# A Framework for Constructing a Thai Medical Knowledge Base

**Thanaruk Theeramunkong Pichai Iamtana-anan Cholwich Nattee**
**Arthit Suriyawongkul Ekawit Nantajeewarawat Pakinee Aimmanee**

School of Information and Computer Technology
Sirindhorn International Institute of Technology
131 Moo 5 Tiwanont Road Bangkadi Muang Pathumthani 12000
thanaruk@siit.tu.ac.th, pichai.i@hotmail.com, cholwich@siit.tu.ac.th,
art@siit.net, ekawit@siit.tu.ac.th, pakinee@siit.tu.ac.th

**Abstract**

This paper presents a framework for constructing a medical knowledge base in Thailand, including its progress. The framework covers data acquisition, keyword extraction, link construction, and knowledge base construction, including Thai text analysis and annotation process. To transform medical text data, mostly in Thai language, gathered from various sources available on the Internet into a structured knowledge base, a set of manual and semiautomatic methods are proposed. Extracted information in a structured knowledge base is kept in a form of ontology-based representation. The focused area of interest is general medical information. It involves disease characteristics, its cause and its treatment. The collected knowledge includes 1277 terms with 10530 information details. A system is proposed to link and search related information.

**Keywords:** Medical database, knowledge base system, Ontology, Natural Language Processing, Knowledge Discovery, Data mining

## 1 Introduction

Nowadays there is a large amount of data and information related to medical and health science distributed online in various web sites. However, it has not yet been organized in a systematic way. Most of data are gathered and categorized as links to medical data in various portal web sites. Even those data are related but there is no explicit link between them. Thus, typical users need to surf through those web portals by themselves or use search engines by inputting a set of keywords to get a list of related web pages. This process is inefficient and requires a lot of effort. To provide an efficient way to access required information, it is worth making a study on systematic construction of a knowledge base which is the result of combining medical and pharmaceutical information from several sources. In the past, several medical expert systems, such as MYCIN [1], doctors and knowledge engineers have contributed to clinical interpretations and diagnosis [2-3]. To develop this kind of knowledge-based systems, a process of knowledge acquisition is an important issue in building and expanding efficiently knowledge bases with high quality. Even now the automatic knowledge acquisition is still unsatisfactory due to over-complex algorithms and immature methodology [4-5]. Nowadays most medical knowledge bases still rely on manual knowledge acquisition to obtain knowledge but they are usually small and very specific to a relatively narrow medical area [6-7]. Compared to previous frameworks, the Internet provides a more effective method to create knowledge base for expert systems via its accessibility and popularity. Several attempts [8-9] have been done towards the transformation of information on the Internet to knowledge base with success to some extent, especially in English language.

This paper describes a framework of constructing a Thai medical knowledge base and its exploitation systems from diffuse and uncategorized medical data on the Internet.

The system is composed of three different processes; data acquisition, analysis and visualization. In the data acquisition, the medical text data on the Internet are collected by passing some keywords to search engine to get their reference addresses. Then more sophisticated information is extracted from the data by information extraction techniques. In the data analysis, the gathered data are analyzed to obtain some relation in the form of ontology. A set of links are extracted and represented in the form hypertext.

In the rest, Section 2 describes the framework to construct medical knowledge base in Thai language. The structure analysis of Thai running texts is given in Section 3. In Section 4, our implementation of Thai knowledge base system is shown in detail. Finally, conclusion and future work are given in Section 5.

## 2 The Framework to Construct a Thai Medical Knowledge Base

To construct a knowledge base, Thai medical texts are collected from the Internet and processed. Figure 1 shows the framework of construction the knowledge base.

