตู	บัฟเฟอร์ระบบหุ้น	อนุญาต	คลาสสิก	ทำแผนที่หลักสูตร	กอร์ดอน
ถกเถียงทำแผนที่	รับรู้	ความระมัดระวัง	วงจรธุรกิจ	อนุรักษ์พลังงาน	ความคิดทาง	ปฏิรูประบบใน	กั้น
บริโภค	ลงคะแนนเสียง	เครื่องหมาย	เมืองหลวง	ใกล้สูญพันธุ์	ความผิดสมัย	ประเมิน	กระจายปกติ
แบ่งงาน	ละทิ้ง	งานปลีกย่อย	เที่ยวบินทุน	ของเสียจาก	ความเห็นอกเห็น	ประเมินเกณฑ์	ขอข้อเสนอ

Fig. 3 Excerpt of domain concepts extracted from domain glossary consists of eight domains. Note that some terms are common among several domains.

TABLE III
EXCERPT OF A CODEBOOK

Potential Term	Concept	Domain Knowledge
รับผิดชอบต่อสังคม Social responsibility	ความรับผิดชอบต่อสังคม Social responsibility	Social
ประชาธิปไตย Democracy	ปฏิวัติประชาธิปไตย Democratic revolution	Politic
จุดยิงลูกโทษ Penalty shootout	จุดโทษ Penalty	Sport
คลังหดตัว **Deflationary**	**นโยบายการคลังหดตัว** **Contractionary fiscal policy**	**Economic**
อนุรักษ์ Conservation	อนุรักษ์พลังงาน Energy conservation	Environment
นครประวัติศาสตร์ Historic town	ประวัติศาสตร์ History	Culture
บูรณาการ Integration	สอนแบบบูรณาการ Integrate Curriculum	Education
บริหารความเสี่ยง Risk management	กองทุนบริหารความเสี่ยง Hedge fund	Finance

G. *Phase 7: Visualization of co-occurrence networks*

To manifest a multidisciplinary relation in sources of knowledge, a visualization of co-occurrence network (*Phase 7*) is generated. The generated network is utilized in analytical processes concerning the semantic meaning of cross-disciplinary concepts in news contexts.

In this phase, the co-occurrence network can be calculated by using the following terms: cross-domain concept name, adjacent concepts, incident edges, and domain knowledge (Takhom et al., 2020). We compute the edge score by equation (1).

$$cov(d_1, d_2) = \sum_{(d_1,v),(d_2,v)\in D} W(v_1, v_2) \tag{1}$$

where W is a function that maps each edge, D is a relation of each vertex and its corresponding domain, a concept name v is said to be *cross-domain* if there exist (v, d_1) and (v, d_2) in D. Two cross-domain concept names v_1 and v_2 are said to be *adjacent* if there exists an edge $(v_1, v_2) \in E$. Two edges $e_1 = (v_1, v_1')$ and $e_2 = (v_2, v_2')$ are said to be *cross-domain incident* if $v_1' = v_2$, $(v_1', d_1) \in D$, $(v_2, d_2) \in D$ and $d_1 \neq d_2$. We define the coverage score of two distinct domains d_1 and d_2, denoted by $cov(d_1, d_2)$

IV. EXPERIMENT RESULTS AND DISCUSSION

A. *Dataset*

Thai Newspapers Summarization Project was organized by Thailand Ministry of Higher Education, Science, Research and Innovation (MHESI) during August 1, 2020 to October 30, 2020. This project aimed to summarize news contents and annotate domain knowledge.

157 skilled volunteers, Thai native-speakers, who were unemployed during the COVID-19 pandemic, read the collected online news. Afterwards, they manually summarized and carefully annotated into eight news-domains: politic and economic, social and sport, education and culture, environment and finance.

B. *Results and Discussion*

As a result, 4,513 summarized news, from various online-sources during August to September 2020, were retrieved and preprocessed with the aforementioned NLP steps. The resultant dataset consists of 21,153 segmented words. Consequently, 574,001 pairs of bigram and frequency distribution were extracted.

Next, we visualize the knowledge graph of network text analysis. In Fig. 4(a), there are eight domains of knowledge, where each of which is represented by different colors. The color-mapping table for each domain is displayed in Fig 3. We observe that keywords and key phrases are classified into domain-specific knowledge networks. Some terms are used in multiple domains, e.g. ‵โจร′ /tɕo:n/ ‘thief’ is used in three domains: social, economics, and environment. Some terms are domain-specific, e.g. ‵แม่น้ำเจ้าพระยา′ /mæ̂:-ná:m-tɕâo-pʰrá-ja:/ ‘Chaopraya River’ falls into the environment domain. By discerning domain-specific terms from the common ones, it is now much easier to visually identify the domains of knowledge and emphasize niche terms in the text.

