Engineering Requirements with Communication-Based Vector Space Representation

P. Lertthasanawong[1], T. Chandarasupsang[1], N. Chakpitak[1], P.P. Yupain[2]

[1]College of Arts Media and Technology, Faculty of Knowledge Management, Chiang Mai University, Chiang Mai 50200, Thailand,
[2]Quantum Life Institute, Sainoi, Nonthaburi 11150, Thailand, E-mail: <pornpen@neo-netsoft.com>, <kypreech@kmitl.ac.th>

Abstract: Engineering Requirements are a complex process, which oftenly works with the domain expert to develop the planned system. The processes involve a large amount of Natural Language (NL) communication. It is always lack of clarity, unambiguous and misconceptions especially Thai language. This is because they do not have any explicit word boundary. The research proposes the framework of analyzing natural language requirement in term of linguistic and mathematic approach. The methodology is contributed for Very small Enterprises (VSEs) who have little experiences and resources. The results have shown that the Longest matching gives higher accuracy for word segmentation. The domain dictionary which is developed from domain expert provides the precision for requirement understanding. The vector model represents the usage in both Thai and English language.

[Lertthasanawong R, Chandarasupsang T, Chakpitak, Yupapin PP. **Engineering Requirements with Communication-Based Vector Space Representation.** *Life Sci J* 2014;11(8):941-949]. (ISSN:1097-8135). http://www.lifesciencesite.com. 138

Keywords: Communication Model; Thai Word Segmentation; Vector Space representation; Very Small Entities.

1. Introduction

It is a typical practice that software requirements are specified in natural languages (NL). Software requirements specifications is a common known that 71.80% are captured in NL. (Luisa, 2004) More over different data come from multiple resources (social networks, sensors, web mining, logs, etc.). The main obstacle is the ambiguity of Natural Language which is understandable by human and not machine. Even English is also the problems in machine processing of English specification of the software requirements. It is more difficult in Asean languages because of the special characters with word boundaries. (Berry, 2004)

Requirements elicitation is described as the first stages of building an understanding of the problem that the software is required to solve. Requirements elicitations techniques may also be classified into traditional, group, formal, semiformal, and natural language. In traditional ways, requirements elicitation process is performed face to face such as through interviews, whether individually or in a group among customer or manager. There have been several difficulties conducting interview session such as it is time consuming and higher cost. There may exists conflict between key ,users and sometimes software house with regards of perception, assumption, problem defined, and even objective of a system. There are the different personalities and behavior, as well as background and terminology which used during communication between both parties.

The paper presents the alternative way of communication for requirement elicitation. The requirements are represented in term of mathematic approach for more exactly understanding. The Vector Space representation is requirement for calculation what words are meaning. The main challenges of the paper are to provide frameworks and tools that ensure any data in the process without losing.

The remaining paper is structured into the following sections: Section 2 states the specific characters of Thailand software industry. Section 3 Research Methodology Section 4 Case Study Finally, the results and conclusion of the approach is presented in section 5.

2. Challenges in Small Software Company and Thai Words in Requirement Engineering

In Thailand, most of the software houses (with approximately 1,300 local and international software companies in 2012) are classified as very small entities (VSEs). This makes their business styles different characteristics when compare to large companies. They have the registered capital of less than US$ 2 million. (Software Park Thailand Newsletter, 2011) VSEs have common problems related to the management of risk and quality of software projects. This generates cost overruns, time delay and cancelled projects. (Richardson, 2007) From the studies (Habra et al., 2008), it is clear that the majority of ISO/IEC standards do not address the needs of VSEs. Conformance with these standards is difficult. Subsequently VSEs have no and very limited ways to be recognized as entities that produce quality software in their domain. Some model like the CMMI is not affordable for the small organizations. It is found that VSEs find it difficult to relate ISO/IEC standards to their business needs and to justify the application of the standards to their business practices.

Moreover Thai words are implicitly recognized and in many cases, they depend on the individual judgments. They are more complex because the language does not have any explicit word boundary such as a space to separate between each word. The software requirement writing in Thai is continuously without the use of word delimiters. For example the writing is in an English sentence "You ate an apple" as"Youateanapple". This is even more difficult in Thai language because it contains 44 characters, 4 levels of structure, 21 consonant sounds, 18 single vowels and 6 diphthongs. (Nitisaroj, 2010) These may cause the multiple customers interpretation of the requirements in different ways. Each customer concludes that his or her interpretation is correct but it is conflicted. The ambiguity remains undetected until later in software development process. It is more expensive to resolve. The communication-based vector space representation (CVSR) is the name of proposed model.

The CommonKADS methodology is selected because it provides sufficient tools such as a model suite and templates for different knowledge intensive tasks. However, creating and representing knowledge model create difficulties to knowledge engineer caused the ambiguity and unstructured of the source of knowledge. This enables to spot the opportunities and bottlenecks in how organizations develop, distribute and apply their knowledge resources, and so gives tools for corporate knowledge management.

3. Communication Vector Space Representation

Software development is distinct from other types of engineering because the product is intangible, progress is not explicit. Moreover the team members both customer and software house rely on the documentation of others to review progress. The typical activities of the requirements elicitation process can be divided into a set of fundamental activities as described later. The precision of requirement depends on the identification and definition of keywords that exactly reflect the user's needs and wants. The requirement elicitation is the knowledge intensive information processing task. It provides a specification of the data and knowledge structures needed for the application. The conceptual frame work of the research is designed to use both linguistic perspective for better understanding, reducing communication gap and mathematic perspective for more accuracy and quality of software requirements. It is also provides the methods to perform a detailed analysis of knowledge-intensive tasks and processes. Finally, this will supports the development of knowledge systems that support selected parts of the business process.