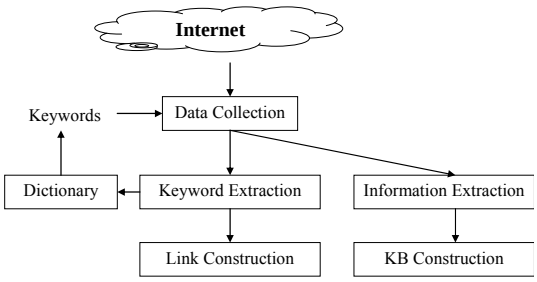


**Figure 1**: The framework for constructing a Thai medical knowledge base.

In the first place, a set of seed keywords are provided in the dictionary. In the data collection process, each keyword is selected to be an input for a search engine to gather information from the Internet. This step aimed to collect medical data (web pages) that diffuse on the Internet. In the keyword extraction, important keywords are detected and kept in the dictionary for further data collection. The extracted keywords are used for construction of hyperlinks among text data that we collected from the Internet. In parallel, the collected web pages are analyzed with information extraction techniques. The output is used to construct knowledge base. For Thai language which has no word boundary, it is necessary to consider some additional process, compared to English language processing, for both keyword and information extraction. In Thai language, there are still very few works [10] on basic language processing, especially word segmentation and part-of-speech tagging. The next section will focus on these issues. Moreover, to build knowledge base, a preliminary version of the knowledge base is designed. Table 1 illustrates the current scheme used for constructing the knowledge base in this work. It can be extended later to include more related information in medical area.

**Table 1**: A sample scheme for medical knowledge base

| Disease | Drug | Organ |
|---|---|---|
| - Appearance<br>- Cause<br>- Germ<br>- Contact<br>- Symptom<br>- Diagnosis<br>- Incurrent disease<br>- Treatment<br>- Drug<br>- Suggestion<br>- Prevention | - Explanation<br>- Characteristic<br>- Commercial name<br>- usage instruction<br>- Side effect<br>- Caution<br>- Exclude | - Appearance<br>- Function<br>- Disease<br>- Symptom |

The collected web pages or documents are manually analyzed by experts and/or automatically interpreted by an AI agent. The result will be stored in the above scheme.

## 3 Structure Analysis

In the Thai writing system, words are consecutively written without delimiters. Even there are spaces in Thai; they are occasionally inserted between phrases or

sentences or some components within a sentence. There are no standard rules of how to use spaces in Thai language. This chaotic phenomenon triggers difficulty in extracting keyword from Thai texts. In most cases, there are multiple possibilities to segment meaningful candidate words from a running text. More difficulty occurs when the running text includes words that do not exist in dictionary (called unknown). It is almost impossible to segment that unknown words from the text. This problem is more serious when we deal with Thai medical texts since they include plentiful terms that are not available in the form that we can access electronically. To extract keywords and their relationships, we need this morphological analysis and name entity extraction. In our work, first each lengthy sentence is manually split based on basic delimiting words, e.g. spaces and conjunctions, and verbal tokens. In addition to conjunctions, verbs in medical texts are not varied, comparing to nouns. By considering this heuristic, we can analyze the sentence by breaking each text chunk into the basic form of [subject + verb + object] where a verb indicates a relationship between the subject and the object of the sentence, as shown in Figure 2. Based on this syntactic structure, further relation extraction can be performed. Moreover, the result of the process can be used as statistics and training examples for learning rules used for later morphological analysis and name entity extraction.

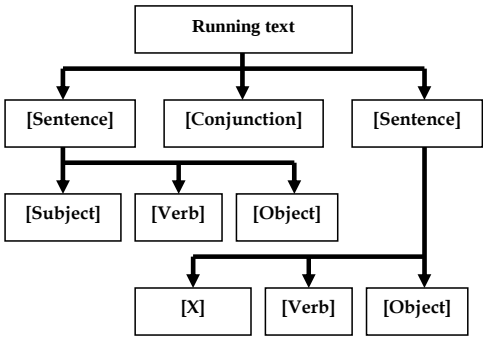


**Figure 2**: An example of Structure Analysis

Several techniques were proposed for segmenting Thai words. They can be separated into rule-based and statistical-based techniques. Those techniques are not appropriate for analyzing medical texts since the medical texts frequently contain various unknown words due to technical aspects. Naturally, the output from the naïve word segmentation techniques is likely to be short words that are substrings of the actual word and they are not meaningful in relation extraction.

In this work, we propose a new technique to segment Thai words by combining statistical and unsupervised machine learning approach. Firstly a Thai running text is split into Thai Character Cluster (TCC) [10] based on Thai character combination rules. Each TCC denotes a group of inseparable Thai characters. Then, a word is constructed from connecting neighboring TCCs based on their usage in sentences. TCCs are connected with each other by analyzing homogeneousness of tokens before and after the connected TCCs. Then, a group is identified as a word with the idea that a word is generally used with various words, thus it should come with inhomogeneous TCCs before and after. Neighboring TCCs tend to form a word if they are often co-occurring in many contexts. We apply statistical data and unsupervised machine learning technique to analyze and classify words. With the proposed approach, we can segment words based on their usage in medical texts and unknown words can be extracted. Moreover, name entities can also be identified by using training examples prepared by manually splitting sentences, and analyzing their spread throughout medical documents. With the above process, a medical running text will be split into small components, but some components, such as adverb or adjective, are still unnecessary. These words can be omitted without the loss of meaningful semantics. Figure 3 shows an example of Thai running text with the structure analysis.