Note that the extracted knowledge domains correspond to the theory of FrameNet (Fillmore, 1985, 2008; Geeraerts, 2010). Here, a domain of knowledge is recalled once specific keywords and key phrases, aka. *evoking words*, are used. In this paper, we use Thai FrameNet (Leenoi et al., 2010) as a reference. As shown in Fig 4(b), ‵ฝาก′ /fà:k/ ‘deposit’, ‵โอน′ /ʔo:n/ ‘transfer’ and ‵เงิน′ /ŋɤ:n/ ‘money’ belong to the finance domain. On the other hand, ‵อ่าง-เก็บ-น้ำ′ /ʔà:ŋ kèp ná:m/ ‘reservoir’, ‵ระบาย′ /rɑʔ-ba:i/ ‘drain’, ‵ท่วม′ /tʰû:am/ ‘flood’ and ‵เขื่อน′ /kʰɯ̀:an/ ‘dam’ belong to the environment domain.

As a byproduct, we also discover *conceptual metaphors* (Lakoff & Johnson, 2008), the understanding of one domain in terms of another, which are used unconsciously in Thai. For example, we revealed that the conceptual metaphor of [MONEY IS LIQUID], the understanding of money being flowable, exists via the following evidence: ‵‵ไหล′ /lǎ:i/ ‘flow’, ‵กระแส′ /krɑʔ-sǣ:/ ‘stream’, or ‵ละลาย′ /lɑʔ-la:j/ ‘melt’ are words used in water domain, a subclass of the environment domain, but also used in the fiscal and financial domain as a subclass of the economic domain.

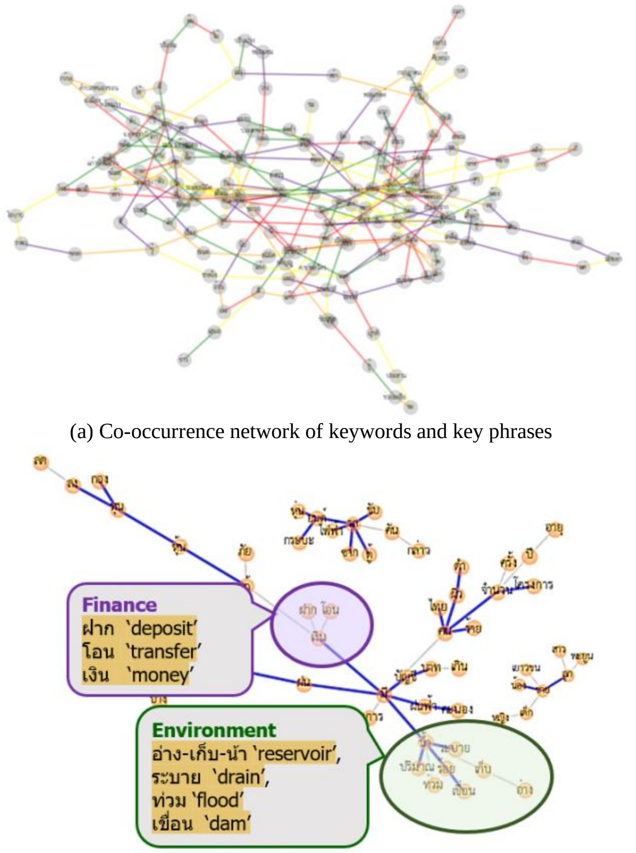
V. CONCLUSION

This paper presents an approach of network analysis for extracting knowledge through question and Thai newspapers contexts by employing the network-text analysis method.

The key contributions are in the following. Firstly, we discovered potential terminologies and their relations in the context of the newspaper, and analyzed the domain concepts, also. Secondly, the network analysis approach represents multidisciplinary knowledge and found the codebook can transpose misunderstandable terminologies to domain concepts. Finally, the experimental results were interpreted and discussed the result causes of miscommunication.

ACKNOWLEDGMENT

The authors are special thanks to reviewers for fruitful technical comments. This research is partially supported by National Electronics and Computer Technology Center (NECTEC), Thailand.