The key communication-model component describing such communicative acts is called a transaction. A transaction tells what information objects are exchanged between what agents and what tasks. Transactions are the building blocks for the full dialogue between two agents, which is described in the communication plan. Transactions themselves may consist of several messages, which are then detailed in the information exchange specification. This specification is based on predefined communication types and patterns, which make it easy to build up message protocols in a structured way. The research is set up and focuses on the dialogue between agents especially for software requirement elicitation (Knowledge Acquisition). There are three steps for setting which are (Schreiber , 1994)

1. The overall communication plan, which governs the full dialogue between the agents;
2. The individual transactions that link two (leaf) tasks carried out by two different agents;
3. The information exchange specification that details the internal message structure of a transaction.

To become effective, produced knowledge has to be transferred to the various parties that use it to perform their own tasks. Accordingly, the process of constructing the CommonKADS communication model goes in terms of three subsequent layers, from global to detailed specifications. This does not claim to be a full knowledge-management methodology. It is in practice used successfully as a powerful tool to support knowledge management. It gives a clear roadmap of how knowledge analysis and knowledge-system development can be used as techniques within an overall knowledge management approach as figure 1.

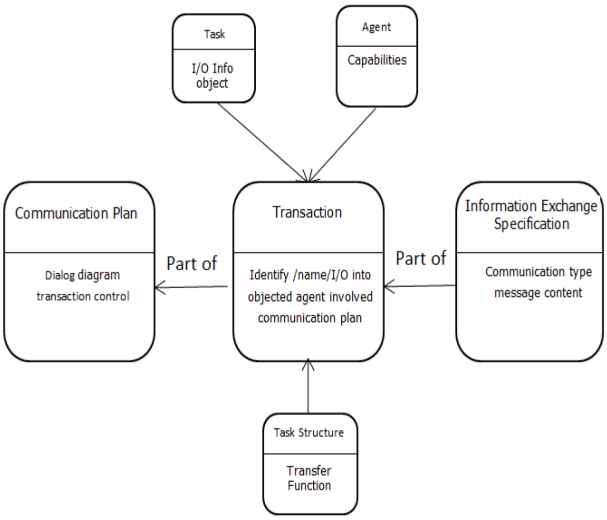

Task
I/O Info object
Agent
Capabilities
Communication Plan
Dialog diagram transaction control
Part of
Transaction
Identify /name/I/O into objected agent involved communication plan
Part of
Information Exchange Specification
Communication type message content
Task Structure
Transfer Function

Figure 1.The communication model. (Schreiber, 2000)

The above diagram represents the information exchange of customers. The data need to transform into information, to knowledge, and finally to the understanding that support the transition from each stage to the next. At the end of communication model, the better consistent, necessary and more complete of user requirement will be met.

This brings the concept of the communication-based with structure form of natural language. The vector space is used for requirement representation. The Communication Vector Space Representation (CVSR) is a new method that contributes designing for requirement transformation. It starts with constructing the dialogue diagram for software requirement by listing all tasks which carried out by considered agents with their input/output information objects. The dialogue diagram presents the complete information flow for requirement engineering. Each requirement is gathered in web based communication with giving identification for traceability and transformation. There are three stages for Communication Vector Space Transformer which a listed bellowed.

Stage1: The communication plan defines the communication protocol used in the project. Which persons communicate which issues to whom? The communication plan should explicitly define the team and member responsibilities. Transaction Dialog construction is developed. The dialogue diagram presents all information flow part of communication plan between the customers and software house.

Stage2: Requirement Segmentation. Each software requirement is tokenized into a series of terms before it can be further analyzed and translated. For information retrieval, dictionary based is used for comparison of suitable segmentation approach. The unknown words from software requirements are trained for the new vocabularies and collected for specific business dictionary meaning which is called Domain Dictionary (DD).

Stage 3: Vector Space representation is based on the previous stage, computers can understand very little of the meaning of human language. This hints that the transformation of each requirement is needed. Vector space model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers for nonambiguity.

At the end of model, every single requirement is represent in form of vector which is called Basic requirement. The approach decreases the ambiguity. The precision of requirements when using DD gives higher value. The conceptual framework of Requirement Engineering with Communication-Based Vector Space Representation in the alternative approach for requirement elicitation is shown in figure 2 bellowed.

