Using Conceptual Graph to Represent Semantic Relation of Thai Facebook Posts in Marketing

Kwanrutai Saelim
Department of Computer Science, Faculty of Science and Technology, Thammasat University
Pathumthanee, Thailand
kwanrutai.saclim@gmail.com

Taweewat Luangwiriya
Department of Computer Science, Faculty of Science and Technology, Thammasat University
Pathumthanee, Thailand
taweewatlu@gmail.com

Rachada Kongkachandra
Department of Computer Science, Faculty of Science and Technology, Thammasat University
Pathumthanee, Thailand
rdk@cs.tu.ac.th

***Abstract*—Conceptual Graph is a representation commonly used to express semantic relationship of natural language. This work presents a method to translate Thai natural language text to conceptual graphs regarding semantic relations based on semantic roles between predicate and its arguments. Shallowing parsing of Thai text and verb patterns as case frames are utilised in identifying core entities in a context and their semantic roles. Then, the argument with annotated roles are translated into conceptual graphs that are able to logically and visually represent relations of core terms. As a result, conceptual graphs of Thai natural texts from Facebook posts in a marketing group were generated. In the study, found issues regarding Thai specific natural style are encountered and discussed.**

***Index Terms*—Semantic Relation, Conceptual Graph, Information Extraction, Semantic Role**

I. Introduction

A semantic role (also known as thematic role or theta role) is the studying of underlying relationship of entities involved with the main predicate in a clause [1] . Semantic roles are an attempt to capture similarities and differences in verb meaning with generalizations that contribute to the mapping from semantics to syntax. Analysing semantic role is a basic step to understand semantic of context and gives comprehensive power to a computer system. Semantic role labelling (SRL, shallow semantic parsing) [2] is the computational process to analyse and assign labels to words or phrases in a clause for indicating their semantic role. It is a challenging task studied for decades. The clause (or sentence) level semantic analysis of text is concerned with the characterisation of events including ”who” does ”what” to ”whom” with ”whom” at ”where” and ”when” [3]. The main task of SRL is to identify semantic relations from a predicate and its associated participants (textual entities). The SRL are labels to let know about action each clause. The SRL have labels indicate each words in the sentence such as Agent, Patient, Location of other entities and Temporal of event each clause.

Shallowing Parsing is an analysis of a sentence which analyze using components of sentence structure. Therefore, Shallow parsing will be done after Part-of-speech tagging(POS). Part-of-speech tagging is the process of marking up a word in a text in the sentence(nouns, verbs, adjectives, adverbs, etc.) [4]. The POS procedure we get results is that smaller units. There are called chunks. Shallow parsing combines units into larger units and that have grammatical meanings. So we get phrases, phrases can reflect the relationship between the meanings of the basic components. And when the Shallowing Parsing process is completed, the results will be next step is to the Semantic Role Labeling process as follow Figure 1. When receiving the SRL result the final step is to create with Conceptual Graph as follow Figure 2.

A conceptual graph (CG) is a graph representation the meaning which reply knowledge of logic based on the semantic networks. Which a conceptual graph one of artificial intelligence. The research of conceptual graph have explored the techniques for reasoning, knowledge representation from natural language into understanding natural language semantics. The semantics of the important base of the conceptual graph and conceptual graph is defined by a concept mapping to and from the ISO standard for Common Logic [5]. Mathematic is the important role of knowledge one in artificial intelligence (AI), Knowledge representation, their formal languages were not sufficiently expressive. Therefore the possibility of automated reasoning from the perspective of knowledge representation they were various in the use

Identify applicable funding agency here. If none, delete this.