(a) Co-occurrence network of keywords and key phrases

(b) Multidisciplinarity in knowledge graph

Fig. 4 Excerpt of a generated co-occurrence network of multidisciplinarity.

REFERENCES

[1] Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., & Shadbolt, N. R. (2003). Automatic Ontology-Based Knowledge Extraction from Web Documents. *IEEE Intelligent Systems*, *18*(1). https://doi.org/10.1109/MIS.2003.1179189

[2] Aviv, R., Erlich, Z., Ravid, G., & Geva, A. (2003). Network analysis of knowledge construction in asynchronous learning networks. *Journal of Asynchronous Learning Networks*, *7*(3), 1–23. http://www.academia.edu/download/43215417/v7n3_aviv_1.pdf%5Cnhttp://www.ravid.org/gilad/v7n3_aviv.pdf

[3] Bernard, C. K. C., & Anita, W. P. P. (2006). Multidisciplinarity Interdisciplinarity and Transdis Ciplinarity in Health Research Services Education and Policy: 1. Definitions, Objectives, and Evidence of Effectiveness. *Clinical and Investigative Medicine*, *29*(6), 351–364.

[4] Bird, S., Klein, E., & Beijing, E. L. (2009). Natural Language Processing with Python. In *Journal of Chemical Information and Modeling* (1st ed.). O'Reilly Media, Inc. https://doi.org/10.1097/00004770-200204000-00018

[5] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

[6] Boonkwan, P., & Supnithi, T. (2018). Bidirectional deep learning of context representation for joint word segmentation and POS tagging. *Advances in Intelligent Systems and Computing*, *629*. https://doi.org/10.1007/978-3-319-61911-8_17

[7] Carley, K. M., Pfeffer, J., Reminga, J., Storrick, J., & Columbus, D. (2013). *ORA user's guide 2013*.

[8] Chaudhry, A. S., & Higgins, S. (2003). On the need for a multidisciplinary approach to education for knowledge management. *Library Review*, *52*(2), 65–69. https://doi.org/10.1108/00242530310462134

[9] Daems, O., Erkens, M., Malzahn, N., & Hoppe, H. U. (2014). Using content analysis and domain ontologies to check learners' understanding of science concepts. *Journal of Computers in Education*, *1*(2–3), 113–131. https://doi.org/10.1007/s40692-014-0013-y

[10] Diesner, J., & Carley, K. M. (2004). Using Network Text Analysis to Detect the Organizational Structure of Covert Networks * Jana Diesner, Kathleen M. Carley. *Communication*.

[11] Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In *Causal mapping for research in information technology* (pp. 81–108). IGI Global.

[12] Engelmann, T., Dehler, J., Bodemer, D., & Buder, J. (2009). Knowledge awareness in CSCL: A psychological perspective. *Computers in Human Behavior*, *25*(4), 949–960.

[13] Engelmann, T., & Hesse, F. W. (2010). How digital concept maps about the collaborators' knowledge and information influence computer-supported collaborative problem solving. *International Journal of Computer-Supported Collaborative Learning*, *5*(3), 299–319.

[14] Fillmore, C. J. (1985). Frames and the semantics of understanding. In *Quaderni di Semantica* (Vol. 6).

[15] Fillmore, C. J. (2008). Frame semantics. *Cognitive Linguistics: Basic Readings*, *34*, 373–400. https://doi.org/10.1075/hop.2.fra1

[16] Geeraerts, D. (2010). Theories of Lexical Semantics. In *Theories of Lexical Semantics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198700302.001.0001

[17] Hecking, T., & Hoppe, H. U. (2015). A network based approach for the visualization and analysis of collaboratively edited texts. *CEUR Workshop Proceedings*, *1518*, 19–23. http://ceur-ws.org/Vol-1518/paper4.pdf

[18] Horrocks, I. (2008). Ontologies and the semantic web. *Communications of the ACM*, *51*, 58–67. https://doi.org/10.1145/1409360.1409377

[19] Kawtrakul, A. (1997). Automatic Thai unknown word recognition. *Proc. 4th Natural Language Processing Pacific Rim Symposium (NLPRS-97), Phuket, Thailand, Oct.*, 341–346.

[20] Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

[21] Leenoi, D., Jumpathong, S., Porkaew, P., & Supnithi, T. (2010). Thai FrameNet Construction and Tools. *Int. J. Asian Lang. Process.*, *21*(2), 71–82.

[22] Meknavin, S., Charoenpornsawat, P., & Kijsirikul, B. (1997). Feature-based Thai Word Segmentation. *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97)*, *97*, 41–46.

