MURDOCH RESEARCH REPOSITORY
http://researchrepository.murdoch.edu.au/

This is the author's final version of the work, as accepted for publication following peer review but without the publisher's layout or pagination.

Thanadechteemapat, W. and Fung, C.C. (2011) *Thai word segmentation for visualization of Thai Web sites.* In: International Conference on Machine Learning and Cybernetics, ICMLC 2011, 10 - 13 July, Guilin, China.

http://researchrepository.murdoch.edu.au/6016/

Copyright © 2011 IEEE
Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

THAI WORD SEGMENTATION FOR VISUALIZATION OF THAI WEB SITES

WIGRAI THANADECHTEEMAPAT AND CHUN CHE FUNG

School of Information Technology, Murdoch University, Murdoch, Western Australia
E-MAIL: W.Thanadechteemapat@murdoch.edu.au, L.Fung@murdoch.edu.au

Abstract:

Information overload is a problem in the Information Age and Information visualization is an approach to provide an overview of the content of a web site. Tag cloud is one of the ways to represent information as an image of a group of words. However, there are limitations on tag cloud generation, and one of them is due to the characteristics for the language. In order to extract tags or words for tag cloud, word segmentation is required. This paper proposes a Thai word segmentation approach for the visualization of Thai Web sites. The proposed Thai word segmentation technique is based on the longest matching technique together with a refined corpus. The results of Thai word segmentation are compatible with the results from previous BEST's contests in Thailand.

Keywords:

Thai Word Segmentation, Tag cloud, Web Page Visualization

1. Introduction

Information resources on the Internet have been rapidly expanded around the world. There are a few indicators on the growth of the Internet. The first one is the number of Internet users has drastically increased more than 440% to nearly two billion from 2000 to 2010 [1]. The other indicator is the number of Web sites, which has also increased ten times to approximately around 298 million from 2001 to 2011 [2]. This leads to information overload, which has been considered as a problem in the Information Age. Although search engines have played an important role helping users to look for their required information, the users have to examine the content on each page in order to look for their desired information [3]. Moreover, Web pages normally have not only content, but they also have non-content in other parts such as header, footer, navigation, advertisement, etc. Therefore, the users may require extra time on each page in searching for their expected results.

One approach to assist the understanding of the content is Information Visualization. This is a way to represent the information as a image such as the use of Tag clouds [4]. Some Web sites such as WordCrowd[1], Wordle[2], WordItOut[3] provide a service to generate a tag cloud from text supplied by the users. However, most of the services do not accept URLs as input parameters as references to particular Web pages. There are many steps involved before a tag cloud can be generated from a website. Furthermore, those tag cloud services support only certain languages as there are unique characteristics and features for each language. One of the problems is how to segment the words or phrases in order to present the information in a tag cloud.

There are many research reports on word segmentation techniques in different languages such as English, Chinese, Japanese or Thai. In Thailand, researchers are challenged to work on Thai language processing, and one of the research areas is *Thai word segmentation*. A competition has even been organized in Thailand since 2009 [5]. The competition is called *Benchmark for Enhancing the Standard for Thai Language Processing* (BEST)[4]. The research is a part of a major initiative to enhance Thailand's computing capability and experience among the Thai Internet users, which have grown to roughly 18 million users in 2009 [6]. Growth of the number of Internet users in the last two decades is shown in Figure 1.

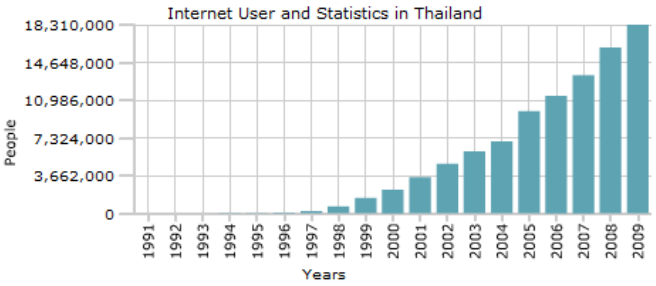

Figure 1. Internet User and Statistics in Thailand

[1] http://tagcrowd.com/
[2] http://www.wordle.net/
[3] http://worditout.com/
[4] http://thailang.nectec.or.th/best/

This paper proposes a Thai word segmentation approach and it is applied for the visualization of Thai Web sites. The objective is to address the fundamental problem of visualization such as Tag cloud for Thai Web sites due to several characteristics in Thai language. The aim of this approach is to reduce the time required by Thai Internet users to locate specific information on particular websites.