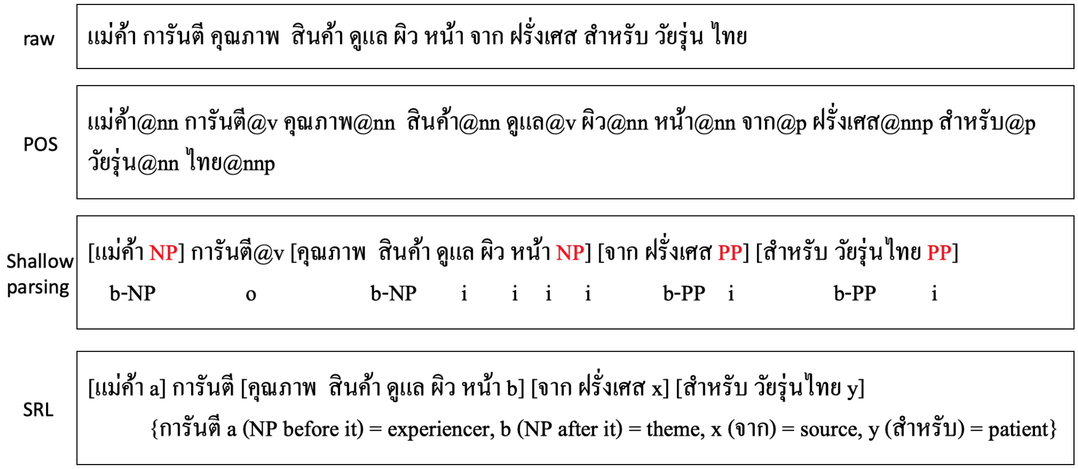

Figure 1. All step to get semantic labelling.

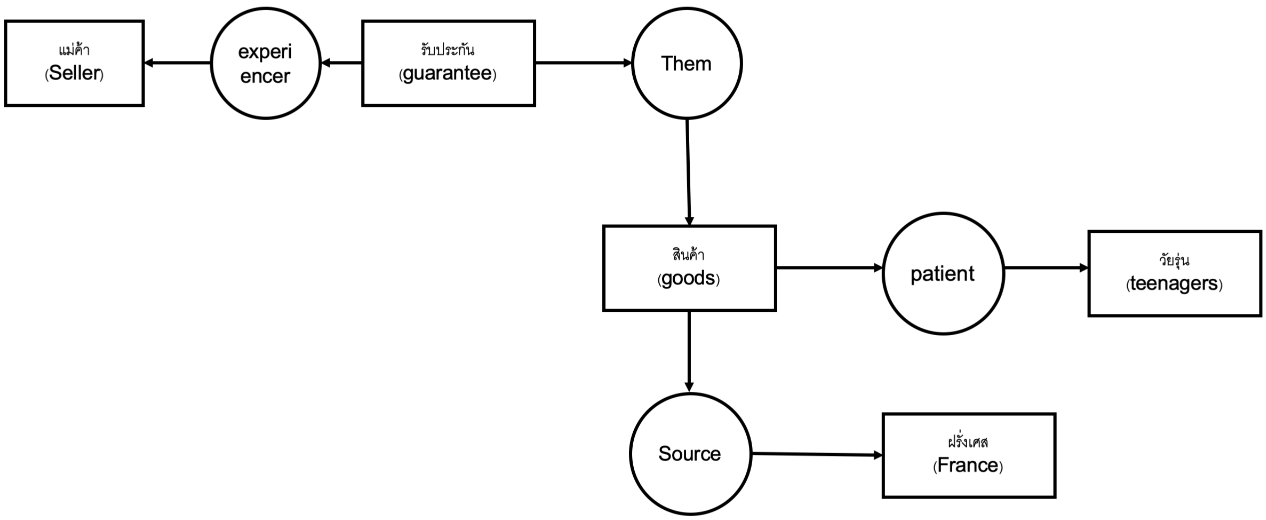

Figure 2. Examples of relations and a conceptual graph from marketing posts

of syntactic constructs. knowledge representation to explain representing information in a form that a computer system can utilize to solve complex tasks. Knowledge representation and findings from logic to automate various kinds of reasoning the application of rules or the relations of sets and subsets.

In the past, there are several representations for labelling semantic roles, such as ”case frame” [1] [6] , FrameNet [7] and Conceptual Graph (CG) [8] . The CG is the most convenient on visualization and comprehensible while the FrameNet is designed to collect rich linguistic details. The most challenging point in SRL is that there is no specified set of roles that generalise for all contextual events. Moreover, ambiguity from natural language including subject-object omission and multi-meaning terms could render the sentence to non-analysable. For Thai language, very few studied and reported on semantic role labelled corpus [9] and Thai FrameNet as a resource for semantic analysis [10] . Their works were specified for a target domain such as agricultural context and tourist context in written Thai documents. Differently, the language used in social network, a major digital text resource nowadays, is more natural and vigorously stylish. Analysing of the semantic relation of words in social media is thus tough but necessary step towards understanding of Thai natural language.