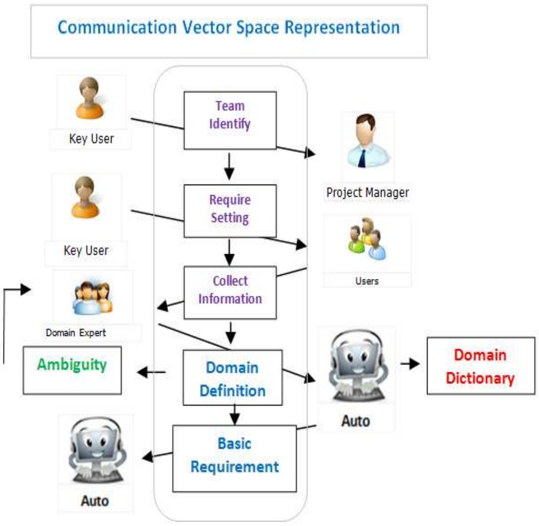


Figure 2. The Communication Vector Space Representation

Moreover, Knowledge Engineering describes the transaction by using the transaction description (worksheet CM-1), shown in the table below, specifying the transaction name, objective, agent involved, communication plan name and the constraints of each transaction (McMorran, 2007; Schwarz, 2004; Batarseh, 2009; Monica, 2005). Meanwhile, each transaction description also uses information exchange specification (worksheet CM-2). In this worksheet is seen the transaction name and agent involved, identifying the sender and receiver of this transaction. Moreover, it also describes the information item that classifies the layer of each part of the information, separating core and support information, and the message specification, which describes the communication message type that makes up the transaction of each individual message. In order to analyses the responsibility area and the information possess by one agent, the worksheet CM-1 and CM-2 have been used to identify the information description and agent involve of each task. The result for applying the CM-1 and CM-2 is show in the case study in part IV.

4. Case Study

The research is applied into three cases. The case studies of this research are divided into two main groups which are Thai language and English. The first case is tested with English requirements for robustness (Project A). The second has two sub groups which are Project B and C. They are tested with Thai language but different characters.

The project A is developed for accounting requirements in English. This testing is for quality assurance methodology and framework. The project A is tested for the expectation outcome and performance. Most of them is package software and is usually used for enterprise application.

The two data sets are from two Domain experts of (1) Natalin Group co, ltd (Project B) and (2) Neo SME Project (Project C). The former is represented the medium project with 361 requirements. It is customized software which is more complicated than package software. The latter is represented the small project with 96 requirements. All requirements are in Thai language. There is no large project because it is the limitation of VSEs.

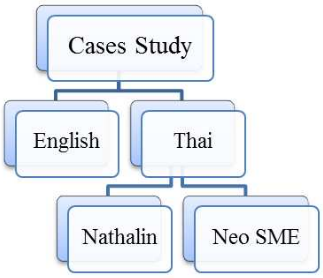


Figure 3. Research Cases Studies

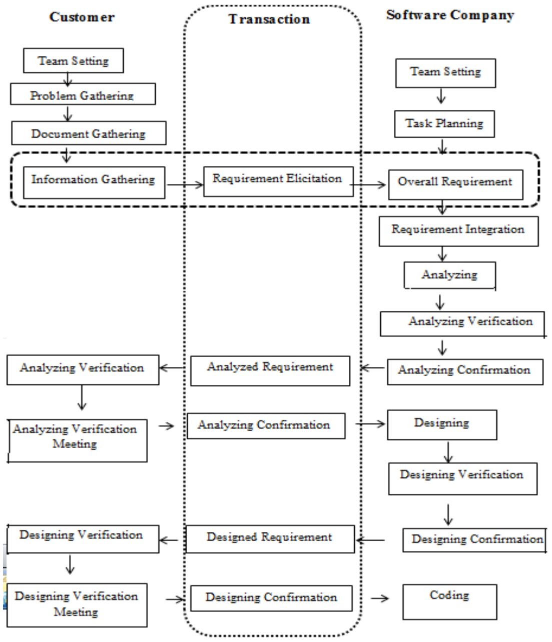


Figure 4. The proposed lay-out of a dialogue diagram for VSEs Communication Model

4.1 Starting with Transaction Dialog constructions.

The research starts with constructing the dialogue diagram for software requirement by listing all tasks which carried out by considered agents (the involved customers in the organization) with their input/output information objects. The plan is focused on the sequence of information). The dialogue diagram presents the complete information flow part of communication model for specifies knowledge/information transfer from customers and software house as see in figure 4.

There are two templates of communication model for gathering of transaction. The first is transaction description worksheet (CM1) and the second is information exchange specification (CM2). The individual transactions in CM1 are specified of the transaction description. The information exchanged specification in CM2 is separated for details.

4.2.Business Requirement Segmentation

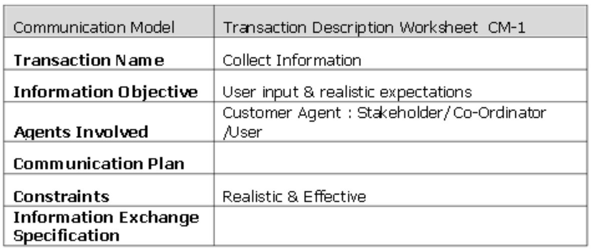
One important problem of Natural Language Processing is figuring out what a word means when it is used in a particular context. The different meaning of a word is listed as its various senses in a dictionary. Recently, the Human Language Technology Laboratory (HLT) under the National Electronics and Computer Technology Center (NECTEC) has designed and released the Thai word segmentation corpus to the Thai NLP research community. There are two popular tools (Haruechaiyasak , 2008) as following.

1. Longest Matching is the method that scans an input sentence from left to right and select the longest match with a dictionary entry at each point. In case that the selected match cannot lead the algorithm to find the rest of the words in the sentence, the algorithm will backtrack to find the next longest one and continue finding the rest.

