IEICE TRANS. INF. & SYST., VOL.E101–D, NO.4 APRIL 2018

SURVEY PAPER *Special Section on Data Engineering and Information Management*

A Survey of Thai Knowledge Extraction for the Semantic Web Research and Tools

Ponrudee NETISOPAKUL[†a)], ***Member and*** **Gerhard WOHLGENANNT**[††b)], ***Nonmember***

SUMMARY As the manual creation of domain models and also of linked data is very costly, the extraction of knowledge from structured and unstructured data has been one of the central research areas in the Semantic Web field in the last two decades. Here, we look specifically at the extraction of formalized knowledge from natural language text, which is the most abundant source of human knowledge available. There are many tools on hand for information and knowledge extraction for English natural language, for written Thai language the situation is different. The goal of this work is to assess the state-of-the-art of research on formal knowledge extraction specifically from Thai language text, and then give suggestions and practical research ideas on how to improve the state-of-the-art. To address the goal, first we distinguish nine knowledge extraction for the Semantic Web tasks defined in literature on knowledge extraction from English text, for example taxonomy extraction, relation extraction, or named entity recognition. For each of the nine tasks, we analyze the publications and tools available for Thai text in the form of a comprehensive literature survey. Additionally to our assessment, we measure the self-assessment by the Thai research community with the help of a questionnaire-based survey on each of the tasks. Furthermore, the structure and size of the Thai community is analyzed using complex literature database queries. Combining all the collected information we finally identify research gaps in knowledge extraction from Thai language. An extensive list of practical research ideas is presented, focusing on concrete suggestions for every knowledge extraction task – which can be implemented and evaluated with reasonable effort. Besides the task-specific hints for improvements of the state-of-the-art, we also include general recommendations on how to raise the efficiency of the respective research community.

key words: *knowledge extraction, Thai language text, landscape analysis, semantic web*

1. Introduction

The majority of information available on the Web is in the form of natural language text, making text the most abundant source of human knowledge. Knowledge Extraction (KE) generates knowledge from structured and unstructured data in machine-readable and machine-interpretable formats. In the area of the Semantic Web, knowledge extraction has a long history. One central area is the (semi-)automatic creation of domain models, called ontology learning. Ontology learning from text includes various tasks, including terminology extraction, finding synonyms and lifting terms into domain concepts, the extraction of taxonomic relations, and others [1]. On the other hand, the information extraction (IE) and KE community focuses on more narrow tasks of varying task complexity, for example named entity recognition, relation extraction, event detection, and many more.

In this work, we study the research and tools designed for KE for Thai language text and compare it to English language KE, concentrating specifically on the set of tasks which can be directly translated into Semantic Web constructs. We roughly follow the task categorizes and task mappings suggested by Gangemi [2] in his survey of KE for English language, and evaluate the state-of-the-art in KE from Thai text based on this list of tasks.

This publication is motivated by a number of factors. Firstly, there has been a lot of interesting work in KE from English text in the last couple of years, including increasing support of various KE tasks with a number of powerful tools. A detailed assessment of the KE for Thai language landscape is currently missing, as well as an evaluation of research gaps and research opportunities.

Here, we want to assess the state-of-the-art of the landscape of research and tool support for Thai language in order to find gaps and research directions. We aim not only to identify broad gaps, but rather on practical research ideas which can be executed and implemented with reasonable effort (low-hanging fruits). Secondly, we want to analyze the structure of the Thai research community and to detect limiting factors in the research culture in Thailand.

In order to address the research goals we first define nine KE tasks which already exist in KE from English text. Then, we combine a survey of available literature and tools with a questionnaire-based study about Thai KE in which we approach members of the Thai research community. The results from the questionnaire provide a self-assessment of the Thai community and pointers to additional research work in the field. A number of complex Scopus queries are the base for analyzing the structure and absolute and relative size of Thai research community. Finally, based on the information gathered, we provide an interpretation of the Thai language state-of-the-art per KE task, and hope to give helpful suggestions on research directions and concrete ideas to the Thai IE and KE communities.

The remainder of this paper is organized as follows: Sect. 2 presents related work. Section 3 introduces the research and data acquisition methods used in this survey, and describes the nine KE tasks. In Sect. 4 we discuss the struc-

Manuscript received June 17, 2017.
Manuscript revised October 24, 2017.
Manuscript publicized January 18, 2018.

[†]The author is with the Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand.
[††]The author is with the Intern. Lab. of Information Science and Semantic Technologies, ITMO University, St. Petersburg, Russia.
a) E-mail: ponrudee@it.kmitl.ac.th
b) E-mail: gwohlg@corp.ifmo.ru

DOI: 10.1587/transinf.2017DAR0001

ture and size of the respective Thai research community, present the results from the literature survey, and provide research ideas. Finally, Sect. 5 concludes the paper with a summary and a list of contributions.

2. Related Work

This section provides some background on the knowledge extraction tasks selected for this publication, and also about the peculiarities of Thai language processing and AI research. The section does not include the detailed analysis of work done on KE for Thai language text, as in this survey article it will be part of Sect. 4 (*Survey Results*).

In 2013 Aldo Gangemi [2] published a landscape analysis for English language knowledge extraction for the Semantic Web (KE4SW) tools. First he selects various KE tasks and suggests a mapping of tasks to Semantic Web (SW) constructs. Available tools are evaluated on a small text sample with measures such as accuracy, recall, precision and F1. One of the main findings is that automatic KE4SW is feasible in principle, but KE tools lack standardized output formats, and often the translation to ontological design constructs is unclear. In this work, our list of tasks is inspired by the work of Gangemi, although we tried to simplify the set of tasks, in order to merge tasks that have a similar mapping to SW constructs, and to make the tasks self-explanatory for researchers participating in our questionnaire study. As the number of available KE tools is very limited for Thai language, we have a stronger focus on existing research methods and publications for the specific tasks than on tools, and also we aim at giving insights into surrounding aspects such as the structure of the Thai research community.

Most work on Thai language NLP and KE states that the automatic processing of Thai written language is challenging due to a number of aspects: First of all, Thai language is a continuous stream of characters and does not include marks for word or sentence boundaries [3]. Therefore, tools for automatic word and sentence segmentation are necessary. Also, analysis is made difficult by features such as flexible word order, zero anaphora, the absence of upper-/lower-case characters, the high ambiguity in compound words, and serial verbs [4]. There has been a lot of work on Thai word segmentation, for example Haruechaiyasak et al. [5] report a task accuracy of 88% when combining Conditional Random Fields (CRF) with a dictionary-based approach. In Thai language word segmentation is connected to the problem of word sense disambiguation (WSD), making WSD also part of preprocessing. The correct word segmentation often depends on the context and sense of words [6]. As stated, the general tool support for Thai language is very limited, and Thai is not supported by well-known NLP toolkits like GATE[†] or NLTK. However, there are tools available for basic NLP tasks like word segmentation and POS-tagging, such as PyThai[††], PyThaiNLP[†††], or WordCut[††††] and RDRPOStagger[†††††] [7]. As the accuracy for word segmentation is only around 90%, some work on high-level tasks (such as relation extraction) omits it altogether. Other basic NLP tools such as chunkers (shallow parsing) or full parsers are currently not available for Thai [8].

Regarding the context and history of AI research in Thailand, including research in NLP, IE and KE, Kawtrakul and Praneetpolgrang [9] give a number of interesting insights. They distinguish a pioneering period (before 1999), with initial work mainly by Kasetsart University and NECTEC. After 2000, research was mostly driven by research road-maps from governmental organizations (National Research Council of Thailand) or organizations such as NECTEC, with national research programs and specific application domains. After 2011 there was a stronger focus on practical applications of AI technology. The work of Kawtrakul and Praneetpolgrang helps to provide background information on the historical research directions in AI in Thailand and more specifically NLP, IE and KE, and on the evolution and structure of the research community. The past federal research road-maps have a strong focus on some application domains such as agriculture, tourism and medicine, which is reflected in our literature survey. The focus on application domains also partly helps to understand the lack of generic KE tools for Thai language text.

3. Methods and Tasks

In this section, first we specify the research methods employed in this study. Those include an extensive literature review combined with a survey in order to combine our analysis with a self-assessment by the Thai research community. In the main part of the section, we enumerate and describe the set of knowledge extraction tasks which we use to analyze the state-of-the-art in Thai formal KE.

3.1 Research Methods

The main research methods and processes applied in this work are as follows:

1. **Literature study**: We conducted an extensive literature study focusing on methods and tools related to knowledge extraction from Thai language text. The literature study had three phases: (i) Phase one focused on the study of relevant research work found in literature databases (mainly Scopus[††††††]). In phase one we analyzed and categorized publications according to the predefined list of knowledge extraction tasks (see Sect. 3.2). Furthermore, we searched for relevant KE

[†]https://gate.ac.uk

[††]https://pypi.python.org/pypi/pythai
[†††]https://travis-ci.org/wannaphongcom/pythainlp
[††††]https://gitlab.com/veer66/wordcutpy
[†††††]http://rdrpostagger.sourceforge.net
[††††††]https://www.scopus.com

tools for Thai language and collected a list of relevant persons (authors) and organizations. (ii) In phase two, utilizing the list of researchers, a questionnaire-based survey was conducted (see below). The survey provided links to additional research work, which was integrated into the results of the research survey in phase (iii).

2. **Survey research**: With the survey-based study in form of a questionnaire, we pursued two goals: Firstly, to have a community assessment by Thai researchers – for any of the nine KE tasks, additionally to our assessment. The other reason for the questionnaire survey was to get a more complete picture of existing research (methods and tools) in Thai KE, so we asked for links to persons and organizations, and for tools, for any of the KE tasks.
3. **Research community metrics**: With the help of database queries (Scopus) we assess the organizational structure of the Thai research community and also compare the quantity of Thai research on each of the tasks to the respective quantity for the Japanese community in order to measure the Thai community.

3.2 The Knowledge Extraction Tasks

Loosely based on the work by Gangemi [2], in the following we distinguish nine knowledge extraction for the Semantic Web (KE4SW) tasks. Gangemi [2] evaluates the state-of-the-art of tools for those tasks for English language, the language where research is most advanced. In Sect. 4 we use the blueprint by Gangemi [2] to evaluate the literature and tool landscape of Thai language KE for any of the tasks.