In this work, we aim to create a framework for using conceptual graph to represent Thai social network posts. This work attempts to provide a standard framework to analyse and label Thai natural text with semantic role sets. The syntactic parser is applied to identify a predicate and its related entities.

II. Background

A. Semantic Role Labelling

There are many researches on manually created lexical and semantic resources as a lexical resource for natural language processing. Those works are guaranteed for their accuracy with the draw-back of labour-intensive task and restricted specific domains. For Thai, Suktarachan et al. presented a construction of Thai concept frames applied in language processing for agricultural domain based on the verb centric approach resulting in 5,784 Thai sentences annotated with POS and semantic roles. This is a good resource for semantic analysing for Thai text but apparently specified for agriculture documents. Their case frames and semantic roles however are adoptable as a standard for semantic role labelling task.

Semantic role labelling (SRL) is a task to identify the latent predicate argument structure of a clause/sentence, providing representations that answer basic questions about sentence meaning, such as ”who” does ”what” to ”whom”. General roles used in SRL are labels such as Agent, Patient, and Location for the entities participating in an event with temporal and manner details. These labels therefore provide a ground-level of semantic relation representation of the text. There are commonly used semantic roles [11] [12] as follows Figure 3.

- Agent: The ”doer” or instigator of the action denoted by the predicate.
- Patient: The ”undergoer” of the action or event denoted by the predicate.
- Theme: The entity that is moved by the action or event denoted by the predicate.
- Experiencer: The living entity that experiences the action or event denoted by the predicate.
- Goal: The location or entity in the direction of which something moves.
- Benefactive: The entity that benefits from the action or event denoted by the predicate.
- Source: The location or entity from which something moves
- Instrument: The medium by which the action or event denoted by the predicate is carried out.
- Locative: The specification of the place where the action or event denoted by the predicate in situated.

B. Conceptual Graph

A conceptual graph (CG) is a graph representation for logic based on the semantic networks [2]. The CG has been not only used to represent natural language semantics, but also knowledge representation and reasoning. In 1976, Sowa [2] proposed to use conceptual graphs (CGs) as an intermediate language for transforming natural language texts to machine-readable graph form. For example, CG of the sentence ”Kids went to Bangkok by train” is illustrated in Figure 3 3. In the CG, the rectangles refer to concepts representing entities of the text while the circles called conceptual relations are used to denote relations between concepts. Conceptual relations usually are based on the semantic role relation (see Section 2.1). An arc pointing toward a circle signifies the first argument of the relation, and an arc pointing away from a circle signifies the last argument.

CGs normally keeps the core information with concrete meaning of the text for simplification and ability of easy-to-comprehension; thus, not all the entities in the input text are kept. There are several researches on how to make use of CGs such as Question-answering, Diagrammatic reasoning, Entity-relationship model, Semantic web, etc.

III. Methods

We design a three-step approach for generating a conceptual graph representation of Thai texts. Same to other Thai natural processing tasks, pre-processing is required to handle Thai term-boundary and remove non-text entities such as emoticons and symbols. First, we identify the syntactic roles in a sentence using shallow syntactic parser. Second, we design a set of syntactic rules to semantic roles. Third, we construct a conceptual graph following the assigned semantic roles. An overview architecture of the system is drawn in Figure 4 1.

In this work, the focused natural Thai texts are the Facebook posts about marketing. The posts related to product-selling advertisement and good details are collected from a marketing group. The target posts are text-based explanation excluding images and digital emotion expressions, i.e. emoticons and stickers. For pre-processing step, word segmentation is necessary to mark words ”boundary”. The preferable boundary approach of segmentation is a concept level since the entities should be understandable in themselves and do not need to combine for completing a word sense.

1) Firstly, for each clause in the sentence, we identify the main verb and build a sentence pattern using the parsed tree.

2) Secondly, for each verb in the sentence we extract a list of possible semantic frames from VerbNet, together with restrictions for each semantic role.