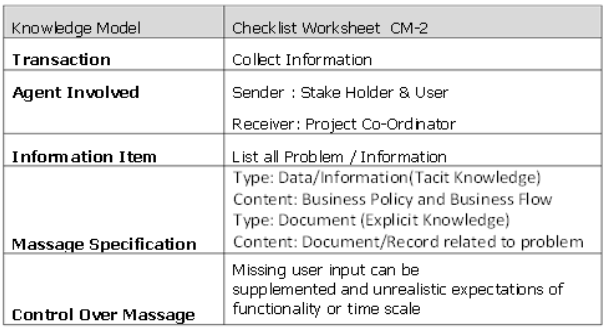
2. Maximum Matching is algorithm that first generates all possible segmentations for a sentence and then select the one that contain the fewest words, which can be done efficiently by using dynamic programming technique. Because the algorithm actually finds real maximum matching instead of using local greedy heuristics to guess, it always outperforms the longest matching method. Nevertheless, when the alternatives have the same number of words, the algorithm cannot determine the best candidate and some other heuristics have to be applied.

There are two part of this stage of the experiment. The first part is to find out which segmentation technique is better for Thai language. The second part is to comparison the precision of using general dictionary and specific dictionary. The experiment starts with the requirement of the Project C for comparative study of word segmentation techniques. The first fifty requirements are used with a pre-segmented corpus of NECTEC LexiTron Dictionary with the two segmentation techniques. The result in table 1 shows the significant to Thai language because the more of unknown word is increasing when use the longest matching. The reason of the results is from Thai language does not have any explicit word boundary delimiters, such as a space, to separate between each word when segment. When the system runs the algorithm actually finds real maximum matching instead of using local greedy heuristics to guess, this bring the misunderstanding and not accurate.

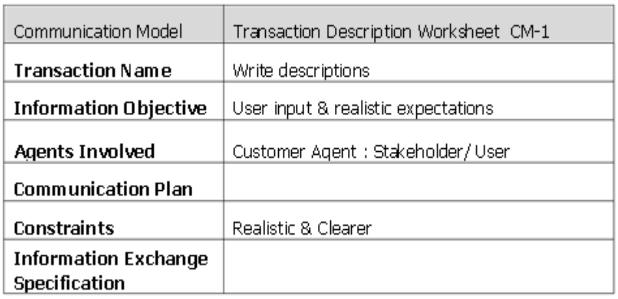
Table 1. The transaction description worksheets (CM1) and the information exchange specification worksheets (CM2) for requirement elicitation



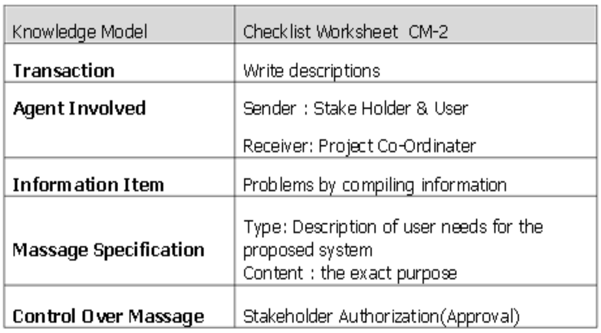
Communication Model	Transaction Description Worksheet CM-1
Transaction Name	Collect Information
Information Objective	User input & realistic expectations
Agents Involved	Customer Agent : Stakeholder/ Co-Ordinator /User
Communication Plan	
Constraints	Realistic & Effective
Information Exchange Specification	



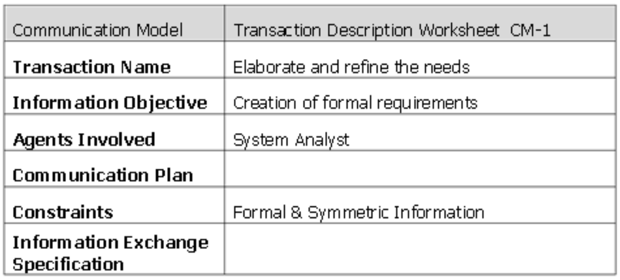
Knowledge Model	Checklist Worksheet CM-2
Transaction	Collect Information
Agent Involved	Sender : Stake Holder & User Receiver: Project Co-Ordinator
Information Item	List all Problem / Information
Massage Specification	Type: Data/Information(Tacit Knowledge) Content: Business Policy and Business Flow Type: Document (Explicit Knowledge) Content: Document/Record related to problem
Control Over Massage	Missing user input can be supplemented and unrealistic expectations of functionality or time scale



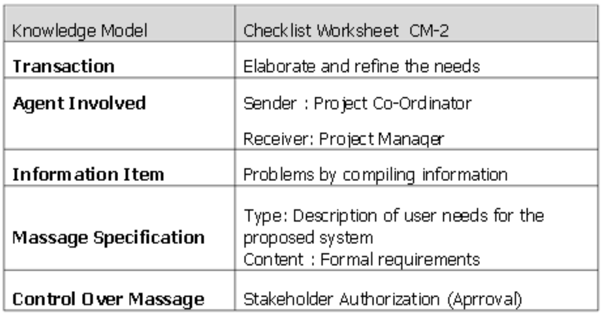
Communication Model	Transaction Description Worksheet CM-1
Transaction Name	Write descriptions
Information Objective	User input & realistic expectations
Agents Involved	Customer Agent : Stakeholder/ User
Communication Plan	
Constraints	Realistic & Clearer
Information Exchange Specification	