[23] National Electronics and Computer Technology. (2021). *AI for Thai Platform*. https://aiforthai.in.th/about.php

[24] Novak, J. D., & Cañas, A. J. (2008). *The theory underlying concept maps and how to construct and use them*.

[25] Popping, R. (2003). Knowledge graphs and network text analysis. *Social Science Information*, *42*(1), 91–106. https://doi.org/10.1177/0539018403042001798

[26] Rajman, M., & Besançon, R. (1998). *Text Mining - Knowledge extraction from unstructured textual data*. https://doi.org/10.1007/978-3-642-72253-0_64

[27] Ritter, A., Clark, S., Etzioni, O., & others. (2011). Named entity recognition in tweets: an experimental study. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1524–1534.

[28] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5). https://doi.org/10.1136/jamia.2009.001560

[29] Schreiber, M., & Engelmann, T. (2010). Knowledge and information awareness for initiating transactive memory system processes of computer-supported collaborating ad hoc groups. *Computers in Human Behavior*, *26*(6), 1701–1709.

[30] Sornlertlamvanich, V., Potipiti, T., & Charoenporn, T. (2000). *Automatic corpus-based Thai word extraction with the c4.5 learning algorithm*. https://doi.org/10.3115/992730.992762

[31] Sornlertlamvanich, V., Potipiti, T., Wutiwiwatchai, C., & Mittrapiyanuruk, P. (2000). *The state of the art in Thai language processing*. https://doi.org/10.3115/1075218.1075296

[32] Takhom, A., Boonkwan, P., Ikeda, M., Usanavasin, S., & Supnithi, T. (2017). Reducing miscommunication in cross-disciplinary concept discovery using network text analysis and semantic embedding. *The 6th Joint International Semantic Technology Conference, CEUR Workshop Proceedings 1741, 2000*, 20–31. http://ceur-ws.org/Vol-2000/paos2017_paper3.pdf

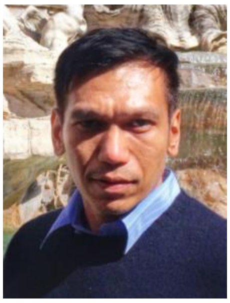
[33] Takhom, A., Usanavasin, S., Supnithi, T., Ikeda, M., Hoppe, H. U., & Boonkwan, P. (2020). Discovering cross-disciplinary concepts in multidisciplinary context through collaborative framework. *International Journal of Knowledge and Systems Science*, *11*(2). https://doi.org/10.4018/IJKSS.2020040101

[34] Tapsai, C., Meesad, P., & Unger, H. (2019). An Overview on the development of Thai natural language processing. *Information Technology Journal*, *15*(2), 45–52.

[35] Tapsai, C., Unger, H., & Meesad, P. (2020). *Thai Natural Language Processing: Word Segmentation, Semantic Analysis, and Application* (Vol. 918). Springer Nature.

[36] Tirasaroj, N., & Aroonmanakun, W. (2011). The effect of answer patterns for supervised named entity recognition in Thai. *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, 392–399.

Akkharawoot Takhom received the B.S. degree in Management of Information Technology from Mae Fah Luang University in 2009, and M.Eng. degree in Information and Communication Technology for Embedded Systems from SIIT in 2013. He received Ph.D. degree in Knowledge Science from the JAIST in 2018 and received Ph.D. degree in Engineering and Technology from SIIT in 2019.



Dhanon Leenoi received B.A. from Thammasat University in 2000 and M.A. in Linguistics from Chulalongkorn University in 2008. Since 2009, he has been with Language and Semantic Technology Laboratory, NECTEC, Thailand.

Chotanunsub Sophaken is studying Grade 7 in Princess Chulabhorn Science High School. He has been the internship student in Language and Semantic Technology Laboratory, NECTEC, Thailand.

Prachya Boonkwan received B.Eng. and M.Eng. degrees in Computer Engineering from Kasetsart University in 2002 and 2005, respectively. He received a Ph.D. degree in Informatics from the University of Edinburgh, UK, in 2014. Since 2005, he has been with Language and Semantic Technology Lab at NECTEC in Thailand.

Thepchai Supnithi received the B.S. degree in Mathematics from Chulalongkorn University in 1992. He received the M.S. and Ph.D. degrees in Engineering from the Osaka University in 1997 and 2001 , respectively. He is currently head of language and semantic research team artificial intelligence research unit, NECTEC, Thailand.