The tasks include both basic IE and KE tasks such as Named Entity Recognition (NER), as well as more complex tasks such as semantic role labeling and event detection. The tasks are concerned with various aspects of extracting and classifying knowledge from text, some tasks operate on the document level (parts of task 1 to task 3), but mostly the extraction and classification operates on the level of keywords, multi-word phrases, and structures within a sentence.

As we are interested in KE4SW, the results of each task are mapped to Semantic Web constructs. When mapping to a Semantic Web construct, the output of the tool is described as one or multiple RDF subject-predicate-object triples using the established Semantic Web vocabulary. This allows the use of existing Semantic Web technologies to process and analyze the output, and to potentially integrate it into the Web of Data. The mapping suggestions are inspired by Gangemi [2]. In the following we refer to Semantic Web vocabulary with the common namespace prefixes `rdf`, `rdfs`, and `owl`, more details are found in the official documentation of the W3C[†].

A listing of the KE4SW tasks is found below. The listing includes a (i) description of the task, (ii) a mapping of the task results to the Semantic Web (*SW Mapping*), (iii) and to make the tasks more graspable we include an example of output produced by a KE tool for a input sentence. The tools (see below) were applied to this simple sentence: `"Parents were driving their young children to a zoo in the U.S. state of Colorado."` The tools mentioned below all support English language text, some of them support other languages as well, but none of the tools can process Thai text. In this section we define the KE tasks based on tools existing for English language, and Sect. 4 presents an evaluation of work existing on those tasks for Thai text.

There are many tools available for most KE4SW tasks for English language, our selection of tools is a rather random choice from the options, serving the purpose to give examples. For the sake of brevity, we can only present snippets of the results produced by the tools, the full results are found online[††]. We used the following tools to generate the output for the sample sentence:

- *Fred*: Fred [10] is a "machine reader for the Semantic Web" and can read text from 48 languages and transform it to linked data[†††].
- *Topia*: Topia is a Python package which includes linguistic tools such as POS-tagging and some simple statistical analysis to determine terms and their strength[††††].
- *Alchemy*: AlchemyLanguage is a collection of APIs that offer text analysis through natural language processing. It can analyze text and extract its concepts, entities, keywords, sentiment, and more. Alchemy is commercial product, with a demo online[†††††].
- *AMALGrAM 2.0*: AMALGrAM[††††††] analyzes English sentences for multiword expressions (MWEs) and noun and verb supersenses [11].

The list of tasks is as follows:

1. **T1 – Topic and keyword extraction:** In this task, for a single document, the algorithm extracts keywords, or assigns topics from a list of topics to the document. So the input are single documents, which are then annotated with keywords or topics (result).
   SW mapping: document `dc:subject` *topic*
   As a reminder, *SW mapping* stands for the mapping of tool output to Semantic Web constructs, and `dc:subject` refers to the popular Dublin Core vocabulary[*]. The input to the task is a sentence (or a longer text), the output is a list of terms (keywords): For the sample sentence, *Topia* produces the following list of keywords: `['Colorado', 'Parent', 'U.S', 'children', 'state', 'zoo']`
2. **T2 – Terminology extraction:** Here, the goal is to extract relevant terminology for a domain. The input to

[†]https://www.w3.org/standards/semanticweb/ontology

[††]https://aic.ai.wu.ac.at/~wohlg/thaiKE/
[†††]http://wit.istc.cnr.it/stlab-tools/fred
[††††]https://pypi.python.org/pypi/topia.termextract
[†††††]https://alchemy-language-demo.mybluemix.net/
[††††††]https://github.com/nschneid/pysupersensetagger
[*]http://dublincore.org

this task is a domain corpus, ie. a collection of text documents. The task output is a set of relevant terms – which can be transformed for example to concepts in a domain ontology.
SW mapping: term `rdf:type owl:Class`
In this task, the input is a full domain corpus, so we cannot execute the task on our example sentence.

3. **T3 – Taxonomy extraction:** T3 aims at the extraction of subclass (is-a) relations between terms or concepts. Some approaches address this task simultaneously with T2. The input to this task is a text (or corpus), typically from a specific domain. The output are triples with is-A (rdfs:subClassOf) relations connecting subject and object.
SW mapping: concept-a `rdfs:subClassOf` *concept-b*
From our example sentence, *Fred* generates the following subclass relations:
```
@prefix fr: http://ontologydesignpatterns.org/ont/fred .
fr:domain.owl#YoungChild rdfs:subClassOf fr:domain.owl#Child .
```
In the example output, the classes `YoungChild` and `Child` are abstractions found by Fred for the sentence part "young children".
4. **T4 – Binary relation extraction (RE):** In this task, the focus is on the extraction of relation triples, again in the form of (subject, predicate, object), from text. Typically, there is a distinction between methods for closed RE, where the types of relations are predefined, and open RE, where the method or tool extracts arbitrary relations. The triples are extracted from an input sentence.
SW mapping: term-a relation-type term-b
Alchemy extracts two relations from the example sentences:
```
Parents locatedAt zoo .
children locatedAt zoo .
```
As seen in the example, for a textual input (eg. a sentence), the output is a list of relation triples, with the relation connecting subject and object.
5. **T5 – Named Entity Recognition (NER):** This task is concerned with the detection of Named Entities (NE) in text, and their classification, typically as Person, Organization, Location or time – or using a more sophisticated schema.
Again, the input is for example a sentence, and here the system detects and categorizes named entities in that input.
SW mapping: term/URI `a` *Person*|*Organization*|...
Fred generates the following output about the term *Colorado* from the example sentence:
```
fr:domain.owl#Colorado owl:sameAs <http://dbpedia.org/resource/Colorado> .
<http://dbpedia.org/resource/Colorado> a schema:Place, schema:AdministrativeArea .
```
The tool detected the named entity *Colorado*, and classified it as `schema:Place`.
6. **T6 – Named Entity Linking (NEL):** In this task, the tool or method links (*grounds*) a NE into an existing knowledge base such as DBpedia[†]. The input is text and the NE-annotations, the result consists of links from annotations to KBs.
SW mapping: term/URI `owl:sameAs` *URI in KB (eg. DBpedia)*
Often, tasks T5 and T6 are connected and conducted together; for our example sentence, *Fred* has already linked Colorado to the corresponding entry in the knowledge base DBpedia in T5.
7. **T7 – Supersense Tagging (SST):** A central challenge of computational lexical semantics is to abstract from the surface form of words to their general meaning [11]. NER (Tasks 5) tackles part of this challenge, but is limited to specific types of words. A more general and dense approach is Supersense tagging (SST), which is an NLP task that consists in annotating each significant entity in a text, like nouns, verbs, adjectives and adverbs, according to a general semantic taxonomy defined by the 41 WordNet lexicographer classes (called Supersenses) [12]. A task related to T7 is word sense disambiguation (WSD). As the name suggests, in WSD the goal is to select the correct word sense of words in a text, based on word context. WSD is typically utilizing lexical databases such as WordNet, which enumerate and define word senses. Due to missing word boundaries in Thai language, WSD affects word segmentation [6] and is part of the NLP preprocessing pipeline. Therefore it is not included in this survey on high-level knowledge extraction. Supersense tagging is an established task in between low-level WSD and NER. As in the example below, Schneider and Smith [11] do not categorize single words, but rather multi-word expressions. For an input text, the tool maps the constituents to WordNet basic classes.
SW mapping: term/URI `rdf:type` *WordNet basic class*
For the input sentence, the *AMALGrAM* supersense tagger generates the following annotations: *Parents*$_{[PERSON]}$, *were driving*$_{[motion]}$, *their young children*$_{[PERSON]}$, *to a zoo*$_{[ARTIFACT]}$, *in the U.S. state-of-Colorado*$_{[COMMUNICATION]}$.
In the example sentence, every multiword expression (phrase) is annotated with the corresponding WordNet Supersense.
8. **T8 – Semantic role labeling (SRL):** Sometimes also called shallow semantic parsing, SRL consists of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles. Often FrameNet [13] is used as a base for the roles and frame elements.
The input is a text (eg. a sentence), the result consists of extracted frames and their respective frame elements.
SW mapping: verb/URI `rdf:type` *frame-type*

[†]http://wiki.dbpedia.org

verb/URI role-type frame-element
For the example sentence, *Fred* extracts the "drive" frame. The drive frame has various frame elements defined, like the *Destination*, the *Agent (Driver)*, etc.

```
@prefix ns1: <http://ontologydesignpatterns.
org/ont/fred/domain.owl#> .
@prefix ns3: <http://ontologydesignpatterns.
org/ont/vn/abox/role/> .
ns1:drive_1 a ns1:Drive ;
ns3:Destination ns1:zoo_1 ;
ns3:Agent ns1:parent_1 ;
ns3:Source ns1:U.s._state ;
ns3:Theme ns1:child_1 .
```

In the example output, multiple assumptions about the subject `ns1:drive` are made: The subject is of type `ns1:Drive`, therefore a "drive" frame, and has definitions for frame elements, for example the *Destination* is the *zoo*.

9. **T9 – Event extraction (EE):** This usually includes the detection of event types (for example the event "company acquisition") in text, and the extraction of the participants in the event.
   Again, the input is a sentence, and the tool outputs the events discovered.
   SW mapping: event `rdf:type` *Event-type*
   Fred extracts the *Drive* event from the our sentence:
   `ns1:Drive rdfs:subClassOf ns2:Event .`

3.3 Natural Language Processing and Knowledge Extraction Tasks

In this paragraph, we emphasize the relation and the dependency of KE and general NLP. Typically, natural language processing (NLP) is a series of tasks, performed consecutively. After a basic sentence and word segmentation (wordseg) task, the first step is often morphology analysis, that is, to transform words into their lemma forms. Depending on the task at hand, more steps are applied: POS tagging, NER, syntactic parsing (synpars), and semantic parsing (sempars). Most of our nine KE4SW tasks can be tackled with various approaches which required different NLP preprocessing steps. Table 1 presents an overview showing which NLP tasks are generally applied as basis for the KE task. Noted that some research works experiment with totally different approaches and do not fit this categorization. Also, there are other non-standard NLP tasks like word sense disambiguation (WSD) and pronoun resolution which are relevant to complex KE tasks like T8 and T9.

Table 1 NLP tasks typically executed in pre-processing for KE tasks.