3) Thirdly, we match the sentence pattern to each of the available semantic frames, considering the semantic role's constraints. As a result, we are presented with a list of all

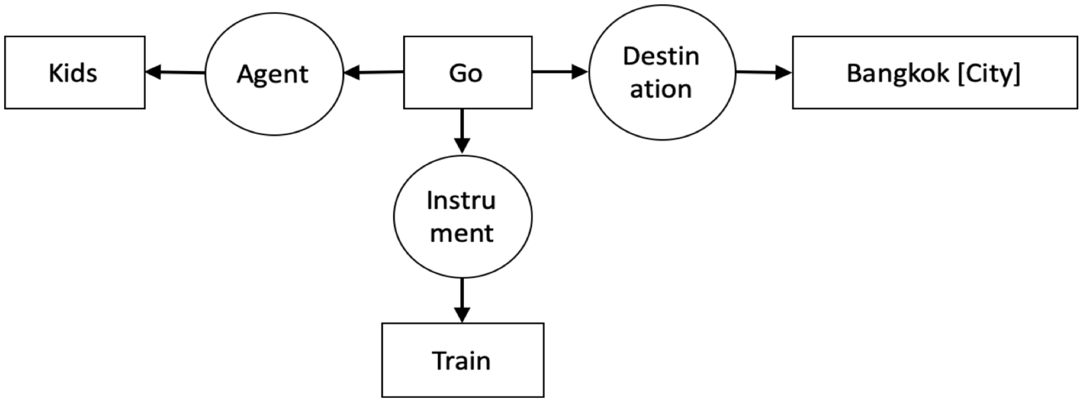

Figure 3. Conceptual graph representing a sentence "Kids went to Bangkok by train"

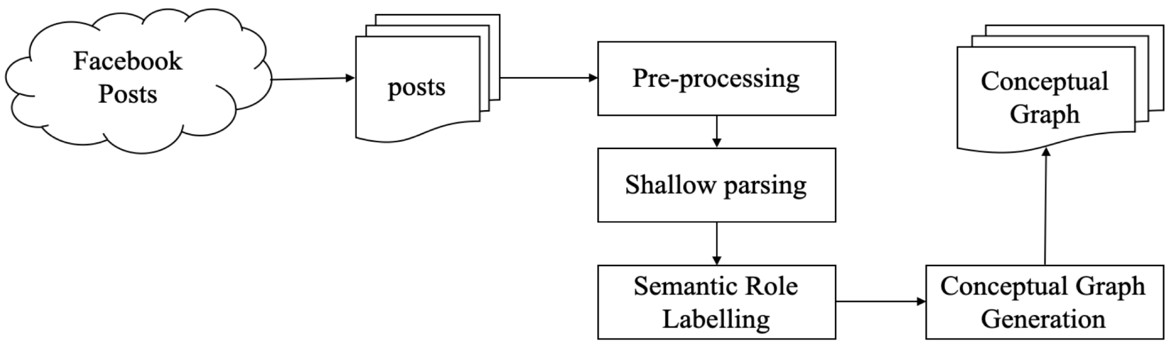

Figure 4. An overview architecture of Thai text based conceptual graph generation.

possible semantic role, s assignments, from which we have to identify the correct one.

A. *Shallow Parsing for Detecting Clausal Core Entities*

In SRL, the main entity to connect other entities is a predicate (typically a main verb of a clause) while the other entities such as a subject of the predicate and a direct object is handled as arguments of the predicate with roles. Thus, detection of these entities is undoubtedly essential for later processes. In fact, the other entities that represent a little to none meaning in the text, such as interjections and politeness ending markers, are ignored in this process to reduce complexity. A shallow parser is thus applied to separate an input text into phrases including Noun phrase (NP), Prepositional phrase (PP) and Verb phrase (VP). The process is conducted from left-to-right manner following a paradigm of Thai word composition. Unlike a normal syntactic parser, this shallow parsing only handles the VP for consecutive verbs (Serial Verbs) that has no NP and PP among them. In a case that NP belongs to a preposition, they will be grouped as PP while the NP with PP attached is counted separately as two distinct entities. The demonstration of the parsing results is given in Figure 5 4.

A result of this process is phrasal chunks of Thai text. Please be noted that although shallow parsing in general has a fine accuracy result, shallow parsing for Thai is still a difficult and complex task from a nature of Thai in which is ambiguous and semantically implicit. Hence, the output chunks are required for manual post-editing to be reliably usable.