This paper starts with a description of the aim of this paper in the introduction. Section 2 provides some of the related work and background of Thai word segmentation techniques as well as web page visualization. Section 3 outlines the proposed approach and Section 4 reports on the results and compared them with other results. The final section concludes this paper with a discussion on future work.

2. Related work

2.1. Background on Thai Word Segmentation

By definition, a word is the smallest language unit that carries specific meaning [7], and it can be usually distinguished based on different features [8] [9]. The first one is orthographic words, which refers to use written marks to segment words such as a space, but it is not an exact criterion. For example, *ice cream* is considered as one word. This feature cannot be applied to Thai language since there is no written mark to segment Thai word, including spaces. The second feature is phonological word, which means using a part of speech or a unit of pronunciation to segment the words. This is most useful in English because one English word has only one main stress. However, this is not applicable to Thai. Next, this feature segments words by using lexical items or items contained in a dictionary. Nevertheless, a lexical item can exist in different grammatical forms. For example, *write* can be shown in different forms such as writes, wrote, written, or writing. There are some other forms as regard to the lexical items such as inflection, derivation, phrasal verb, prepositional verb as well as short forms. As a result, word segmentation is not easy due to the characteristics of the languages, and even manual word segmentation may produce different results [9] [10].

Ideally, Thai word segmentation process should provide results in two types [9]: simple words and compound words. The former refer to words with one minimal meaningful unit of word, also called *morpheme*. Examples are such as ภาพ (image/picture) or รถ (car). Compound words consist of more than one morpheme or one word stem and they may have different meaning when combined. Examples are such as ยินดี (glad : ยิน – hear, ดี – good) or เสียสละ (sacrifice : เสีย – broken, สละ – to discard). However, compound words having not much different meaning should be segmented into multiple words such as คนจน (poor: คน – people, จน – poor) or บทนำ (introduction/preface: บท – chapter/part, นำ – to lead). To verify whether a compound word is a true compound, syntactic criterion can be applied. In other words, structure of its sentence can be applied for indentifying the validity of true compounds. In summary, Thai word segmentation programs should produce results based on minimalist approach since the words can be grouped together easily in later tasks of the language processing.

There are many techniques that can be applied to Thai word segmentation. They can be grouped into two categories which could be dictionary based or non-dictionary based. Dictionary based approaches include techniques such as longest matching [11] [12] [13] [14], maximum matching [12] [13] [14] and decision tree [15]. Non-dictionary based include rule-based [16], Hidden Markov Model (HMM) [15] [17] and Native Bayesian [14]. This paper proposes a technique based on the longest matching technique by using a refined hybrid corpus instead of a dictionary. The technique is described Section 3.

The following section describes information visualization, which is a process after getting results from the Thai word segmentation process.

2.2. Information Visualization

Information visualizing is a means to explore and deliver new insight on large amount of data [18] [19]. There are three steps involved, which are *data preparation*, *data transformation*, and *data visualization* [4].

Tag cloud can be considered as an image to symbolize the frequency of the words in a passage [3],. This can also be considered as a representation of the characteristics of the information [20] . Tag cloud can be generated from predefined tags or words specified by the Web masters or the users. There is no special technique involved as the Web masters or users will have to enter their own tags. Another way to generate tag cloud is to base on words automatically extracted from a passage. Word segmentation is therefore required in the data preparation step for tag cloud generation.

In order to present tag cloud in the data visualization stage, different techniques such as spatial clustering [21] could be adopted to layout the tags on a limited area. This paper displays the tag cloud by arranging the tags sequentially.