**Original text**

สารโฟเลต หรือ กรดโฟลิกหรือโฟลาซิน เป็นสารอาหารจำพวกวิตามินบี ร่างกายต้องการสารโฟเลตวันละ 400 ไมโครกรัม ซึ่งสามารถรับประทานได้เพียงพอ ในอาหารประจำวัน

| **Line inserted text (Thai)** | **Line inserted text (English translation)** |
|---|---|
| สารโฟเลต หรือ กรดโฟลิกหรือโฟลาซิน เป็นสารอาหารจำพวกวิตามินบี | (Folate or Folic acid or Folacin is a vitamin B-typed nutrient.) |
| ร่างกายต้องการสารโฟเลตวันละ 400 ไมโครกรัม | (Normally our body needs Folate at least 400 microgram a day.) |
| ซึ่ง | (Which) |
| สามารถรับประทานได้เพียงพอ ในอาหารประจำวัน | ([we] can take sufficiently from our daily food.) |

**Structure-annotated text**

[S สารโฟเลต หรือ กรดโฟลิกหรือโฟลาซิน] [V เป็น] [O สารอาหารจำพวกวิตามินบี]

[S ร่างกาย] [V ต้องการ] [O สารโฟเลต] [X วันละ 400 ไมโครกรัม]

[C ซึ่ง]

[S ?] [V สามารถรับประทานได้เพียงพอ] [X ในอาหารประจำวัน] // S = Folate but it is omitted

**Figure 3**: An example of a Thai running text with the structure analysis
(C: Conjunction, S: Subject, O: Object, X: Other constituents, [..] and ? =ellipsis)

In the figure, the running text is analyzed into a shorter chunk by considering spaces and conjunctions, and verbal tokens. Next each component is annotated manually and/or automatically with S (Subject) or V (Verb) or O (Object) or C (Conjunction) or X (Others). Focusing on verbs, higher analysis techniques are needed to analyze these syntactic structures in order to obtain semantic structures, such as agent-theme relations, referential relations, anaphora relations and ellipsis relations. The output will be in the form of ontology or semantic web. This point is left as our future work.

## 4 Implementation

The objective of this work is to construct a large knowledge base from resources on the Internet. Still in the preliminary stage, we start collecting medical texts and implement a set of supporting tools for this purpose. In data collection, a number of medicinal and pharmaceutical web sites (2759 URLs) are selected as seeds to obtain medical text data.

**Table 2**: Collected medical text database
(numbers in brackets are the number of data)

| Type | Detailed Information | |
|---|---|---|
| Disease (474) | Appearance (1142) | Treatment (1036) |
| | Contract (162) | Prevention (466) |
| | Incurrent disease (436) | Germ (59) |
| | Suggestion (570) | Diagnosis (412) |
| | Cause (884) | Drug (79) |
| | Symptom (1052) | Others (303) |
| Organ (33) | Appearance (108) | Function (68) |
| | Symptom (37) | Others (46) |
| | Related Disease (137) | |
| Drug/ Chemical (770) | Description (1210) | Characteristic (382) |
| | Dose/usage (741) | Side effect (215) |
| | Exclusion (102) | Caution (550) |
| | Commercial name (180) | Others (153) |

From the seed web pages, we collected the data manually and analyzed them to store under the categories listed in Table 2. In this process, one web page may include more than one topic. Therefore, each part of the web page can be stored under different categories. Moreover, a web-based engine is developed to support navigation of related information via keyword linking.