B. *Semantic Role Identification*

The algorithm for semantic role identification of a sentence that we propose consists of the following three steps:

1) Firstly, for each clause in the sentence, we identify the main VP and build a sentence pattern using the heuristic rules.

2) Secondly, for each verb in the sentence, we extract a list of possible semantic frames from VerbNet, together with restrictions for each semantic role.

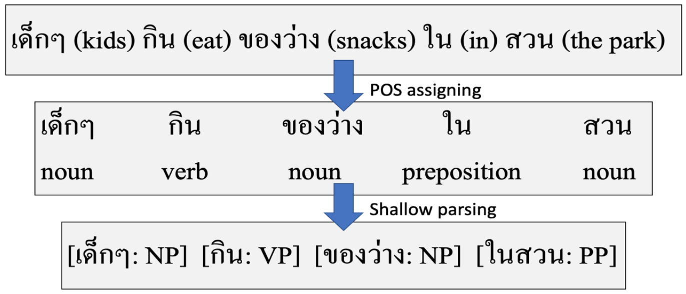

Figure 5. A demonstration of Thai shallow parsing

Verb	Preceding NP	Following NP	Related PP
"ซื้อ" (by)	Agent	Patient	"จาก"(from) – source
			"ที่" (at) –,locative
"ส่ง" (send)	Agent	Patient	"จาก" (from) – source
			"ถึง" (to) – destination
			"ที่" (at) –,locative
			"ด้วย" "โดย" (with) – instrument
"เห็น" (see)	Experiencer	Theme	"ด้วย, โดย" (with) – instrument
			"จาก" (from) – source

3) Thirdly, we match the pattern to each of the available semantic frames, considering the semantic role constraints. As a result, we are presented with a list of all possible semantic role assignments, from which we have to identify the correct one.

The rules for building pattern used in the step 1 mainly considers the verbs in a context. According to Thai serial verb construction, auxiliary verbs including temporal auxiliary (tense marker such as "กำลัง" (progressive) and "แล้ว" (past)), probability auxiliary (such as "น่าจะ" (should) and "อาจจะ" (may)), and directional auxiliary (such as "ขึ้น" (upward) and "ไป" (outward)) are ignored unless there are no other verbs. The remaining verbs with semantic meaning are treated equally. For the VerbNet, we collect Thai verbal information along with restriction to identify semantic relationship to other entities. We exemplify the information as shown in Table 1. For the current information, we have only collected verbs from the gathered corpus, but we plan to expand the database from other sources.

C. Conceptual Graph Generation

With annotated semantic roles, a conceptual graph can be generated accordingly. The CG as a whole is kept in a formula using logical expression. For an example from the context in Equation 1, the sentence can be formulated into the following formula.

$$(\exists x)(\exists y)(go(x) \wedge kids) \wedge City(Bangkok) \wedge train(y) \wedge$$

$$Agent(x, kids) \wedge Destination(x, Bangkok) \wedge Instrument(x, y))$$

As the exemplified formulation, the logical operators are conjunction and the existential quantifier. Those two operators are the most common in translations from natural languages. The formulae thus can be linked to form more detailed relationship as shown in Figure 6 5.

With the conceptual graph and its logical reasoning, the obtained information can be utilised to classify sellers together based on their predicates such as type of products, locations, and marketing manners. Furthermore, this information could be used as knowledge base for further AI based application and data analytics.

IV. Discussions

In generating CGs from Thai Facebook posts on marketing, we encountered several issues. We found that the natural Thai textual expressions were difficult to handle perfectly. Firstly, 27 % of the inputs lack their subject part in which leads to incomplete graph. This circumstance in fact is a usual style of Thai natural language that typically causes ambiguity and incorrectness in automated processing. In this work, we handled the issue by analysing the verb and decide to add the subject part manually. There were two common solutions as 1) adding the post owner name as the subject of the predicate, and 2) considering NP previously mentioned in prior sentence as a subject of the predicate. We found that about 85 % of the cases were handled with the former solution. Secondly, many emerging terms or jargons were used with specific meaning such as "บ่องตง" (transformation of "'บอกตรง" (frankly speak)) and "ตะมุตะมิ" (stylish euphemism referring to "being cute").