3. Proposed Thai Word Segmentation Approach

3.1. Corpus Preparation

This paper uses the BEST 2009 word segmentation corpus, which is [22] referred as *CMain* in this paper. The corpus comprises four genres; novels, academic articles, Thai encyclopedia and online news. The proportions of them are 27%, 21%, 21% and 31%, respectively. There are approximately 5 million words together. Moreover, this paper also uses another corpus called *CName* for name entity which comprises of proper nouns of name of places, people and events. *CName* is provided in the BEST contest's website and it is then mixed with the name entities, abbreviations and poems extracted from *CMain*.

In order to ensure the accuracy of the corpus, a test set named *CTest80* is randomly extracted from 80% of the words from each genre and merged together. The remaining 20% of the words in *CMain* are used for testing the proposed approach. It is called *CTest20*. The objective is to identify any inconsistent words within the corpus. Words in the *CTest20* are firstly combined by removing the separation symbols. They will then appear as a normal text in Thai. The question is whether they can be separated using the *CTest80* corpus.

In order to assess the accuracy of *CTest80,* it is used to segment the combined words in *CTest20*. The results were then saved as a list of segmented words. Each segmented word was compared with the original answer. Any inconsistent or incorrect words were then used to refine the *CTest80* dataset. This concept is similar to that described by Kruengkrai [23]. Examples of some of inconsistent or wrongly segmented words in the corpus are shown in Table1. Consequently, this step is able to improve the correctness of the final result.

TABLE 1. EXAMPLES OF INCONSISTENT WORDS IN THE CORPUS

Original words Segmented as two	Freq	Wrongly Segmented as one word	Freq
ไม่|ได้	6,875	ไม่ได้	3
จะ|มี	4,016	จะมี	4
จะ|เป็น	3,258	จะเป็น	4
ดัง|กล่าว	3,085	ดังกล่าว	1
ชาว|บ้าน	1,719	ชาวบ้าน	12

After *CTest80* is refined, it is reordered according to the length of each word entry, and there are 32,476 items. In the meantime, 44,159 word entries in *CName* are also ordered by the same way. Next, *CTest80* and *CName* are ready for use in the next stage.

3.2. Thai Word Segmentation Technique

The Longest matching technique [11] [12] is applied in this paper, and there are few steps included in order to improve the matching speed. The first is to get each item from *CName* and *CTest80* to match the testing text from left to right. The item will be marked if the item matches in the testing text. There are two steps in marking the item. First, put an annotation at the end of each word and mark the position of each character if it is matched. If the next item is the same, it will not be marked inside the marked position again. In the meantime, the marked positions are counted in order to end the matching process early, if the numbers of unmarked positions are longer than the remaining item in *CName* or *CTest80*. Finally, all symbols are segmented in the last step, and the result of the process is written in a file. The correctness of the result is then measured. The proposed approach is illustrated in Figure 2.

3.3. Measurement

Precision, *recall* and *F-Measure* are used as assessment in this paper. They are the same criteria used in the BEST contest [5] and they are shown in Equation 1 below.

$$F - Measure = \frac{2 \times precision \times recall}{(precision + recall)} \quad (1)$$

Precision refers to the number of correct segmented words produced by the system divided by the number of total segmented words produced by the system. Recall is the number of correct segmented words produced by the system divided by the number of total answered words.

3.4. Visualization of Thai Web Sites

Information from a Web site is downloaded by a crawler, and non-content such as data in the navigation area are removed in order to identify the main content. However, the detailed process of Web content extraction is not described in this paper. The content is then segmented by the word segmentation process based on the corpus, and segmented words are finally produced.

The segmented words are then transformed into a list of words together with their frequency of occurrence. The words in the list are represented as tags in the tag cloud, and the layout of tags is based on the number of words.