KE Task	wordseg	senseg	POS	NER	synpars	sempars
T1	+	+/-	+/-	+/-	-	-
T2	+	+/-	+	+	-	-
T3	+	+	+	+	+/-	+/-
T4	+	+	+	+	+	+
T5	+	+	+	+	+/-	+/-
T6	+	+	+	+	+/-	+/-
T7	+	+	+	+	+	+
T8	+	+	+	+	+/-	+
T9	+	+	+	+	+	+

+ means the NLP task is usually applied before the KE task
+/- means the NLP task may benefit the KE task
- means the NLP task is normally not necessary for the KE task

4. Survey Results and Discussion

In this section we present the main results of this research, which include (i) an analysis of the structure and size of the community, (ii) a detailed survey of methods and tools for any of the KE tasks including the community self-assessment, and a discussion section presenting the research gaps and practical research ideas.

4.1 Structure of the Community

Here we analyze the structure of the Thai IE and KE research community, and confirm the impression that the community is small and limited to a few organizations. An analysis of the Thai research community and of historical developments helps to understand why the state-of-the-art evolved this way, and shows which are the leading organizations in specific subfields. Furthermore, we compare the Thai community to the size of the Japanese community, in order to get a feeling for relative sizes. The purpose of this survey article is to evaluate the state-of-the-art in KE from Thai text. In order to measure the Thai KE community size and community focus, we compare it to another Asian research community. Like in Thai KE, also for KE from Japanese text, research is mostly performed by researchers from the respective country, and not the global community (like in English language KE). This allows for an objective comparison of community structure.

We use queries to the Scopus[†] scientific literature database to estimate the number of contributions. Scopus will not include all relevant scientific work, but provide a sufficient number of results to gain insights into the community structure. The number of publications was chosen to structure the Thai KE community as it is a good indicator for the scientific output and quality of scientific work of research units.

Figure 1 provides an overview of the ratio of research contributions in the KE field by Thai research organizations. It is based on a complex Scopus query, which involves keywords from all our KE tasks. All underlying data, queries and more detailed results presented in this section are available online[††]. The publications gathered from the query were filtered manually for relevance. In total, we found 104 publications relevant to our nine KE tasks. Among the eight leading institutions, SIIT has the biggest share of publications, with around 27%. Further important organizations are

[†]https://www.scopus.com
[††]https://aic.ai.wu.ac.at/~wohlg/thaiKE/stats.xlsx

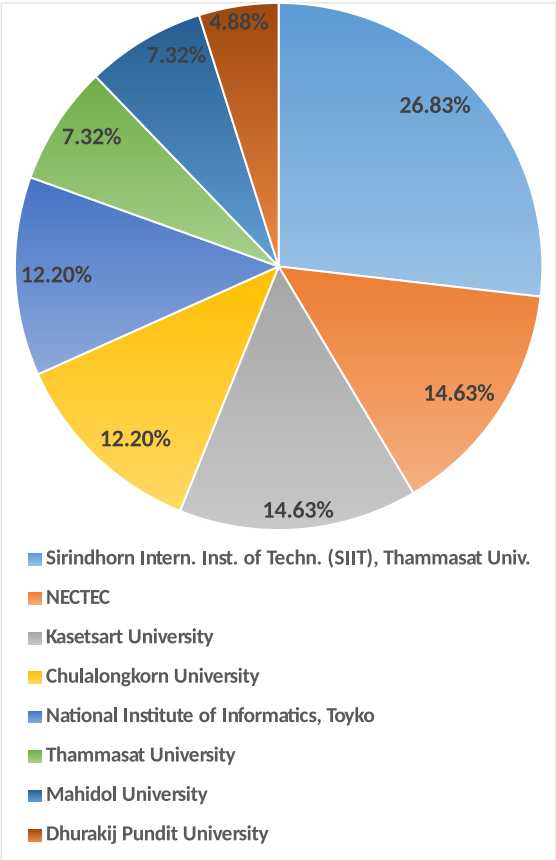

Fig. 1 Share of research publications of the leading research organizations on KE for Thai language.

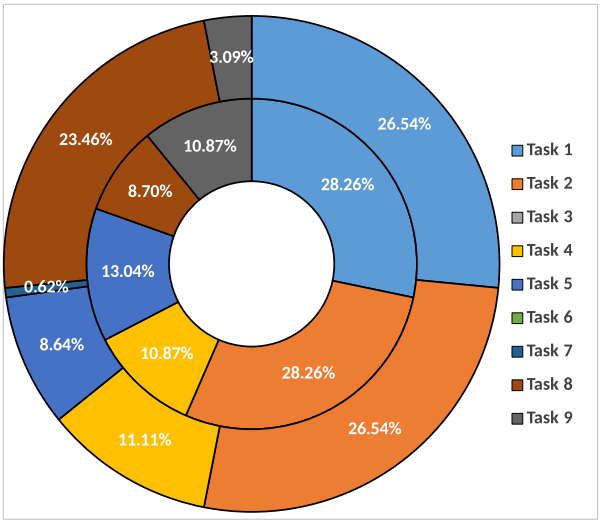

Fig. 2 A comparison of the ratios of relative research work done per KE tasks. Research on Thai text (inner circle) and Japanese text (outer circle).

NECTEC and Kasetsart University (both around 15%), and others. All publications found were in the period between 2006 and 2017.

Note that according to the Scopus database, SIIT, although named under the umbrella of Thammasat University, is a separate institution from Thammasat University. This accurately reflects the differences between both institutions, for example regarding the faculty recruitment and evaluation methods, research funding and internationalization level. While Thammasat University has broad range of faculties and is recognized as one of the leading traditional Thai Universities, SIIT only focuses on science and engineering programs and is operating based on faculty competence and contracts.

A detailed investigation of the literature reveals that Thai IE and KE research community can be grouped around four important researchers, surrounded by their graduate students and associates. Those prominent researchers are Prof. Theeramonkong from SIIT, Assoc. Prof. Kawtrakul and her colleagues from Kasetsart University, Dr. Supnithi from NECTEC, and Assoc. Prof. Aroonmanakun from Chulalongkorn University. These four groups comprise around 68% of the total IE and KE publications in the Scopus database. Also note that three of them are Japanese graduates. The ongoing collaborations with their Japanese mentors and colleagues undoubtedly contribute to their success and consequently, to Thai IE and KE research community. This helps to explain the significant number (12.2%) of Thai IE and KE publications published by a Japanese institution, the National Institute of Informatics (NII).

A comparison between the Thai and Japanese research communities shows quite consistently that the number of publications is around 5-7 times higher in the Japanese community. With Scopus queries, we found factors of 5.0 for "text processing", of 5.5 for "semantic web / ontologies", and of 6.8 for "language processing". These numbers confirm the impression that the Thai research community is comparably small.

Additionally to the comparison of the absolute quantity of publications, Fig. 2 shows the relative research effort per KE task, again comparing the Thai and Japanese community. The figure gives the percentage of publications per task (from the total number of publications). While it is clear that the Japanese IE and KE community has a higher number of research papers published in absolute numbers (see above) it is quite interesting to note that, except task 8 and task 9, the ratio of research on each task is surprisingly similar for both communities. For example, as seen in Fig. 2, task 1 (keyword and topic extraction), as well as task 2 (taxonomy extraction), comprise around 28% for the Thai community and 26.5% for the Japanese community. For both communities, we found very little work on taxonomy extraction (tasks 3), named entity linking (task 6), and supersense tagging (task 7), while both communities show a number of 11% for task 4, relation extraction. NER (task 5) attracted more effort for Thai language (13% versus 9%). The largest gaps are found for task 8 (SRL) and task 9 (event extraction). There has been substantial work on Japanese for tasks 8 with 23% (Thai with only 9%). On the other hand, the Thai community has addressed event extraction (task 9) with 11% of total work, while the Japanese community only published 3% of its work on event extraction.

Although these figures can change over time, it can be

Table 2 The number of publications per task and the application domains, as well as the total number of publications per KE task.

Domain vs. Task	1	2	3	4	5	6	7	8	9
General	3	1			4	1		2	2
Social Media	4								
Medicine / Chemistry	1				2	1		1	1
Agriculture		2	2	2					1
Culture and Tourism				2	2				
News				2	5			1	
Other	2			2	1				
No. of Publications	10	3	2	8	14	2		4	4

an indicator of some common research interests among Thai and Japanese researchers, as well as highlighting the different focus areas of each community.

4.2 Existing Methods and Tools

In this section we present the analysis of literature on KE from Thai language text. First we give an overview of existing work regarding the *application domain*, and also regarding the *methods used* in the publications. The main part of the section consists of a survey of existing work for each of the nine KE tasks. As already mentioned, the list of relevant publications was determined by a combination of queries to literature databases, and a questionnaire-based study. As mentioned in the previous section, we started from an initial number of 104 publications found by a set of Scopus queries. A filtering process removed publications with little relevance to KE4SW. This included a lot of work on basic NLP tasks like segementation, POS-tagging or phrase extraction, which sometimes applied basic keyword extraction, etc., but not in context of information or knowledge extraction. Furthermore, we excluded work which had been superseded by extended work of the same authors, for example conference papers that later became parts of journal publications. Overall, with the input from the initial literature search, relevant publications linked in those paper, and the findings from the questionnaire-based survey, the result were 33 publications, which are analyzed in this survey article. In Tables 2 and 3 publications are classified into multiple groups, leading to higher aggregate numbers.

Table 2 first of all gives an overview of the total number of publications per KE tasks, and organizes the publications according to their *application domain*. Task 1, 4, and 5 have gained to most attentions, whereas for tasks 3, 6, and 7 there exists little work. The number and ratios of total research work differ from the analysis in Sect. 4.1, where the attribution to a task was conducted automatically by the Scopus query terms. Here, we classified the work manually depending on its content, and for example although some research contained the word terminology extraction in the title, it was rather concerned with keyword extraction, therefore we grouped it into task 1 instead of task 2. Also in contrast to Sect. 4.1, as a lot of Thai KE work is using NER as a preprocessing step or component in systems for tasks like relation extraction – we classified it into multiple categorizes, as NER was a major part of the publication.

Table 3 Publications by types of methods used in their algorithms.