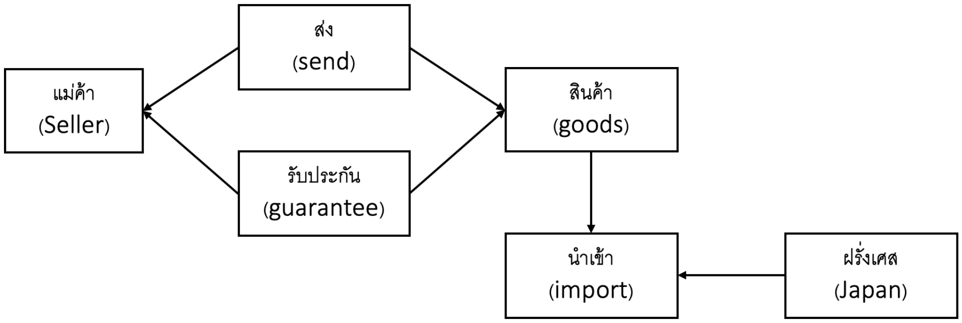

Figure 6. Examples of relations and concepts from marketing posts

These terms were used sparingly in many contexts in which requires specific design of patterns. Thirdly, the confusion between Thai serial verb construction and running causes. Thai language does not have an explicit marker for cause boundary. With the omission of subject part and continuous written text, verbs can be placed adjacently and lead to ambiguity in both shallow parsing and identifying semantic roles. For instance, Thai text as "ลูกค้า" (customer: NP) "สนใจ" (interesting: verb) "ชิ้น ไหน" (which product: NP) "อินบอค" (inbox: verb) "เลย" (mood marker: ignore in shallow parsing)" should be analysed for two main verbs sharing the subject "customer". However, the pattern can be confused that the second verb might take the object of the first argument as its own subject. This issue though can be solved if there is a dictionary to help declaring a concept of the noun as it is an object and cannot be considered as an Agent for the predicate.

V. Conclusions

This work presents a method to translate Thai modern natural texts to conceptual graph via the use of semantic role identification. The proposed method requires the shallow parsing to detect the predicate of the clause/sentence and verb information to identify semantic roles of predicate arguments. With the boundless style of Thai natural language, the design of processes is considerate of incomplete sentence and being flexible. In an attempt to generate conceptual graph, several issues leading to failure were found and discussed. The generated conceptual graphs can be used to classify posts according to predicates and their arguments such as type of products, locations, and marketing manners. Furthermore, this information could be used as knowledge base for further AI based application and data analytics.

References

[1] S. Padó and M. Lapata, "Cross-lingual annotation projection for semantic roles," *Journal of Artificial Intelligence Research*, vol. 36, pp. 307–340, 2009.

[2] J. F. Sowa, "Cognitive architectures for conceptual structures," in *International Conference on Conceptual Structures*. Springer, 2011, pp. 35–49.

[3] T. Cohn and P. Blunsom, "Semantic role labelling with tree conditional random fields," in *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2005, pp. 169–172.

[4] G. Neumann and J. Piskorski, "A shallow text processing core engine," *Computational Intelligence*, vol. 18, no. 3, pp. 451–476, 2002.

[5] J. Sowa, "Chapter 5 conceptual graphs," *Foundations of Artificial Intelligence*, vol. 3, 12 2008.

[6] C. J. Fillmore, "The case for case." 1967.

[7] D. Leijen, W. Schulte, and S. Burckhardt, "The design of a task parallel library," vol. 44, 10 2009, pp. 227–242.

[8] S. Hensman, "Construction of conceptual graph representation of texts," 01 2004.

[9] Z. Lin, Y. Duan, Y. Zhao, W. Sun, and X. Wan, "Semantic role labeling for learner chinese: the importance of syntactic parsing and l2-l1 parallel data," *arXiv preprint arXiv:1808.09409*, 2018.

[10] D. Leenoi, S. Jumpathong, and T. Supnithi, "Building thai framenet through a combination approach," in *2010 International Conference on Asian Language Processing*, Dec 2010, pp. 277–280.

[11] C. J. Fillmore and C. F. Baker, "Frame semantics for text understanding," in *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, 2001.

[12] C. J. Fillmore and C. Baker, "A frames approach to semantic analysis," in *The Oxford handbook of linguistic analysis*, 2010.