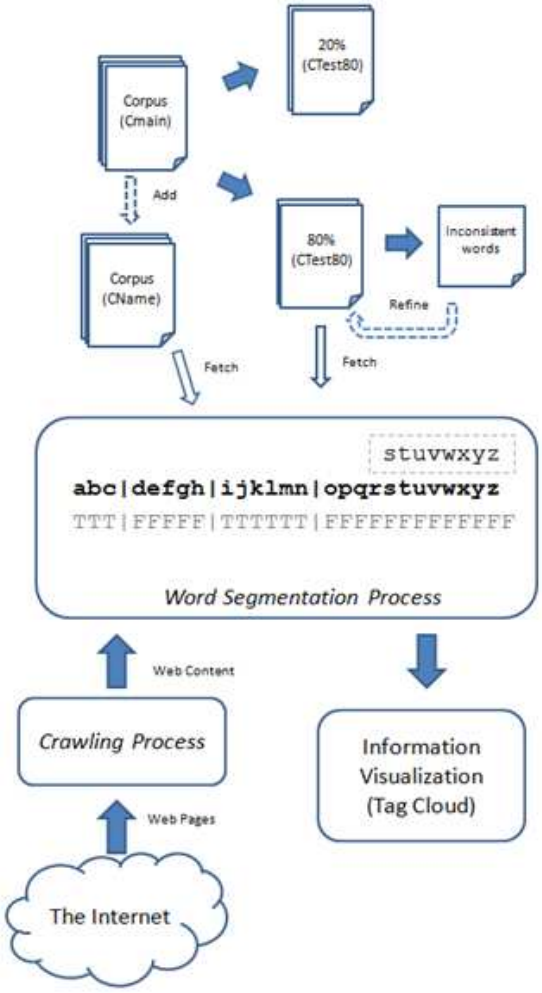

Figure 2. The proposed Thai word segmentation approach

4. Result

As regard to the proposed Thai word segmentation approach, this paper used 20% of the content in *CMain* to assess the proposed approach. However, this approach has not been submitted to the BEST Contest due to implementing was only carried out after the BEST 2009, InterBest2009, and Best2010 contests. Therefore, there is a lack of information on the test set that was being used in the contests. In other words, the results in this paper may not be under the same test criteria as that used in the contest.

However, the final results of the proposed approach are compared with the result of participants from the InterBest2009's Web site [24]. The compared results are shown in Tables 2 to 5. All the results are ordered by the highest F-measure to the lowest F-measure from the top to the bottom. Although the results of proposed approach may not be the best, they are obviously compatible.

TABLE 2. RESULTS ON ACADEMIC ARTICLE DATA SET

Participants	Academic Article Data Set		
	Precision	***Recall***	***F-Measure***
Kruengkrai	98.58	97.10	97.84
Proposed approach	**97.15**	**97.22**	**97.19**
Suesatpanit	96.20	97.26	96.73
Haruechaiyasak	95.71	96.54	96.13
Bangcharoensap	93.12	97.24	95.14
Limcharoen	88.92	94.44	91.60

TABLE 3. RESULTS ON ENCYCLOPEDIA DATA SET

Participants	Encyclopedia Data Set		
	Precision	***Recall***	***F-Measure***
Kruengkrai	98.14	97.26	97.70
Proposed approach	**97.23**	**97.29**	**97.26**
Suesatpanit	96.37	96.60	96.48
Bangcharoensap	93.54	96.52	95.01
Haruechaiyasak	95.15	94.83	94.99
Limcharoen	90.11	94.72	92.36

TABLE 4. RESULTS ON NEWS DATA SET

Participants	News Data Set		
	Precision	***Recall***	***F-Measure***
Kruengkrai	97.78	96.30	97.03
Suesatpanit	94.95	96.13	95.54
Proposed approach	**94.14**	**96.34**	**95.22**
Haruechaiyasak	93.54	94.99	94.26
Bangcharoensap	87.42	94.84	90.98
Limcharoen	77.43	92.61	84.34

TABLE 5. RESULTS ON NOVEL DATA SET

Participants	Novel Data Set		
	Precision	*Recall*	*F-Measure*
Kruengkrai	97.37	96.20	96.78
Haruechaiyasak	95.43	95.72	95.57
Suesatpanit	94.68	95.46	95.07
Bangcharoensap	92.64	96.31	94.44
Proposed approach	**91.82**	**94.13**	**92.96**
Limcharoen	87.03	93.99	90.38

An example of a Thai Web page[5] is shown in Figure 3, and tag cloud for that page is also illustrated in Figure 4. Note that there is no Thai sentence in Figure 4 and there are spaces between the tags or words. The tags or words are ordered by the highest word frequency from the content on the Web page. All words in the tag cloud are automatically segmented by the proposed word segmentation approach.