Method vs. Task	1	2	3	4	5	6	7	8	9	∑
Machine Learning	3			2	8	1		3	3	20
Patterns	1	2	2	7	8	1		2	4	27
Dictionaries	2	1	1	2	3	1		1		11
Ontologies		2	2	4	3	1		3	3	18
IR (like tf-idf)/IS	9	3	2							14

Table 2 indicates the lack of domain-independent work. A lot of existing research was limited to specific application domains. This can be explained by the Thai federal research roadmaps and their focus on domains such as agriculture, tourism and medicine (see Kawtrakul and Praneetpolgrang [9], Sect. 2). As for other languages, more recent work often is applied to text from news and social media.

Another interesting aspect of analysis are the technical methods used in existing work on Thai language KE, exemplified by the 33 publications included into this survey. Table 3 shows that most systems, 27 out of 33, rely at least in part on pattern-based methods. Machine learning and ontologies as background knowledge are also applied very often (in 61%, and 55% of publications, respectively.) A bit less frequent is the use of Information Retrieval and Information Science methods, as well as the integration of information from dictionaries. Most work combines multiple methods.

A more detailed summary of results with tables that classify individual publications regarding their application domain and the methods used is available online[†].

The following subsections present the survey of work on the individual KE tasks on Thai language.

4.2.1 Task 1 – Keyword and Topic Extraction

As already mentioned, in task 1 we investigate methods that annotate documents with keywords or topics, which can be translated to one or multiple SW triples of the form `document <dc:subject> topic`. This task is probably the most basic. This task includes work on general keyword extraction from documents, and the annotation of documents with specific keywords or topics within document classification. Specialized keyword classification and extraction of important Semantic Web constructs is included in tasks 2–9.

We found very few publications solely dedicated to plain *keyword extraction*. Haruechaiyasak et al. [14] propose a language-independent keyword extraction method specifically for non-segmented languages like Thai. The basic idea is to circumvent word segmentation by finding frequently occurring substring-sets in the unsegmented text with the help of pattern mining.

Oftentimes the tf-idf measure is used as a preprocess-

[†]https://aic.ai.wu.ac.at/~wohlg/thaiKE/tabs.pdf

ing or feature extraction step in topic extraction or document classification systems. Classification is then typically done with supervised machine learning (ML) techniques such as SVM. For text classification of social media messages (Twitter tweets), Jotikabukkana et al. [15] present a method which applies semi-supervised learning. They start with the extraction of keywords with tf-idf and a word-article-matrix from pre-classified news media documents. Then they collect tweets with the keywords found previously, and use those to extend the set of keywords per class. Chirawichitchai et al. [16] investigate which term weighting methods and which ML classifiers are best suited for Thai language text classification. They compare for example tf-idf, tfc or entropy weighting for term weighting. With most ML classifiers, ltc term weighting leads to the best results. The work of Viriyavisuthisakul et al. [17] has in a similar direction. The authors investigate which of the well-known similarity measures (such as Euclidean distance, Jaccard distance, etc.) is best suited to be used with a kNN classifier to categorize social media postings into four predefined classes. The work is related to (single-)topic extraction from documents, and uses the well-known tf-idf measure to build a term-document matrix. Lertnattee and Theeramunkong [18] perform text classification of Thai medicinal texts into 8 classes, The specific properties of Thai medical text complicate word segmentation. Among various segmentation and classifier candidates, the authors have the best results with tf-idf as a term representation model and the SVM classifier. Daowadung and Chen [19] classify textbook texts aiming to predict the readability for primary school students. They apply word segmentation and mutual information (MI) to choose terms, and then use tf-idf and a SVM classifier to classify the input documents regarding document readability.

Furthermore, there has been work on emotion classification of Thai text, the strategies are similar to the document classification work already mentioned. For example, Chirawichitchai [20] classifies single sentences from social media posts into six emotional classes (sentiment detection). After NLP preprocessing and bag-of-words feature extraction with tf-idf from input sentences, he reduces the feature space with information gain, and uses various classifiers in a supervised ML approach trained on 1800 manually labeled sentences. SVM provides the best results. Inrak and Sinthupinyo [21] propose the use of *bi-words* features instead of just uni-grams for emotion classification, and apply LSA for representing the document semantics. Using supervised ML based on 200 labeled examples, they show that the bi-word feature improves classification accuracy significantly.

Topic modeling includes topic assignment to documents. Very recently, there has been work on using LDA topic modeling on Thai language by Jiamthapthaksin [22], who includes dictionaries for slang words to improve the topic models.

Finally, in strong contrast to other approaches, Kiatdarakun and Suksompong [23] apply the symbol-level technique of entropy rate and probability distributions of consecutive characters to attribute long text documents to authors – a task related to text classification.

4.2.2 Task 2 – Terminology Extraction

In their work on ontology learning from Thai text in the domain for agriculture, Kawtrakul et al. [24] do terminology and taxonomy extraction using lexico-syntactic patterns (see next section for more details). Although the approach by Imsombut and Kawtrakul [25] (see below for details) focuses on taxonomy extraction, it also involves the extraction of domain terminology with pattern-matching in text from a specific domain. Furthermore, the S-Sense (social sensing) tool [26] generates tagclouds for a corpus on the fly. The corpus typically consists of social media documents. The tagcloud presents the most important terms in the corpus, which can be seen as a simple method for terminology extraction from text.

4.2.3 Task 3 – Taxonomy Extraction

As part of systems that aim at ontology learning from text, both Imsombut and Kawtrakul [25] and Kawtrakul et al. [24] tackle taxonomy extraction from Thai text.

Imsombut and Kawtrakul [25] apply lexico-syntactic patterns and the extraction of information from list items in text in order to learn concepts and taxonomic relations. For preprocessing the input text in the agriculture domain, they apply word segmentation, POS-tagging, and noun phrase detection. The authors also use various techniques to filter candidate terms, especially to select hypernyms from a list of extracted candidates. The approach itself involves domain-specific background knowledge (eg. a domain ontology) and manual pattern creation and is therefore not easily generalizable. Quite similar in goal and method, Kawtrakul et al. [24] propose a method to extract a taxonomy in the agriculture domain. First they learn patterns with the help of an existing ontology, and then apply those patterns to domain text. In order to increase the precision of the system, they filter term candidates, eg. with mutual information or by looking for matching head nouns.

4.2.4 Task 4 – Relation Extraction

In relation extraction (RE) there is usually a distinction between open and closed RE. All the work we found on Thai text is in the area of closed RE, ie. the set of relations to be extracted is predefined.

The two most common approaches to RE are pattern-based techniques and machine learning-based ones. Patterns can be created manually or learned with the help of background knowledge. We found three publications which apply hand-crafted patterns to domain text. Sornlertlamvanich and Kruengkrai [27] perform RE in the domain of a Thai cultural database. They use the title of the item as subject of the triple, and then extract the objects with patterns from the

description texts of the relation subjects. To achieve high accuracy, they only accept certain named entities as objects (see below on the NE extraction part). The extracted relations are finally visualized in a knowledge graph. The work of Imsombut and Sirikayon [28] uses a similar approach in the domain of tourist attractions, which address both tasks 4 and 5. They extract predefined relations with hand-crafted patterns between named entities. The goal is different in the work of Kowsrihawat and Vateekul [29]. Their service called *JudgeDoll* extracts the main facts and summarizes Thai Supreme Court verdicts. They apply closed RE with lexico-syntactic patterns in order to find the respective information in the legal texts.

In some earlier work, Kawtrakul et al. [30] combine and extend their previous work on knowledge extraction and question-answering in the agriculture domain focusing on cause-effect relations. To extract cause-effect relations they use manually created rules (patterns), with the goal to do frame-based slot-filling to extract facts based on a given domain ontology. The domain ontology is used to integrate data sources and defines the format of the extraction patterns.

Beyond the manual creation of extraction patterns, already in 2005, Kongwan and Kawtrakul [31] did some interesting work on the extraction of facts (closed binary RE) using background knowledge to learn the respective patterns. They apply lexicons to identify interesting object properties and also property values in the agriculture domain, for example that a mangosteen fruit has a weight of 110 grams (object, property, value). The work of Sitthisarn and Bahoh [32] is at the core of KE4SW, they aim at the automatic extraction or annotation of RDF from text sources. With the help of a on a small ontology, their system extracts instanciations of ontology concepts and relations in the domain of Thai official correspondence. On the downside, the application domain is very narrow, and the authors use hand-crafted patterns and dictionaries to find the relevant entries.

As RE can be viewed as a classification problem, in recent work on English text supervised ML is applied often. Also for Thai, there has been work on RE which applies ML to classify relation candidates. Tongtep and Theeramunkong [4] present work on the extraction of four binary relation types from crime-related news. The four relation types are: action-location, location-action, action-person and person-action. They use 9 surface-level features from an annotated training corpus in a ML-based method. To raise accuracy only relations between NEs are classified. An evaluation of various ML techniques to classify the relations yields best results for SVM. As a predecessor of that work, Tongtep and Theeramunkong [33] conduct work on RE between named entities, also in the domain of crime related news. Here, they focus on feature extraction based on lexical patterns and the surrounding context of NEs. Again, the authors apply supervised ML to classify and filter extracted relations.

4.2.5 Task 5 – Named Entity Recognition

Most work on NER in Thai language does not approach classic NER-tagging as a whole, but focuses on the extraction of specific instance types for given application scenarios. Regarding methods, most authors use pattern-based methods, some also apply ML-based techniques.

In the area of pattern-based systems, Imsombut and Sirikayon [28] extract NEs which are instances of *tourism attractions* or *tourism activities* with Conditional Random Fields (CRF) and a hand-crafted tourism ontology for categorization – so the approach is rather a method for domain-specific ontology population. As a preprocessing step before doing rule extraction, Intarapaiboon et al. [34] extract and annotate specific entities, namely chemical reaction names and chemical substances in domain text, or medical entities, respectively [8]. Similarly, Sitthisarn and Bahoh [32] apply NER during preprocessing for their ontology population and relation extraction task. Sutheebanjard and Premchaiswadi [35] have a different goal, they seek to find fast and efficient ways to extract NE only of type person from text, without the necessity of applying word segmentation or POS tagging. The system uses the surrounding context and rules to find person candidates in text in three domains. Focusing on another specific entity type the main goal of Vichayakitti and Jaruskulchai [36] is to find temporal expressions (dates, etc.). They call their approach temporal event extraction, in our task classification it is better placed in task 5. Their method is based on the manual creation of patterns from text tagged with temporal expressions.