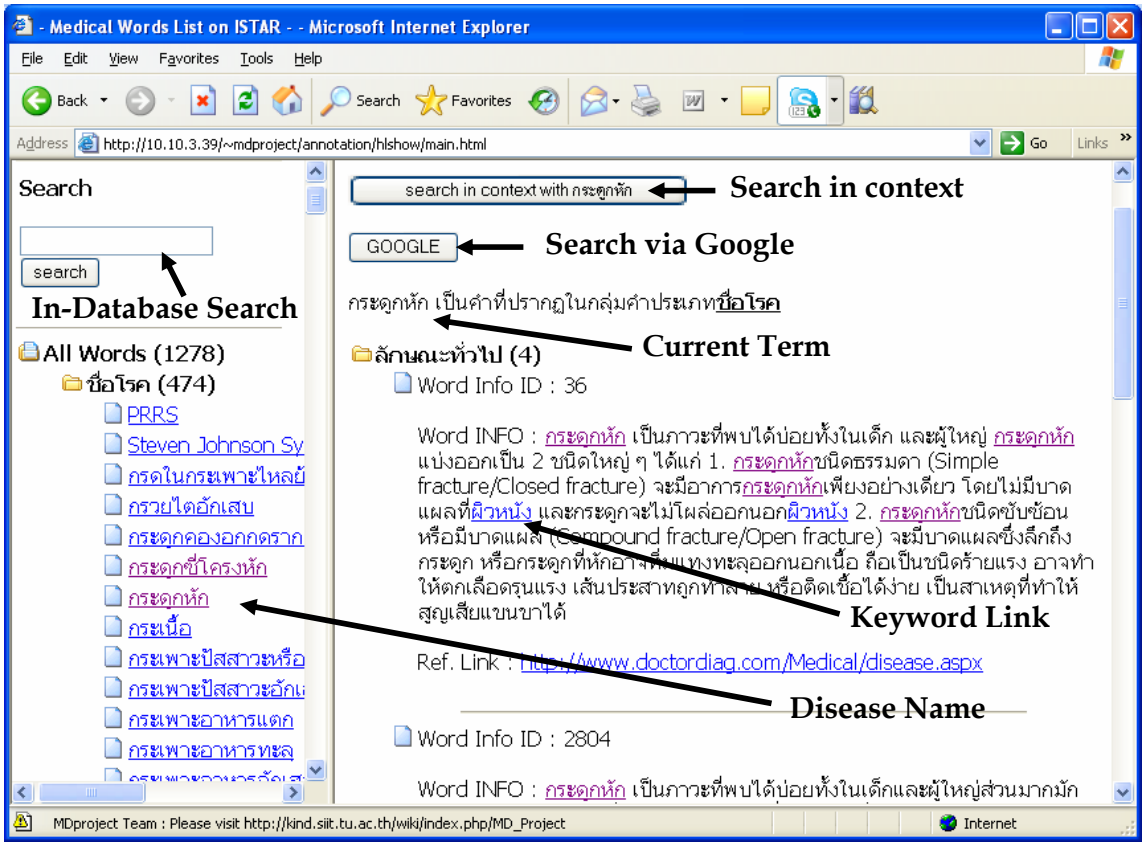


**Figure 4**: The web-based engine for providing medical information
(Keyword Linking, In-Database Search, Google Search, In-Context Search)

Figure 4 shows our web-based engine for providing medical-related information. The basic information in this system comes from manual collection. It also supports in-database searching, in-context search and Google search. On the left part of the screen are search box and the list of medical terms while on the right part are in-context search button, Google search button, the key of the current term, and detailed information about the current term. In the detailed information, there are some links to other related terms. Figure 5 illustrates the result of searching in context with the same term, i.e. after clicking 'search in context' in Figure 4. By this function, we can find terms that are related to the current term, including their types and frequencies. Each related terms possess links to their definitions. This Wiki-like function is helpful for searching relevance information.

## 5 Conclusion and Future Work

In this paper, we introduced a framework for constructing a medical knowledge base in Thailand, consisting of data acquisition, keyword extraction, link construction, and knowledge base construction. A set of manual and semiautomatic methods, such as Thai text analysis and annotation process, were proposed to transform Thai medical texts into structured data. They are disease characteristics, its cause and its treatment. Composed of 1277 terms with 10530 information details, the system supports linking and searching related information. In the future, automatic term extraction and information harvest from the Internet should be considered to automatically build Thai medical knowledge-base.