5. Conclusions and discussion

Information Visualization is an approach to represent information in a graphical manner. At present, there are a lot of information on the Web sites in the Internet, and they are rapidly increasing. A visual group of words known Tag cloud can be used to represent information, but the characteristics of each language may lead to limitations on the tag cloud. The challenge is mainly how to extract the useful words from the information. Therefore, word segmentation is required in order to facilitate the process of tag cloud generation. This paper proposes a Thai word segmentation approach so that it could be applied to information visualization such as tag cloud. The proposed word segmentation is based on the longest matching technique, and it works with a refined corpus instead of a dictionary. Results of the proposed word segmentation are compatible with results from the BEST contest at Thailand. This paper also illustrates tag cloud based on the proposed Thai word segmentation. There are many more ways to improve the approach in this paper. Future improvement could be made on the word segmentation process, but there are also challenges on tag cloud generation and Web content extraction.

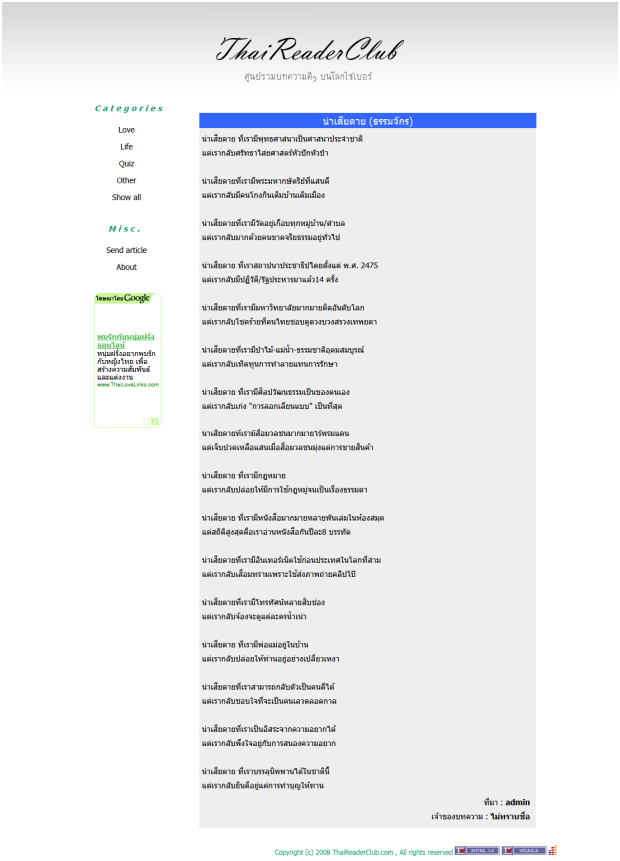

Figure 3. An example of Thai Web page

Figure 4. An example of tag cloud on the content from Figure 3

Acknowledgements

Authors would like to appreciate all staffs and committees in the BEST contest, which provides data resources for public.

[5] http://www.thaireaderclub.com/read.php?id=963 accessed on 31 March 2011

References

[1] World Internet Users and Population Stats. 2010 June 30 [cited 2011 March 30]; Available from: http://www.internetworldstats.com/stats.htm.

[2] Netcraft. March 2011 Web Server Survey. 2011 March 9 [cited 2011 March 30]; Available from: http://news.netcraft.com/archives/2011/03/09/march-2011-web-server-survey.html.