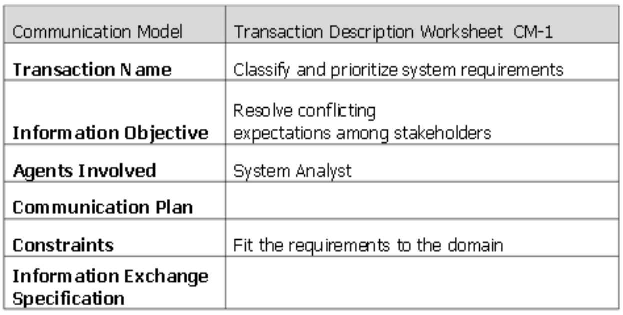
Knowledge Model	Checklist Worksheet CM-2
Transaction	Write descriptions
Agent Involved	Sender : Stake Holder & User Receiver: Project Co-Ordinater
Information Item	Problems by compiling information
Massage Specification	Type: Description of user needs for the proposed system Content : the exact purpose
Control Over Massage	Stakeholder Authorization(Approval)



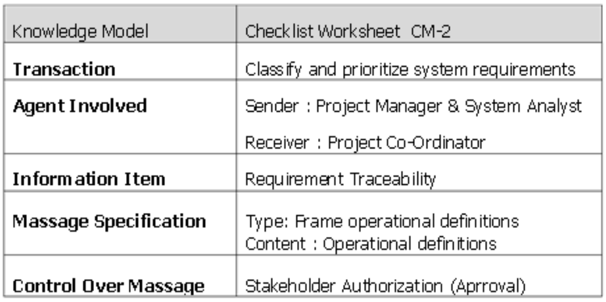
Communication Model	Transaction Description Worksheet CM-1
Transaction Name	Elaborate and refine the needs
Information Objective	Creation of formal requirements
Agents Involved	System Analyst
Communication Plan	
Constraints	Formal & Symmetric Information
Information Exchange Specification	



Knowledge Model	Checklist Worksheet CM-2
Transaction	Elaborate and refine the needs
Agent Involved	Sender : Project Co-Ordinator Receiver: Project Manager
Information Item	Problems by compiling information
Massage Specification	Type: Description of user needs for the proposed system Content : Formal requirements
Control Over Massage	Stakeholder Authorization (Aprroval)



Communication Model	Transaction Description Worksheet CM-1
Transaction Name	Classify and prioritize system requirements
Information Objective	Resolve conflicting expectations among stakeholders
Agents Involved	System Analyst
Communication Plan	
Constraints	Fit the requirements to the domain
Information Exchange Specification	



Knowledge Model	Checklist Worksheet CM-2
Transaction	Classify and prioritize system requirements
Agent Involved	Sender : Project Manager & System Analyst Receiver : Project Co-Ordinator
Information Item	Requirement Traceability
Massage Specification	Type: Frame operational definitions Content : Operational definitions
Control Over Massage	Stakeholder Authorization (Aprroval)

From table 2, the percentages of accuracy with different percentages of unknown words are explored. It is found out that in case of no unknown words, the accuracy is around 97% in both maximum matching and longest matching but the accuracy drops to 54% and 48% respectively, in case that 50% of words are unknown words. As the percentage of unknown words rises, the percentage of accuracy drops continuously.

The Longest Matching is used for the research to improve Rule-Based for Thai Word Segmentation. This is very useful for Thailand software industry to reduce project failure due to requirements factor.

Table 2. The accuracy of two dictionary-based systems vs. percentage of unknown words

Unknown Word (%)	Accuracy (%)	
	Longest Matching	Maximum Matching
10	93.12	92.23
20	86.21	82.60
30	68.07	64.52
40	61.53	57.21
50	54.01	48.67

The base case study of package software

Neo SME is represented the packaged project with less than 100 requirements. The case consists of the standard requirements. The two popular segmentation techniques which are Longest Matching and Maximum Matching are implemented to find the better accuracy for segmentation. Most natural language processing applications require input text to be tokenized into individual terms or words before being processed further.

The first fifty requirements are used with a pre-segmented corpus of NECTEC Lex-iTron Dictionary with the two segmentation techniques. Each segmentation technique gives the different output for separation. When compare each output with the standard, the results show as following example table.

Table 3 Word segmentation of two techniques using NECTEC LexiTron Dictionary

Requirements	Word Segmentation
1.1.1 ต้องการให้ระบบต้องยอมให้แก้ไขรายละเอียดเจ้าหนี้ได้	ต้องการ|ให้|ระบบ|ต้อง|ยอมให้|แก้ไข|รายละเอียด|เจ้าหนี้|ได้
1.1.1 Need: The system must allow creditor details to be edited	The|system|must|allow|creditor|details|to|be|edited
1.1.2 ต้องการให้บัญชีลูกหนี้ไม่สามารถเปลี่ยนมูลหนี้ได้	ต้องการ|ให้|บัญชี|ลูกหนี้|ไม่|สามารถ|เปลี่ยน|มูลหนี้|ได้
1.1.2 Need: The creditor account type cannot be changed from liability	The|creditor|account|type|cannot|be|changed|from|liability

For the experiment of comparison the precision of dictionary, the new dictionary is developed by the domain expert. This approach can improve in the accuracy of identifying the correct word sense which will result in better machine translation systems, information retrieval systems. Domain Dictionary is set up by key users for assign an appropriate sense at the first round of gathering requirement when changed or new needs are required, the requirements.