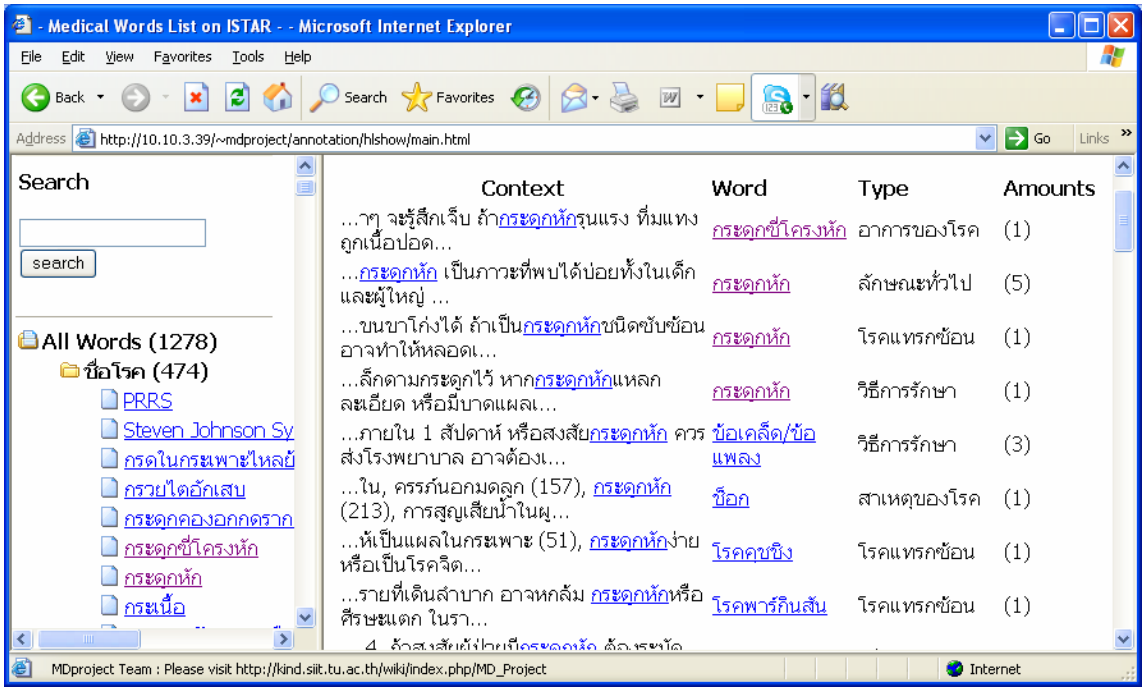


**Figure 5**: The result of in-context search


## Acknowledgement

This work has been supported by NECTEC under Project No. NT-B-22-I4-38-49-05.


## References


[1] E.H. Shortliffe, S.G. Axline, B.G. Buchanan, T.C. Meridan, S.N. Cohen, An artificial intelligence program to advise of physicians regarding antimicrobial therapy, Computers and Biomedical Research, 6 (1973) 544-560.

[2] L.M. Brasil, F.M. Azevedo, J.M. Barreto, Hybrid expert system for decision supporting in the medical area: complexity and cognitive computing, International journal of medical informatics, 63 (2001) 19-30.

[3] Kemal Polat and Salih Güneş, Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection, Expert Systems with Applications, Vol. 33, No. 2 (2007) 484-490.

[4] S. Tsumoto, Automated knowledge acquisition from clinical databases based on rough sets and attribute-oriented generalization, Proceedings of the AMIA Annual Symposium, (1998) 548-552.

[5] M.L. Wong, W. Lam, K.S. Leung, P.S. Ngan, C.Y. Cheng, Discovering knowledge from medical databases using evolutionary algorithms, IEEE Engineering in Medicine and Biology Magazine, 19 (2000) 45-52.

[6] S.L. Achour, M. Dojat, C. Rieux, P. Bierling, E. Lepage, A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development, Journal of the American Medical Informatics Association 8 (4) (2001) 351-360.

[7] Pietro Torasso, Knowledge based expert systems for medical diagnosis, Statistics in Medicine, Vol. 4, No. 3 (2006) 317-325.

[8] R. Grove, Internet-based expert systems, Expert System, 17 (3) (2000) 129-135.

[9] Hongmei Yan, Yingtao Jiang, Jun Zheng, Bingmei Fu, Shouzhong Xiao and Chenglin Peng, The internet-based knowledge acquisition management method to construct large-distributed medical expert systems, Computer Methods and Programs in Biomedicine, Vol. 74, No. 1 (2004) 1-10.

[10] Thanaruk Theeramunkong and Thanasan Tanhermhong, Pattern-based features vs. statistical-based features in decision trees for word segmentation, IEICE Transaction on Information and Systems, Vol. E87-D (5) (2004) 1254-1260.