Finally, in their RE approach, Tongtep and Theeramunkong [4] use NE extraction as a preprocessing step. They present algorithmic improvements such as giving priority to certain patterns based on the length of the matching segments, as well as heuristics for pattern characteristics and pattern ordering. This work is based on the pattern extraction methods studied in Tongtep and Theeramunkong [33], which extracts persons, locations, and actions from crime-related news documents.

Besides pattern-based system, there some authors that rely on ML for NER. For example, Sornlertlamvanich and Kruengkrai [27] apply NER on the description texts of items in a Thai cultural database. They train their own NE detection system with a sequence labeling algorithm using an annotated corpus. In contrast to most work in NER, which focuses on feature selection to detect named entities, Tirasaroj and Aroonmanakun [37] experiment with different ways and levels of complexity in annotation of NEs in text, for example just having one annotation tag for a person, or splitting into first name and surname. Their aim is to see which type of NE-annotation leads to the best results using CRF in a supervised ML setting. A completely different approach is proposed by Tongtep and Theeramunkong [3], who identify NEs directly from unsegmented text. They utilize statistics of characters and their clusters to find Thai words and NEs simultaneously to classify them with CRF models.

Training and evaluating NER systems requires large annotated corpora. Those efforts to create large annotated corpora lay the foundation for tool support in the future. Theeramunkong et al. [38] present a framework for NE tagging called Thai-NEST[†]. They define a specific tagging process and provide GUI-based tagging tools. The specifications include a tag set and tagging guidelines. The goal of Thai-NEST is to provide a large NE corpus, to extend the set of NE types, and to support a wider area of domains as compared to previous work. Among the results of Thai-NEST are 10,000 articles annotated with about 45,000 NEs in seven domains (eg. sports, politics, etc.). In a more recent effort, Aw et al. [39] created a big corpus for word segmentation, POS-tagging, and NER. A special emphasis in their project is the tagging of foreign and loan words in Thai language, which become more frequent over time. The corpus includes two domains, and in total around 4 million words, which are tagged with 35 POS tags and 10 different NER categories. Obviously, the creation of annotated corpora is a very expensive process, therefore Tongtep and Theeramunkong [40] suggest a method for semi-automatic annotation of POS and named entity information. With their combination of lexical patterns and statistical methods they reduced the number of unknown tokens in the corpus to 16%. Unfortunately, also for known tokens the annotations are not always correct, so manual verification is still necessary.

4.2.6 Task 6 – Named Entity Linking

Related to the task, but applied only in a narrow field, Intarapaiboon et al. [34] link specific entities in the chemistry domain to two background ontologies as part of their semantics-based IR system. The authors include no specific details on the linking process in their work.

In contrast, Rungsawang et al. [41] propose the only generic work related to NEL that we found. They try to *wikify* Thai documents, ie. to link terms in documents to Wikipedia. First the authors are extracting terms from the Wikipedia title and anchor texts in order to build a controlled vocabulary. The central and most challenging aspect is to disambiguate between link target candidates, the authors apply ML trained on various features extracted from the link contexts existing in Thai Wikipedia documents. The approach is even more general than classic NEL, as any kind of term is linked to Thai Wikipedia, not only NEs.

4.2.7 Task 7 – Supersense Tagging

For this task, there currently exists no work for Thai language text – to the best of our knowledge.

4.2.8 Task 8 – Semantic Relation Labeling

The most relevant work on this task has been done by Leenoi et al. [42]. They created a Thai version of FrameNet called TFN. In the process, on the one hand, the authors translate Berkeley FrameNet [13] into Thai with bilingual dictionaries, and partly manually. Additionally, the authors add new Thai-specific frames. The TFN resource contains around 1300 frames, and over 22,000 lexical units and 2300 annotated sentences. Obviously, TFN is an important initial step towards automatic Semantic Role Labeling algorithms and tools for Thai.

The other approaches do not directly refer to FrameNet. Tongtep and Theeramunkong [4] state that their work on relation extraction is strongly related to SRL, however, it is limited to a small number of specific roles, and also restricted to relations only between NEs. Within the chemistry domain the method for semantics-based IR of Intarapaiboon et al. [34] uses pattern-based rule extraction and multi-slot frames. They generate frame extraction rules with the WHISK algorithm. Extracted frames are represented with description logic, and also can be directly encoded in OWL. Finally to provide background knowledge, the concept entities are linked to existing ontologies about chemical reactions. In a similar approach, Intarapaiboon et al. [8] extract semantic frames using rules learned with WHISK from hand-tagged training data. Again, the rules are applied using a sliding window on unsegmented text, and the method is evaluated in multiple domains.

4.2.9 Task 9 – Event Extraction

Event extraction is similar to (binary) relation extraction, but focusing on specific event types and allowing n-ary relations. Due to the situation of missing chunk parsers for Thai language, Intarapaiboon et al. [8], [34], [43] modify the WHISK information extraction technique using sliding windows to operate on Thai text. Those supervised approaches for pattern learning in multiple domains are related to event or frame extraction, they aim at detecting predefined semantic frames from text. In contrast, the work of Kawtrakul et al. [30] is focused on specific relations and events in the agriculture domain as part of their cause-effect relation extraction work.

Although we grouped the existing publications into specific tasks, a lot of the publications involve multiple tasks, except for work on task 1, which is mostly restricted to keyword and topic extraction only. As an example, the work of Imsombut and Kawtrakul [25] addresses both tasks 2 and 3 (terminology and taxonomy extraction), and Intarapaiboon et al. [34] even touches four different tasks (task 5, 6, 8 and 9).

4.2.10 Tools

The set of tools which are openly available for KE from Thai language is still very limited. Most past research work was focused on underlying NLP tasks, which include work on word segmentation [5], [44] and part-of-speech tagging [45], [46]. There are also a number of easy-to-use tools existing for basic Thai NLP processes, for example

[†]http://saki.siit.tu.ac.th/kindml/thainest

Table 4 An overview of tools and datasets for Thai KE from text, including a short description and the tasks addressed.

Tool or Dataset	Description	Tasks
Tool: S-Sense	Social Media sentiment and tagcloud analysis	(1), 2
Tool: Thai-NEST	GUI-Tool and tagset for NER corpus annotation	5
Dataset: Thai-NEST	NER dataset created with Thai-NEST tool 10.000 annotated articles in 9 domains	5
Dataset: Aw et al. [39]	NER dataset with about 4M words and 2 domains	5

PyThaiNLP[†].

In the KE domain, to our knowledge, there exist no open source tools at all. For example, for the NER task some tagged training corpora are available from annotation projects eg. by Aw et al. [39]. According to our literature study the common practice seems to be to train a NER model on-the-fly, and not to use existing tools or pretrained NER models. Regarding tools, tool support only exists for creating annotated corpora from the Thai-NEST project [38].

The most sophisticated tool for KE from Thai language text, which can be seen as a Web Intelligence platform, is S-Sense [26]. The tool[††] uses Twitter as data source to collect information about specific topics. S-Sense computes the sentiment of social media data, and it also generates tag clouds for the respective topics. Tag could generation is related to task 2 (terminology extraction) as it aims to determine the central terms of a domain or data set.

Table 4 provides an overview of datasets and tools existing for knowledge extraction from Thai text.

4.3 Community Assessment

After the initial assessment of the research landscape on KE from Thai text by querying the Scopus literature database and by following links in research papers found, we conducted a questionnaire study with Thai researchers that have published in the KE field to tackle two goals: (a) to perform a comparison of the state-of-the-art assessment by the Thai research community to our own assessment, (b) to get a more complete picture by identifying additional researchers and organizations and also tools for any of the tasks. This additional information was used to extend the literature and tool survey.

72 Thai researchers were asked via email to fill the questionnaire. We selected all authors of papers in the KE and IE field from our initial set of publications obtained by queries to scientific databases. 12 researchers filled the questionnaire. Amongst other things, we attribute the low number of responses to the fact that probably many of the authors moved away from academia in the meantime or to other research fields. This is especially true for PhD students, who often move into other positions and fields after finishing their degree. Given the goals of the questionnaire-based survey, in particular the goal of getting input on relevant work and tools in Thai language KE which we might have missed, the number of 12 participants is sufficient to collect effective feedback.

Table 5 Assessment of the state-of-the-art in Thai KE by the 12 participants of the questionnaire-based survey; on a scale of 1–5 (low – high).

Task	Mean	Std.Dev.	Min	Max
1	**3.08**	.90	2	5
2	2.75	.87	1	4
3	2.50	.91	1	4
4	2.50	.91	1	4
5	**3.00**	.95	1	4
6	**1.42**	.52	1	2
7	**1.58**	.67	1	3
8	1.75	.75	1	3
9	2.25	.75	1	4

The questionnaire is available online[†††].

In the questionnaire, for each of the nine tasks, the respective researcher is asked to: a) estimate the state-of-the-art for the task for Thai language text – on a scale from 1–5, where **1** stands for "almost no research done on this task" and **5** stands for "high quality research and tools available", b) name persons and organizations doing research on the respective task. And finally, in c) the researcher is asked about tools available for the task, if any. 12 researchers responded on the question about the state-of-the-art (a), 9 researchers give input on persons and organizations (b), and 6 researchers provided feedback on tools available (c).

According to the assessment by Thai IE and KE researchers which is summarized in Table 5, the state-of-the-art is highest for tasks 1 and task 5. This conforms to our analysis, that *Keyword and topic extraction* (task 1) and NER (task 5) have gained the most attention. Also regarding the tasks where the state-of-the-art is low, there is high agreement about task 6 (NEL) and task 7 (SST). There is some disagreement about task 2 (terminology extraction), which might stem from different interpretations of the task. Interestingly, the inter-rater agreement is rather high on tasks which participants rated as having a low state-of-the-art, and higher if the state-of-the-art is more advanced.

On question (b), persons and organizations, there has been major agreement between responders. In total, 12 different researchers, and 6 different organizations were suggested. In terms of tools (c), the participants only mentioned two tools (LexToPro, S-Sense). LexToPro performs only very basic NLP tasks such as word segmentation, and S-Sense is already described in Sect. 4.2.10. In total, the results from the questionnaire regarding points (b) and (c) contained little novelty. The answers helped to identity a couple of additional researchers, otherwise the questionnaire con-

[†]https://github.com/wannaphongcom/pythainlp

[††]http://www.ssense.in.th

[†††]https://goo.gl/forms/Tgv4y4te8Iu1mKvC3

firmed our impression that the research community is small and concentrated within a few organizations.