At the end of the first round, some domain vocabulary is approved by domain expert for domain dictionary (DD). The next round of word segmentation, the vocabularies from both LexiTron Dictionary and DD are used together. The next fifty requirements are segmented. The new vocabulary of DD will be found more. This make the words clear for better understanding.

The results of domain dictionary are presented in term of statistic outputs. They are measured in term of the precision (P), the recall (R) and (F) of the data. The details are bellowed. (Powers, 2011)

Precision (P) is the fraction of the documents retrieved that are relevant to the user's information need.

$$P = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall (R) is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$R = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

This is also known as the F measure (F), because recall and precision are evenly weighted.

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

The results of experiments using Domain Dictionary can be analyzed as follow:

Table 3. The Comparison of Precision, Recall and F Measure between LexiTron and Domain Dictionary of B and C Case

	Natalin Group co, ltd (Project B)					
	LexiTron Dict Controlled Variables			Domain Dict UncontrolledVariables		
	P	R	F	P	R	F
100	61.34	62.01	61.67	68.13	75.57	71.65
200	72.35	77.37	74.47	75.04	70.19	76.72
300	79.67	79.7	79.19	80.79	80.33	79.95

	NeoSME (Project C)

	LexiTron Dict Controlled Variables			Domain Dict UncontrolledVariables		
	P	**R**	**F**	**P**	**R**	**F**
96	62.54	73.78	67.20	68.49	79.38	73.14

The results show that to use the domain dictionary for word segmentation increases the requirement precision value in both medium and small projects (B and C). The F-measure can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

It is found that word segmentation in requirement engineering is total different from documentation as following reasons.

1. Software requirement clustering is at a very fine level of granularity whereas document clustering stems from the need to sort or filter large collections of texts which is usually the sole purpose, of information clustering, is to organize the documents into a limited number of categories to ease a few basic tasks(categories is typically small).

2. Each Software requirement domain has different assumptions and own dictionary about the membership of each datum but document clustering usually assumes each document comes from one of the fixed numbers of categories.

3. Software requirements can be multiple clusters whereas documents are specific cluster. The research segmentation is new and useful for Thailand software industry

Vector Space Representation

Vector space model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers for example, index terms. It is used in information filtering, information retrieval, indexing and ranking.

The research is designed for mathematical structure to form by a collection of elements called vectors which may be added together and multiplied (scaled) by numbers, called scalars in this context. From the requirement segmentation,

each term of word in each requirement is calculated for vector to represent the whole requirement. The coding of vector space transformer is set up as

The calculation begin with

```
    var sigMa = 0.0;
        foreach (var dw in dictWords)
        {
          int tf = _getDictTF(dw.tf, expect.ID);
          sigMa += (tf * tf) * ((double?)(dw.idf
* dw.idf) ?? 0.0);
        }
```

```
        expect.Vector =
findExpectVector(expect.ID, dictWords,
Math.Sqrt(sigMa).ToDecimal().MyRound(4));
        private string _findExpectVector(int
expectID, IEnumerable<req_DomainDict> dictWords,
decimal normalizeValue)
        {
          var result = "";
          if (normalizeValue != 0m)
          {
            //loop นี้กลับไปหาค่า w แต่ละตัว
            foreach (var dw in dictWords)
            {
              var w_ik = (_getDictTF(dw.tf,
expectID) * dw.idf) / normalizeValue;
              result +=
string.Format("{0}:{1}{2}", dw.TWord,
w_ik.ToDecimal().MyFloatToMoney(),
WordSegHelpers.SegSeparator);
            }
          }
          return result;
        }
```

The results of Vector Space Transformer are in the table 4. The results show that it is automatic process for requirement elicitation which is very helpful for VSEs to analyze for the next steps.

The results of Vector Space Transformer are in the table 4. The results show that it is automatic process for requirement elicitation which is very helpful for VSEs to analyze for the next steps.

The vector space representation can reduce misunderstandings and complexity of language. Because of the ambiguity of natural language is in the form of the mathematical structure. The analysis of mathematical structure is easier than in form of language. Using computer is very fast to query the vector for every word in all requirements.

The exact value is of each requirement will be

grouped into requirement features for requirement document for the next steps.

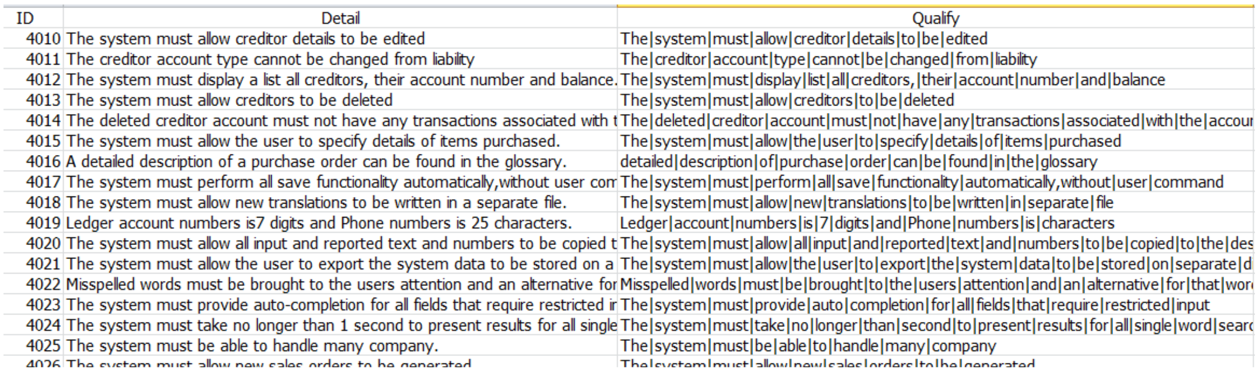
Robustness Test of Vector Space Representation

For making more benefit, the approach is set up for testing in the other language. English is used for the test run. The result of English has shown that the longest matching is used to segment very well. It is very easy and no ambiguity because the punctuation. in the test data. They are period (.) and comma (,). Both of them are very useful as in figure 6.