[3] Chun Che Fung, et al. iWISE, an intelligent Web Interactive Summarization Engine. in 2009 International Conference on Machine Learning and Cybernetics. 2009.

[4] Chun Che Fung and Wigrai Thanadechteemapat. Discover Information and Knowledge from Websites Using an Integrated Summarization and Visualization Framework. in Third International Conference Knowledge Discovery and Data Mining, 2010. WKDD '10. 2010.

[5] Kosawat, K., et al. BEST 2009 : Thai word segmentation software contest. in Natural Language Processing, 2009. SNLP '09. Eighth International Symposium on. 2009.

[6] Statistics summary of the Internet in Thailand. 2010 August [cited 11 September 2010]; Available from: http://internet.nectec.or.th/webstats/home.iir?Sec=home.

[7] Longman English Dictionary Online. 2011 [cited 31 March 2011]; Available from: http://www.ldoceonline.com/dictionary/word_1.

[8] Trask, Larry. What is a word? 2004 [cited 30 March 2011]; Available from: http://www.sussex.ac.uk/linguistics/documents/essay_-_what_is_a_word.pdf.

[9] Aroonmanakun, Wirote. Thoughts on word and sentence segmentation in Thai. in the Seventh Symposium on Natural Language Processing. 2007. Pattaya, Thailand: Citeseer.

[10] Aroonmanakun, Wirote, Collocation and Thai Word Segmentation, in the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop. 2002. p. 68--75.

[11] Poowarawan, Yuen. Dictionary-based Thai Syllable Separation. in the Ninth Electronics Engineering Conference. 1986.

[12] Meknavin, S., et al. Feature-based Thai word segmentation. in the Natural Language Processing Pacific Rim Symposium 1997. 1997. Phuket, Thailand: Citeseer.

[13] Promchan, Pisit and Yunyong Teng-Amnuay. Performance Comparison of Thai Word Separation Algorithms. in the National Computer Science and Engineering Conference 1998 (NCSEC'98). 1998. Bangkok.

[14] Haruechaiyasak, Choochart, et al. A comparative study on Thai word segmentation approaches. in Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on. 2008.

[15] Bheganan, Poramin, et al., Thai Word Segmentation with Hidden Markov Model and Decision Tree, in Advances in Knowledge Discovery and Data Mining. 2009. p. 74-85.

[16] Sutheebanjard, P. and W. Premchaiswadi. Thai personal named entity extraction without using word segmentation or POS tagging. in Natural Language Processing, 2009. SNLP '09. Eighth International Symposium on. 2009.

[17] Jaruskulchai, Chuleerat, An automatic Thai lexical acquisition from text, in PRICAI'98: Topics in Artificial Intelligence. 1998. p. 436-447.

[18] Eppler, Martin J and Remo A. Burkhard, Knowledge Visualization. 2004, NetAcademy Project: Switzerland.

[19] Burkhard, Remo Aslak, Knowledge Visualization - The Use of Complementary Visual Representations for the Transfer of Knowledge. A Model, a Framework, and Four New Approaches. 2005, Swiss Federal Institute of Technology Zurich.

[20] McKie, S. Scriptclud.com: Content Clouds for Screenplays. in Semantic Media Adaptation and Personalization, Second International Workshop on. 2007.

[21] Slingsby, A., et al. Interactive Tag Maps and Tag Clouds for the Multiscale Exploration of Large Spatio-temporal Datasets. in Information Visualization, 2007. IV '07. 11th International Conference. 2007.

[22] Boriboon, M., et al. BEST Corpus Development and Analysis. in Asian Language Processing, 2009. IALP '09. International Conference on. 2009.

[23] Canasai Kruengkrai, et al., A Word and Character-Cluster Hybrid Model for Thai Word Segmentation, in InterBEST 2009 Thai Word Segmentation: an International Episode. 2009: Bangkok.

[24] Workshop, InterBEST2009: Thai Word Segmentation. Paper's result. 2009 [cited 1 November 2010]; Available from: http://thailang.nectec.or.th/interbest/index.php?option=com_content&task=view&id=17&Itemid=32.