4.4 Peculiarities of Thai Language Regarding KE

After presenting the Thai state-of-the-art, and as connection to the upcoming discussion section, in this section we analyze the peculiarities of Thai language which complicate NLP and KE processes. Thai language is an unsegmented language. The basic problem of word segmentation itself can not easily be solved as it is ambiguous, and inter-tangled with other NLP and KE tasks.

Firstly, the continuous conjugated text can be legitimately segmented in more than one way, i.e., it can be segmented into different words with different meanings. For example, ตากลม can be segmented into "ตา-eye กลม-round" (having round eyes) or "ตาก-dry ลม-wind" (drying with wind), depending on the surrounding context. Next, Thai language lacks basic terminology; basic terms are usually composed of multiple words. For example, the word "ดอกไม้ – flower" comes from two meaningful words, "ดอก- a bud" and "ไม้ – wood". A compound term "ไม้ดอกไม้ประดับ – decorative flowers and plants" composes of two other compound terms "ไม้ดอก – floral plants" and "ไม้ประดับ – decorative plants". Therefore, how to segment the conjugated text "ดอกไม้" again depends on the surrounding context. Third, many terms in Thai have the same form as a sentence, i.e., subject-verb-object. For example, scarf is "ผ้า-a piece of cloth พัน-wrapping around คอ-neck", napkin is "ผ้า-a piece of cloth เช็ด-wipe ปาก-mouth" and blanket is "ผ้า–a piece of cloth ห่ม-cover up". These problems directly affect *T1: topic and keyword extraction, T2: terminology extraction and T3: taxonomy extraction*, and obviously also impact other tasks. Note that there is no standard agreed-upon word segmentation guideline for Thai, even among the Thai NLP experts.

Thai language has high degree of ambiguity, both at word level and clause level. At word level, a Thai word usually has more than one meaning and plays more than one role. For example, a word "ดอก" can be a classifier, a noun, a verb, or a part of other words such as "ดอกไม้-flower", "ออกดอก-blooming", "บ่เป็นหยังดอก-no worries", etc. At clause level, there is verb versus adjective ambiguity. Both verbs and adjectives follow nouns, hence their roles are structurally in-distinguishable. Eg., a driver is "คนขับรถ person-drive-car", which can be a sentence by itself. In a complex noun, eg. a truck driver is "คนขับรถบรรทุก – person-drive-car-lade", when car-lade becomes "truck". But the word "lade-บรรทุก" can also mean "carry" used as a main verb. These problems affect both noun phrase and verb phrase extraction. In addition, it also makes it very difficult to determine the boundary of clauses or sentences, as well as to determine the underlying structure of a clause, both grammatically and semantically. Hence, word level and structural level ambiguity problems affect every task from T1 to T9.

There are also other particular features of Thai language that affect mostly T4: binary relation extraction, T7: supersense tagging, T8: semantic role labeling and T9: event extraction. Those problems are (i) no real helping verb and clue word, (ii) serial verbs, (iii) extended sentence, (iv) double-word phrase, and (v) zero anaphora. To briefly illustrate one example, A word "ถูก" which is put before a verb to indicate a passive voice sentence, also has many other meanings – such as cheap (adj), touch(v), occur to (v), right(v) and so on. A preposition ที่-at ใน-in บน-on sometimes becomes an adjective, an adverb, a verb, a noun or even a conjunction. This affects tasks that involve determining the real verb. The same problem exists for the serial verbs.

Note that the situation is somewhat different for task T5: named entity recognition. To identify a proper name such as a place or a person, there are patterns which can be predefined as templates. For example, a common pattern for a place can be *type of places + specific name of the place + possessive word (optional) + organization the place belongs to*.

4.5 Discussion

This section includes major aspects and findings from our analysis of Thai formal KE. First, we discuss the state-of-the-art of Thai KE research and provide reasons for the current status. Next, some general directions for improvement applicable to all tasks are proposed, and finally we address each KE task separately with a short analysis based on the Thai language literature survey and give concrete research ideas per task.

One of the main goals of this work is to assess the state-of-the-art of KE for Thai language. First of all, in a few areas there exist some very interesting and ambitious research works given the limited resources available, for example the work on Thai FrameNet [42], event extraction [34] or on S-Sense [26]. In general, as expected, the state-of-the-art is significantly lower as compared to KE from English text. There is a high level of agreement between the self-assessment of Thai researchers and the analysis of existing work done in this publication. Some of the KE tasks have already been addressed with high-quality approaches, but in most cases the approaches focus only on a specific domain or use case. Overall, tool support is missing, especially tools which are available publicly. There are various reasons for the comparably lower state-of-the-art: (i) the limited amount of research resources in Thailand, which is reflected in a small research community centered around a couple of organizations and research teams (see Sect. 4.1), (ii) the peculiarities of Thai written language which complicate NLP and KE considerably (details in Sect. 4.4), and (iii) like in any country, there are inefficiencies in the research culture, most notably in this case, by not making implementations and tools available.

Before discussing results and research ideas per task,

we want to focus on some general points which are applicable to all tasks, and will also be mentioned within the individual KE task suggestions. One obvious way to improve the state-of-the-art for any tasks is to adopt and transfer current methods being use for English language. Due to the different structure and particularities of Thai language this may not always be successful, but definitely a promising starting point. Many approaches used on English text are based on common machine learning techniques – in some cases training data for Thai language has to be created, in some cases existing resources can be leveraged. Even for pattern-based approaches the general ideas can be transferred, and the patterns adapted to the specifics of the language. And secondly, as mentioned, a main shortcoming of Thai KE research is the absence of high-quality and publicly available tools and datasets. We suggest a gradual shift in research culture where code, datasets and if useful even trained models are made available on platforms such as GitHub[†]. Publishing all results can also be made a requirement by Thai funding bodies for public funding of projects in the NLP and KE field.

In the remainder of this section each of the KE tasks is discussed separately regarding start-of-the-art and research gaps, and especially regarding practical research ideas.

(1) Task 1 – Keyword and topic extraction

For English, a plethora of methods, tools and APIs (both free and commercial) for keyword extraction, topic extraction and text classification are available. The methods include tf-idf, C-value, weirdness, LSA and LDA, various machine learning (ML) techniques, etc. Depending on the specific task, and the domain, users can choose from many options. Also for Thai language, there has been work on applying different techniques from fields such as information retrieval and information science, for example tf-idf, mutual information, information gain, LSA and LDA. Furthermore, various ML classifiers have been used for text and emotion classification. Although Thai research is comparably mature for task 1, as for any task, there is a lack of public tools and APIs.

To advance the state-of-the-art regarding tasks 1, we suggest a couple of research ideas and steps: (i) First of all, it would be helpful to make tools stemming from existing research available to the general public. As stated, a lot of strategies and methods have been already been experimented with, therefore it should be possible to provide the implementations used. Most of these tools will be based on processing tools for word segmentation and POS-tagging. (ii) Some existing approaches were using supervised ML. It will benefit future research and tool evaluation to make the training data (and maybe ML models) available in a well-organized format. (iii) A more ambitious, but interesting, project would be the extension of existing NLP and IE tools such as GATE with support for Thai language.

[†]https://github.com

(2) Task 2 – Terminology extraction

Regarding task 2, except for tagclouds in S-Sense [26], which does not include the technical details about the generation of the tag clouds, we did not find any detailed and in-depth work on terminology extraction. Obviously, the situation is different for English language, among recent work there is for example TBXtools by Oliver and Vazquez [47]. TBXtools is an open source automatic terminology extraction tool written in Python [††]. As most of the applied techniques in TBXtools are straightforward statistical or linguistic methods, it should be manageable to implement them also for Thai language. So as a first step we suggest to apply the approaches for n-gram detection and nesting detection on segmented and POS-tagged Thai text. Giving a thorough evaluation of these methods in various domains and providing an (open-source) implementation will be helpful to advance Thai terminology extraction.

(3) Task 3 – Taxonomy extraction

Although there has been few work on Thai text, the approaches by Kawtrakul et al. [24] and Imsombut and Kawtrakul [25] who use background knowledge to automatically learn Hearst-style patterns are interesting. Some state-of-the-art systems for taxonomy extraction for English still make heavy use of patterns, for example TAXI [48], the winner of the SemEval 2016 challenge on taxonomy extraction evaluation. As TAXI is designed to learn taxonomies in any given domain, it is language-independent, and an open-source implementation is available. It would be promising to apply it to Thai text, evaluate the system, and modify it to the specifics of Thai language.

An idea very different from the use of lexico-syntactic patterns is the application of word embeddings to taxonomy extraction. For English language, Rei and Briscoe [49] study various word embedding models and similarity measures for hyponym generation. A starting point for Thai can be the use of pretrained Thai word embedding models and the hypernym relations as defined in BabelNet[†††] as training data. Pre-trained word embeddings (word2vec and fastText) for Thai language are publicly available https://github.com/Kyubyong/wordvectors. BabelNet is a multilingual encyclopedic dictionary and integrates resources such as WordNet, Wikipedia, Wikidata and others. As BabelNet is such a rich resource and also contains Thai language terms, it is a great source for training data and background knowledge not only for this task, but also other KE tasks.

(4) Task 4 – Relation extraction

Modern systems for relation extraction apply sophisticated NLP-tools for preprocessing, most importantly NER tools and parsers (shallow or deep parsing) [50]. As such tools are not yet available for Thai text, there is no easy route

[††]https://sourceforge.net/projects/tbxtools

[†††]http://babelnet.org

to adopt those approaches for Thai text. But some of the older systems use methods which can be transferred more easily. For example, SnowBall [51] starts from known relation instances and learns text patterns to extract previously unknown instances. By using bootstrapping techniques the need for labeled training data or training patterns can be drastically reduced.

Supervised ML, esp. kernel-based methods, is very successful for closed relation extraction for English language. Given necessary training data (see above), it would be interesting to evaluate these techniques in various domains to determine their suitability for Thai text.

Finally, relation extraction is in many ways related to taxonomy extraction, but obviously the types of relations extracted are more general. Regardless, word embeddings and their analogy feature, ie. vector offset operations, are worth evaluating for Thai language text using pre-trained (or custom) embedding models. The advantage of this approach is that it doesn't need sophisticated NLP preprocessing.