In English, the use of punctuation, particularly the full stop character is a reasonable approximation. However in English, The problem is non-trivial, because while some written languages have explicit word boundary markers, such as the word spaces and full stop. When processing plain text, tables of abbreviations that contain periods can help prevent incorrect assignment of sentence boundaries.

Table 4. Vector of Requirement Representation of Thai language

Detail	Qualify	Vector
1.1.1 ต้องการ เพิ่ม แก้ไข หรือ ยกเลิก ข้อมูลสินค้า เพื่อนำข้อมูลสินค้าไปใช้อ้างอิงในหน้าใบบันทึกข้อมูลใบสั่งซื้อ ฯลฯ (Need to add edit or delete goods for purchasing order.)	ต้องการ|เพิ่ม|แก้ไข|ยกเลิก|ข้อมูล|สินค้า|นำ|ข้อมูล|สินค้า|ไปใช้|อ้างอิง|ใบ|บันทึก|ข้อมูล|ใบสั่งซื้อ|ฯลฯ	ต้องการ:0.01|ข้อมูล:0.19|สินค้า:0.22|แก้ไข:0.18|ยกเลิก:0.20|เพิ่ม:0.21|บันทึก:0.29|ใบสั่งซื้อ:0.30|นำ:0.37|อ้างอิง:0.37|ฯลฯ:0.40|ไปใช้:0.45|
1.1.2 ต้องการระบุได้ว่าสินค้านี้มีผู้ขายใดที่ขายสินค้า และต้องการบันทึกราคาสินค้าแตกต่างกันของผู้ขายแต่ละรายได้ (Need to identify which material, who sale and to record the different price of each buyer.)	ต้องการ|ระบุ|ได้ว่า|สินค้า|ผู้ขาย|ขาย|สินค้า|ต้องการ|บันทึก|ราคา|สินค้า|แตกต่าง|ผู้ขาย	ต้องการ:0.02|สินค้า:0.30|บันทึก:0.26|ผู้ขาย:0.56|ขาย:0.28|ระบุ:0.30|ได้ว่า:0.34|ราคา:0.34|แตกต่าง:0.37|
1.1.3 ต้องการระบุได้ว่าสินค้านี้มีลูกค้าใดที่ซื้อสินค้า และต้องการกำหนดราคาสินค้าแตกต่างกันให้กับลูกค้าแต่ละรายได้ (Need to identify which products, who buy and to set the different price of each customer.)	ต้องการ|ระบุ|ได้ว่า|สินค้า|ลูกค้า|ซื้อ|สินค้า|ต้องการ|กำหนด|ราคา|สินค้า|แตกต่าง|ให้|ลูกค้า	ต้องการ:0.02|สินค้า:0.30|ลูกค้า:0.46|ซื้อ:0.24|ให้:0.25|ระบุ:0.30|ได้ว่า:0.34|กำหนด:0.34|ราคา:0.34|แตกต่าง:0.37|



ID	Detail	Qualify
4010	The system must allow creditor details to be edited	The|system|must|allow|creditor|details|to|be|edited
4011	The creditor account type cannot be changed from liability	The|creditor|account|type|cannot|be|changed|from|liability
4012	The system must display a list all creditors, their account number and balance.	The|system|must|display|list|all|creditors,|their|account|number|and|balance
4013	The system must allow creditors to be deleted	The|system|must|allow|creditors|to|be|deleted
4014	The deleted creditor account must not have any transactions associated with t	The|deleted|creditor|account|must|not|have|any|transactions|associated|with|the|accour
4015	The system must allow the user to specify details of items purchased.	The|system|must|allow|the|user|to|specify|details|of|items|purchased
4016	A detailed description of a purchase order can be found in the glossary.	detailed|description|of|purchase|order|can|be|found|in|the|glossary
4017	The system must perform all save functionality automatically,without user com	The|system|must|perform|all|save|functionality|automatically,without|user|command
4018	The system must allow new translations to be written in a separate file.	The|system|must|allow|new|translations|to|be|written|in|separate|file
4019	Ledger account numbers is7 digits and Phone numbers is 25 characters.	Ledger|account|numbers|is|7|digits|and|Phone|numbers|is|characters
4020	The system must allow all input and reported text and numbers to be copied t	The|system|must|allow|all|input|and|reported|text|and|numbers|to|be|copied|to|the|des
4021	The system must allow the user to export the system data to be stored on a	The|system|must|allow|the|user|to|export|the|system|data|to|be|stored|on|separate|d
4022	Misspelled words must be brought to the users attention and an alternative for	Misspelled|words|must|be|brought|to|the|users|attention|and|an|alternative|for|that|wor
4023	The system must provide auto-completion for all fields that require restricted ir	The|system|must|provide|auto|completion|for|all|fields|that|require|restricted|input
4024	The system must take no longer than 1 second to present results for all single	The|system|must|take|no|longer|than|second|to|present|results|for|all|single|word|searc
4025	The system must be able to handle many company.	The|system|must|be|able|to|handle|many|company
4026	The system must allow new sales orders to be generated	The|system|must|allow|new|sales|orders|to|be|generated

Figure 6. Requirement Segmentation of English languages

The words of standard dictionary and the words of domain dictionary are the same when they are segmented. This gives the result of calculation of the precision recall and F measure equally. The equation of vector is calculated easily. The certain aspects of punctuation are stylistic. The customers use them for make the guide of stop word or separate the software requirements.