(5) Task 5 – Name entity recognition

Name entity recognition is typically treated as a sequence labeling problem, such as POS-tagging, therefore methods like CRF or HMM have been applied successfully for English language for example by the Stanford NER tool[†]. Thai results for this task are comparably advanced, annotated corpora and classifiers exists, but are not readily available. So at first, and most important step, is to make the annotated corpora easily available, and especially to provide (open-source) tools to generate classification models, and also the resulting models. As soon as those tools are available, a next step will be to improve the accuracy by evaluating various ML techniques, potentially also deep learning methods. Improved NER results will also help other tasks such as relation extraction. NER holds a special importance in Thai language, as it is often interwoven with simple word segmentation. And finally, as for any supervised ML tasks the availability of high quality training data is crucial, therefore the creation of annotated corpora in various new domains will help improve the state-of-the-art. The creation of datasets is known to be a costly process, however the Thai community has already shown its capabilities in this area. As tools and tagsets for Thai NLP dataset creation already exist, the main challenge is the manual tagging itself. Paid microtask crowdsourcing has been shown to be an effective method to raise scalability and reduce cost in dataset creation and system evaluation within knowledge engineering [52], and also specifically NER [53]. To raise the quality of crowdsourcing annotations the process can be combined with expert corrections, especially for unclear cases.

(6) Task 6 – Name entity linking

A study by Chang et al. [54] shows that both simple popularity-based as well as classification-based methods can be applied to named entity linking successfully. For Thai language there already exists a classification-based approach for Wikification [41], so a first step would be to make this or similar systems available. Then those systems can be evaluated for specific use cases, and improved gradually.

(7) Task 7 – Supersense tagging

For task 7 we did not find any existing work on Thai language. Therefore it's obviously easy to advance the state-of-the-art by studying existing work in other languages and evaluating suitable concepts on Thai language. Although the AMALGrAM tool (see Sect. 3.2) seems to be rather complex, it will be interesting to port (parts of) it to Thai language and study the effectiveness. Another critical factor will be the creation of sufficient training data for Thai. As the creation of high-quality training data is a complex process, we suggest the reuse and adaption of the training methodology of Schneider and Smith [11], and the consideration of crowdsourcing techniques as mentioned in Task 5.

(8) Task 8 – Semantic relation labeling

Automatic semantic relation labeling is a very complex task, which usually involves sentence parsing, therefore the development of a high-accuracy system for Thai language is not an easy undertaking. However, first steps have been taken with the creation of a Thai FrameNet and around 2300 annotated sentences which can be used for training a classifier. We suggest to experiment with first prototypes using Thai FrameNet inspired by existing open-source tools existing for English language, such as Semafor[††].

(9) Task 9 – Event extraction

The methods for event detection in English scientific literature include statistical (ML-based), pattern-based and hybrid methods. Here, again, the absence of parsers for Thai language, is a hindrance. But there are pattern-based approaches for Thai to circumvent the problem [43]. To improve the state-of-the-art incrementally, Thai researchers can experiment with various pattern-based and statistical approaches in narrow domains and then extend and generalize the work.

We conclude this section with a quick summary of the proposed research ideas per KE task:

- (T1) Make already existing work available as open source, including existing datasets, models and training data
- (T1) Adoption of English language tools to Thai text
- (T1) Implementing support for existing multilingual IE frameworks such as GATE
- (T2) Port the statistical or linguistic methods from TBXtools to Thai text, evaluate the results, and provide a tool
- (T3) Adapt domain-independent pattern-based taxonomy extraction as in Taxi [48] to Thai text
- (T3) Apply word embeddings [49] using pre-trained

[†]https://nlp.stanford.edu/software/CRF-NER.shtml

[††]github.com/Noahs-ARK/semafor-semantic-parser

models and training relations eg. from BabelNet
- (T4) Make existing training data available. Extend where necessary (with domain experts or crowdsourcing)
- (T4) First steps with open relation extraction, by adopting techniques learning extraction patterns from existing instance data, and / or using bootstrapping methods.
- (T4) Evaluate kernel-based methods for relation extraction (classification) problems in various domains.
- (T4) Evaluate methods which leverage analogy operations with word embeddings
- (T5) Make existing corpora, tools and models available to the public (preferably as open source)
- (T5) Evaluate various supervised ML methods on the training data and improve the accuracy
- (T5) Create high quality annotated training corpora in various new domains
- (T6) Implement classification- and popularity based NEL methods and make the tools including the training datasets available
- (T6) Evaluate those approaches and steadily improve
- (T7) Study work in other languages and build first SST prototypes
- (T7) Create sufficient training and testing data
- (T7) Port open-source tools such as AMALGrAM to Thai language
- (T8) Develop first prototypes for SRL with Thai FrameNet data and use ideas from implementations for other languages
- (T9) Implement and evaluate statistical, pattern-based and hybrid approaches existing for English language

5. Conclusions

In this publication we evaluate the state-of-the-art of Thai research in formal knowledge extraction, and we provide an extensive list of research gaps and practical research ideas in order to improve the current status. First, inspired by previous research on knowledge extraction (KE) from English language [2] we distinguish nine specific KE tasks, and assess their state-of-the-art with a review of existing work. A questionnaire-based survey helps to collect data for a self-assessment by the research community, as well as pointers to additional work. Furthermore, queries to literature databases are used to analyze the structure and size of the Thai research community.

The contributions of this work are as follows: (i) Providing a detailed survey of the existing work on formal KE from Thai language text; (ii) Assessing the state-of-the-art for nine predefined KE tasks, both via analysis of scientific work and by self-assessment of the Thai research community, including our interpretation of reasons for the current state; (iii) Providing insights into the structure of the Thai research community, and comparing it to the Japanese community in order to estimate the community size and community focus; (iv) And finally, and most importantly, we combine the insights from contributions (i)–(iii) to identify research gaps, and provide an extensive list of research ideas. The focus is on practical and concrete ideas which can be implemented and evaluated with foreseeable effort, and which help to improve the state-of-the-art in Thai KE step by step.

Acknowledgments

This work is funded by the Academic Melting Pot Program from King Mongkut's Institute of Technology Ladkrabang (KMITL) for the fiscal year 2017. Also, this work was supported by the Government of the Russian Federation (Grant 074-U01) through the ITMO Fellowship and Professorship Program.

References

[1] A. Weichselbraun, G. Wohlgenannt, and A. Scharl, "Evidence sources, methods and use cases for learning lightweight domain ontologies," Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances, ed. W. Wong, W. Liu, and M. Bennamoun, pp.1–15, IGI Global, Hershey, PA, 2011.

[2] A. Gangemi, "A comparison of knowledge extraction tools for the semantic web," The Semantic Web: Semantics and Big Data, Proc. 10th Intern. Conference, ESWC 2013, Montpellier, France, May 26–30, 2013, ed. P. Cimiano, Ó. Corcho, V. Presutti, L. Hollink, and S. Rudolph, Lecture Notes in Computer Science, vol.7882, pp.351–366, Springer, 2013.

[3] N. Tongtep and T. Theeramunkong, "Simultaneous character-cluster-based word segmentation and named entity recognition in thai language," Proc. 5th Intern. Conference on Knowledge, Information, and Creativity Support Systems, KICSS'10, Berlin, Heidelberg, vol.6746, pp.216–225, Springer-Verlag, 2011.

[4] N. Tongtep and T. Theeramunkong, "Discovery of predicate-oriented relations among named entities extracted from thai texts," IEICE Transactions, vol.95-D, no.7, pp.1932–1946, 2012.

[5] C. Haruechaiyasak, S. Kongyoung, and C. Damrongrat, "Learn-lexto: A machine-learning based word segmentation for indexing thai texts," Intern. Conference on Information and Knowledge Management, CIKM 2008, pp.85–88, Napa Valley, CA, Oct. 2008.

[6] T. Supnithi, K. Kosawat, M. Boriboon, and V. Sornlertlamvanich, "Language sense and ambiguity in Thai," 8th Pacific Rim Int. Conf. Artificial Intell. (PRICAI), ed. C. Zhang, H.W. Guesgen, and W.K. Yeap, LNCS, Springer, Aug. 2004.

[7] D.Q. Nguyen, D.Q. Nguyen, D.D. Pham, and S.B. Pham, "A robust transformation-based learning approach using ripple down rules for part-of-speech tagging," AI Communications, vol.29, no.3, pp.409–422, 2016.

[8] P. Intarapaiboon, E. Nantajeewarawat, and T. Theeramunkong, "Extracting semantic frames from thai medical-symptom unstructured text with unknown target-phrase boundaries," IEICE Transactions, vol.94-D, no.3, pp.465–478, 2011.

[9] A. Kawtrakul and P. Praneetpolgrang, "A history of AI research and development in thailand: Three periods, three directions," AI Magazine, vol.35, no.2, pp.83–92, 2014.

[10] A. Gangemi, V. Presutti, D. Reforgiato Recupero, A.G. Nuzzolese, F. Draicchio, and M. Mongiovì, "Semantic Web Machine Reading with FRED," Semantic Web, vol.8, no.6, pp.873–893, 2017.

[11] N. Schneider and N.A. Smith, "A corpus and model integrating multiword expressions and supersenses," Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, pp.1537–1547, Association for Computational

Linguistics, May–June 2015.
[12] S. Dei Rossi, G. Di Pietro, and M. Simi, "Description and Results of the SuperSense Tagging Task," vol.7689, pp.166–175, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
[13] C.F. Baker, C.J. Fillmore, and J.B. Lowe, "The berkeley framenet project," Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th Intern. Conference on Computational Linguistics - Volume 1, ACL '98, Stroudsburg, PA, USA, pp.86–90, Association for Computational Linguistics, 1998.
[14] C. Haruechaiyasak, P. Srichaivattana, S. Kongyoung, and C. Damrongrat, "Automatic thai keyword extraction from categorized text corpus," Proceedings of ECTI-CON 2004, 13–14 May, 2004.
[15] P. Jotikabukkana, V. Sornlertlamvanich, O. Manabu, and C. Haruechaiyasak, "Social media text classification by enhancing well-formed text trained model," Journal of ICT Research and Applications, vol.10, no.2, pp.177–196, 2016.
[16] N. Chirawichitchai, P. Sa-nguansat, and P. Meesad, "A comparative study on feature weight in thai document categorization framework.," IICS, ed. G. Eichler, P.G. Kropf, U. Lechner, P. Meesad, and H. Unger, LNI, vol.165, pp.257–266, GI, 2010.
[17] S. Viriyavisuthisakul, P. Sanguansat, P. Charnkeitkong, and C. Haruechaiyasak, "A comparison of similarity measures for online social media thai text classification," ECTI-CON 2015 - 12th Intern. Conf. Electrical Engineering/Electronics, Computer, Telecommunications and IT, Hua Hin, Thailand, pp.1–6, 2015.
[18] V. Lertnattee and T. Theeramunkong, "Text classification for thai medicinal web pages," 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2007, pp.631–638, Springer, Lecture Notes in Computer Science, Nanjing, China, 2007.
[19] P. Daowadung and Y.-H. Chen, "Using word segmentation and svm to assess readability of thai text for primary school students," Proc. 2011 8th Intern. Joint Conf. CS and Softw. Engin., JCSSE 2011, Nakhon Pathom, Thailand, pp.170–174, 2011.
[20] N. Chirawichitchai, "Emotion classification of thai text based using term weighting and machine learning techniques," 11th Int. Joint Conf. Computer Science and Software Engineering: Human Factors in Computer Science and Software Engineering: eHPC, JCSSE, Pattaya, Chonburi, Thailand, pp.91–96, 2014.
[21] P. Inrak and S. Sinthupinyo, "Applying latent semantic analysis to classify emotions in thai text," ICCET 2010 - Proc. 2010 Intern. Conference on Computer Engineering and Technology, Chengdu, China, pp.V6450–V6454, April 2010.
[22] R. Jiamthapthaksin, "Thai text topic modeling system for discovering group interests of facebook young adult users," 2016 2nd Intern. Conference on Science in Information Technology (ICSITech), pp.91–96, Oct. 2016.
[23] T. Kiatdarakun and P. Suksompong, "Entropy rate of thai text and testing author authenticity using character combination distribution," 2nd Intern. Conference on Digital Information and Communication Technology and its Applications, DICTAP 2012, Bangkok, Thailand, pp.492–497, 2012.
[24] A. Kawtrakul, M. Suktarachan, and A. Imsombut, "Automatic thai ontology construction and maintenance system," Workshop on Papillon 2004, 2004.
[25] A. Imsombut and A. Kawtrakul, "Automatic building of an ontology on the basis of text corpora in thai," Language Resources and Evaluation, vol.42, no.2, pp.137–149, 2008.
[26] C. Haruechaiyasak, A. Kongthon, P. Palingoon, and K. Trakultaweekoon, "S-sense: A sentiment analysis framework for social media sensing," Proc. IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP), Nagoya, Japan, pp.6–13, Oct. 2013.
[27] V. Sornlertlamvanich and C. Kruengkrai, "Effectiveness of Keyword and Semantic Relation Extraction for Knowledge Map Generation," vol.9442, pp.188–199, Springer Intern. Publishing, Cham, 2016.
[28] A. Imsombut and C. Sirikayon, "An alternative technique for populating thai tourism ontology from texts based on machine learning," 2016 IEEE/ACIS 15th Intern. Conference on Computer and Information Science (ICIS), Los Alamitos, CA, USA, pp.1–4, IEEE Computer Society, 2016.
[29] K. Kowsrihawat and P. Vateekul, "An information extraction framework for legal documents: A case study of thai supreme court verdicts," Proc. 2015 12th Intern. Joint Conference on Computer Science and Software Engineering, JCSSE 2015, Hatyai, Thailand, pp.275–280, July 2015.
[30] A. Kawtrakul, C. Pechsiri, T. Permpool, D. Thamvijit, P. Sornprasert, C. Yingsaeree, and M. Suktarachan, "Ontology driven k-portal construction and k-service provision," Proc. LREC2006 Conference, Genoa, Italy, pp.775–779, 2006.
[31] A. Kongwan and A. Kawtrakul, "Know-what: A development of object-property extraction from thai texts and query system," Proc. 6th Symposium on NLP, Chiang Rai, 2005.
[32] S. Sitthisarn and B. Bahoh, "Towards Automatic Semantic Annotation of Thai Official Correspondence: Leave of Absence Case Study," vol.265, pp.273–282, Springer Intern. Publishing, Cham, 2014.
[33] N. Tongtep and T. Theeramunkong, "A feature-based approach for relation extraction from thai news documents," Intelligence and Security Informatics, Pacific Asia Workshop, PAISI 2009, Bangkok, Thailand, April 27, 2009. Proceedings, ed. H.C. et al., Lecture Notes in Computer Science, vol.5477, pp.149–154, Springer, 2009.
[34] P. Intarapaiboon, E. Nantajeewarawat, and T. Theeramunkong, "Extracting chemical reactions from thai text for semantics-based information retrieval," IEICE Transactions on Information and Systems, vol.E94-D, no.3, pp.479–486, 2011.
[35] P. Sutheebanjard and W. Premchaiswadi, "Thai personal named entity extraction without using word segmentation or pos tagging," 2009 Eighth Intern. Symposium on Natural Language Processing (SNLP 2009), pp.221–226, Oct. 2009.
[36] T. Vichayakitti and C. Jaruskulchai, "Automatic temporal event recognition from thai news," IEEE Intern. Symposium on Communications and Information Technology, 2005. ISCIT 2005., pp.938–942, Oct. 2005.
[37] N. Tirasaroj and W. Aroonmanakun, "The effect of answer patterns for supervised named entity recognition in Thai," Proc. 25th Pacific Asia Conference on Language Information and Computing (PACLIC 25), pp.392–399, 16–18 Dec. 2011.
[38] T. Theeramunkong, M. Boriboon, C. Haruechaiyasak, N. Kittiphattanabawon, K. Kosawat, C. Onsuwan, I. Siriwat, T. Suwanapong, and N. Tongtep, "Thai-nest: A framework for Thai named entity tagging specification and tools," Proc. CILC (2010), CILC 2010, pp.895–908, 2010.
[39] A. Aw, S.A. Mahani, N. Lertcheva, and S. Kalunsima, "Talapi – A thai linguistically annotated corpus for language processing," Proc. Ninth Intern. Conf. Language Resources and Evaluation, LREC Reykjavik, Iceland, pp.125–132, ELRA, 2014.
[40] N. Tongtep and T. Theeramunkong, "Multi-stage automatic NE and pos annotation using pattern-based and statistical-based techniques for thai corpus construction," IEICE Transactions, vol.96-D, no.10, pp.2245–2256, 2013.
[41] A. Rungsawang, S. Siangkhio, A. Surarerk, and B. Manaskasemsak, "Thai Wikipedia Link Suggestion Framework," vol.253, pp.365–373, Springer Netherlands, Dordrecht, 2013.
[42] D. Leenoi, S. Jumpathong, P. Porkaew, and T. Supnithi, "Thai framenet construction and tools," Int. J. Asian Lang. Proc., vol.21, no.2, pp.71–82, 2011.
[43] P. Intarapaiboon, E. Nantajeewarawat, and T. Theeramunkong, "Information extraction from thai text with unknown phrase boundaries," 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2009, vol.5476, pp.525–532, Lecture Notes in Computer Science, Bangkok, Thailand, 2009.
[44] C. Haruechaiyasak, S. Kongyoung, and M. Dailey, "A comparative study on thai word segmentation approaches," 5th Intern. Conference on Electrical Engineering/Electronics, Computer, Telecom-

munications and Information Technology, 2008. ECTI-CON 2008, pp.125–128, 14–17 May 2008.
[45] M. Murata, Q. Ma, and H. Isahara, "Part of speech tagging in thai language using support vector machine," CoRR, vol.cs.CL/0112004, 2001.
[46] J. Pailai, R. Kongkachandra, T. Supnithi, and P. Boonkwan, "A comparative study on different techniques for thai part-of-speech tagging," The 10th Int. Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp.1–5, 2013.
[47] A. Oliver and M. Vázquez, "Tbxtools: A free, fast and flexible tool for automatic terminology extraction," RANLP, pp.473–479, 2015.
[48] A. Panchenko, S. Faralli, E. Ruppert, S. Remus, H. Naets, C. Fairon, S.P. Ponzetto, and C. Biemann, "TAXI at SemEval-2016 Task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling," Proc. 10th Intern. Workshop on Semantic Evaluation, Association for Computational Linguistics, pp.1320–1327, 2016.
[49] M. Rei and T. Briscoe, "Looking for hyponyms in vector space," Proc. Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26–27, 2014, ed. R. Morante and W. Yih, pp.68–77, ACL, 2014.
[50] J. Schmidek and D. Barbosa, "Improving open relation extraction via sentence re-structuring," Proc. Ninth Intern. Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26–31, 2014, pp.3720–3723, 2014.
[51] E. Agichtein and L. Gravano, "Snowball: extracting relations from large plain-text collections," Proc. fifth ACM Conf. Digital libraries, DL '00, New York, NY, USA, pp.85–94, ACM, 2000.
[52] G. Wohlgenannt, M. Sabou, and F. Hanika, "Crowd-based ontology engineering with the uComp Protégé plugin," Semantic Web Journal (SWJ), vol.7, no.4, pp.379–398, 2016.
[53] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, May 2014.
[54] A. Chang, V.I. Spitkovsky, C.D. Manning, and E. Agirre, "A comparison of named-entity disambiguation and word sense disambiguation," Proc. Tenth Intern. Conference on Language Resources and Evaluation (LREC 2016), Paris, France, May 2016.

Ponrudee Netisopakul received a bachelor degree in Statistics from Chulalongkorn University in 1989. She was one of the scholars chosen by the Royal Thai Government to study abroad and received her two Master Degrees and a Ph.D. in Computer Science and Information Science from University of Southern California, University of Delaware and Case Western Reserve University in 1994, 1997 and 2002, respectively. She is currently an Associate Professor at the Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. Her current research interests include knowledge engineering, natural language processing, software measurement and data analysis.

Gerhard Wohlgenannt is holding a fellowship position as research professor at ITMO university in St. Petersburg, Russia, where he is working at the "International laboratory of Information Science and Semantic Technologies". Before, Gerhard Wohlgenannt was an assistant professor at the Institute of Information Business of the Vienna University of Economics and Business. He completed his PhD thesis in the field of ontology learning 2010 and in July 2016, he received his habilitation degree (venia docendi) in Business Informatics. His research interests are ontology learning, crowdsourcing and knowledge extraction from text.