From the results of English language, it shown that the communication-based vector space representation model can use well in English. The results give the emphasized evidence of business requirements representation. The model runs well and faster than Thai language. The calculation of both standard and domain dictionary of English give the same results. This means that the complication of English is not high when compared to Thai language.

Table 5. Vector of Requirement Representation of English language

Requirements	Word Segmentation	Vector Representations
The system must allow creditor details to be edited	The|system|must|allow|creditor|details|to|be|edited	The:0.11|must:0.07|be:0.15|system:0.15|to:0.20|allow:0.29|creditor:0.58|details:0.69|

<table>
<tr><td>The creditor account type cannot be changed from liability</td><td>The|creditor|account|type|cannot|be|changed|from|liability</td><td>The:0.15|be:0.20|account:0.58|creditor:0.78|</td></tr>
<tr><td>If an existing customer has previous quotes, these might be made available to allow the sales order to be automatically generated.</td><td>If|an|existing|customer|has|previous|quotes,|these|might|be|made|available|to|allow|the|sales|order|to|be|automatically|generated</td><td>be:0.20|to:0.27|the:0.17|allow:0.19|generated:0.33|order:0.38|sales:0.38|an:0.46|has:0.46|</td></tr>
</table>

5. Conclusion

The research is designed to communicate automatically in requirement elicitation processes. This paper presents that every requirement is concern as rather than at just a group of related requirements which the traditional requirement elicitation has been done.

The new vocabularies are collected for better understanding in domain meaning. Most of them do not appear in the normal dictionary. It is called Domain Dictionary. This approach supports in the accuracy of word segmentation in the real business environment. The longest matching is used before transform the software requirement into vector number.

The method is suitable for many languages. The precision of the right segmentation is very crucial for requirement calculation. Each requirement is represented by vector. This method is also helpful for VSEs. The approach reduces time and resource in requirement analyzing. The vectors can be precisely queried for further grouping in software design process. In the future, it is planned to group the vector into software feature automatically. Most of them are unstructured. The new challenge in requirement engineering is needed to be studied.

References

1. Luisa M, Mariangela F, Pierluigi I. Market research for requirements analysis using linguistic tools, Requirements Engineering, v.9 n.1, February 2004:40-56.
2. Berry DM, Kamsties E. Ambiguity in requirements specification, Perspectives on software requirements. Springer US, 2004: 7-44.
3. Software Park Thailand Newsletter. Readiness of Thai Software Industry for AEC 2015. Software Park Newsletter Vol. 2/2011:1-12.
4. Richardson I. Why are Small Software Organizations Different? IEEE Volume 24(1), 2007:18-22.
5. Habra N, Alexander S, Desharnais JM., Laporte, CY, Renault A. Initiating software process improvement in very small enterprises. Experience with a light assessment tool. Journal Information & Software Technology, Vol. 50, 2008:7-8.
6. Nitisaroj R. Thai, the Tiger of Text Analysis: An Introduction to Thai Text Processing Issues". Government Users Conference. Chantilly, June 8-9, 2010. Virginia: 5-8.
7. Schreiber G, Welinga B, Robert de H, Akkermans H, Van de Velde W. CommonKADS: A Comprehensive Methodology for KBS development, IEEE International Conference on Computer Design, Vol. 9, No. 6, 1994:28-37.
8. Schreiber G, Akkermans H, Anjewierden A, de Hoog R, Shadbolt N, Van de Velde W , Welinga B. Knowledge Engineering and Management: The CommonKADS Methodology, MIT Press, 2000 :215-239.
9. (cMorran W. An Introduction to IEC 61970-301 & 61968-11: The Common Information Model. Institute for Energy and Environment Department of Electronic and Electrical Engineering. University of Strathclyde Glasgow, UK. 2007.
10. Schwarz K. IEC 61850, IEC 61400-25, and IEC 61970: Information models and information exchange for electric power systems. Institute for Energy and Environment Department of Electronic and Electrical Engineering. University of Strathclyde Glasgow, UK. 2004.
11. Batarseh F. Gonzalez A.J, Knauf R. Validation of Knowledge-based System Through COMMONKADS, School of Electrical Engineering and computer science, University of Central Florida, 1(1), 2009:1-6.
12. Monica H.C., J.S. Bayona and V.B. Navarro, Applying the Common KADS-RT Methodology to Analyse Real-Time Artificial Intelligence System, Department of informatica, University of EAFIT, 1(1), 2009: 1-10.
13. Haruechaiyasak C. A Comparative Study on Thai Word Segmentation Approaches, Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 5th International Conference on, 2008:125 - 128.
14. Powers DMW. Evaluation: From Precision, Recall and F-Measure to ROC, Informness, Markedness & Correlation. Journal of Machine Learning Technologies 2 (1), February 27, 2011: 37–63.

8/15/2014