PERFORMANCE COMPARISON OF THAI WORD SEPARATION ALGORITHMS

Pisit Promchan (pisitp@yahoo.com)
Telecom Asia Corp. Public Co. Ltd.
Yunyong Teng-amnuay (Yunyong.T@Chula.ac.th)
Department Of Computer Engineering, Chulalongkorn University
Bangkok 10330, Thailand

ABSTRACT

A performance comparison of word-separation algorithms for Thai language is presented. The research surveyed existing algorithms. A synthesis of performance indicators was attempted together with a development of measurement methodology. A body of Thai reference data was collected to validate the accuracy of Thai word separation. Experimental results show that the longest-word pattern-matching algorithm gives the most accurate output words while the backtracking algorithm gives the least error words. Word-usage-frequency algorithm gives the highest valid words ratio per number of words in its dictionary. The usage of ambiguity dictionary gives the best ambiguous case resolution, whereas the shortest-word pattern-matching algorithm gives the highest number of output words.

INTRODUCTION

Thai and some other Asian languages have no explicit word boundary in the written sentences. Consider the example, “i fyoucanreadthis”. Thus, the processing for these type of languages have the prerequisite in word separations. Thai word separation is not obvious because of ambiguity. There are many possibilities in separating words from one sentence [1], [4], [5], [8]. For example, consider this contiguous sentence “I mportantproductsofregion” may be incorrectly separate as “Im-port-ant-product-so-fregion” or correctly separate as “Important-products-of-region”.

EXISTING ALGORITHMS

Shortest-word pattern-matching This algorithm finds shortest words [8]. We simulate this methodology using linked list of sorted words by ascending word length as the matching dictionary.

Longest-word pattern-matching The longest words are selected in this algorithm, [4], [8]. The same search algorithm and the data structure as the previous algorithm are also used for simulating this method but with descending word order in the dictionary.

Word-usage-frequency This algorithm selects words according to their usage frequencies [5]. We built the dictionary for this algorithm using the longest-word pattern-matching algorithm to separate the 1.3-Mbyte Thai corpus, count and remove the repeated words, then sort descendingly by repetitive value (usage frequency).

Backtracking This algorithm is the longest-word pattern-matching with backtracking to shorter words in case of error in subsequent words [14], [11].

Maximal-matching The longest or shortest approaches are not always best in resolving ambiguity. The maximal-matching algorithm minimizes the ambiguity by choosing the minimum number of words to form the sentence [11], [12].

Ambiguity dictionary The above algorithms are still not adequate to deal with ambiguity in Thai language. For example, in the sentence: “¤³Ð¡ÃÃÁ¡ÒÃ¡ÃÁ¾ÅÐÈ¡ÉÒÍÂ¡ÃÒ§”, none of the above algorithms can separate words correctly. This can be corrected using the algorithm with ambiguity dictionary containing the ambiguous cases in Thai language such as the mentioned sentence. The ambiguous cases are extracted from 500 lines or 20 pages of Thai corpus in this research. This algorithm lookups the ambiguity dictionary first and then uses the longest-word pattern-matching for ordinary cases.

PERFORMANCE METRICS

Separatibility is the number of words from each algorithm which contains both valid and invalid words.

Word validity is the number of valid word from each algorithm. The valid words are words exist in the reference dictionary mentioned later.

Valid ratio to word separated is the ratio of valid words to separated words .

Valid ratio compared to Longest is the ratio of valid words over the valid words from the longest-word pattern-matching. From the experimental result, the longest-word pattern-matching appears to be the upper bound of all algorithms (which does not mean it is the best).

Valid ratio compared to Shortest is the ratio of valid words to the valid words from the shortest-word pattern-matching.

Similarly, the shortest-word pattern-matching appears to be the lower bound.

Valid ratio per number of words in its dictionary where each algorithm has its own dictionary with the different sets of words.

Error ratio is measured by number of error words per words output by each algorithm.

Ambiguous cases error indicates the number of error for ambiguous cases.

Resources utilization is based on the memory size used to hold the dictionary.

MEASUREMENT METHODOLOGY

The measurement has done by proceeding these steps, running the word separation programs against the reference Thai corpus. The separate character (hex 04) is inserted. The output is reformatted using each word per line. Then, sort the reformatted output and remove the repeat words. The reference dictionary is used to validate the accuracy of Thai words separation. The ambiguous cases error is manually reviewed for the 500 lines or 20 pages of input Thai text. The statistics data is collected along with each step for further analysis.

REFERENCE DATA

The Thai input corpus used for the experiment comes from two sources i.e. general Thai data from various subject including National economics and society development master plan, Research list, etc. It is 6500 lines in size. Another one come from the Thai Data Bank which the subject of computer journal has selected for the measurement. This source of input data is used to compare the difference of each set of experimental results. Furthermore, we have another reference data, that is the reference dictionary. This dictionary is constructed by Mr. D. Cooper at Southeast Asian Software Research Center, Bkk, http://seasrc.th.net/sealang, based on "R ajabandithayasatharn", the Royal Thai language standard organization. It contains 17,889 Thai words.

EXPERIMENTAL RESULTS

Experimental data were collected at 500-line intervals up to the 6,500 lines limit of the Thai corpus. The x-axis is number of line in the input Thai corpus while the y-axis is the performance result.

The shortest-word pattern-matching algorithm produces the highest separatibility while the backtracking is the worst as shown in Fig. 1. Fig. 2 demonstrates that the longest-word pattern-matching algorithm gives the highest number of valid output words while the shortest-word pattern-matching is the worst. The ratio of valid words to separated words is illustrated in Fig. 3 where backtracking algorithm performs best and shortest-word pattern-matching is the worst. The ratio of valid words to words from longest-word pattern-matching is demonstrated in Fig 4 where backtracking is the best while shortest-word pattern-matching is the worst. fig.5. Longest-word pattern-matching gives the best valid ratio compared to shortest-word pattern-matching while word usage frequency is the worst as shown in Fig. 5. Word usage frequency algorithm produces the best valid ratio per number of words in its dictionary as illustrated in Fig. 6. The shortest-word pattern-matching is the worst. The error ratio is demonstrated as Fig. 7, backtracking algorithm produce the lest error while the shortest-word pattern-matching is the worst. Fig. 8. presents the ambiguous cases error, the ambiguity dictionary algorithm produces the lest ambiguous cases error while the shortest-word pattern-matching is the worst. The word usage frequency algorithm is the best case if we consider the resource utilization perspective. The ambiguity dictionary consume resource more than all others caused of the additional dictionary to resolve ambiguous cases as illustrated in Fig. 9.

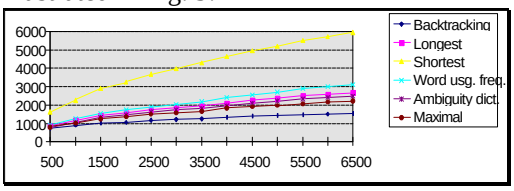

Fig. 1. Separatibility

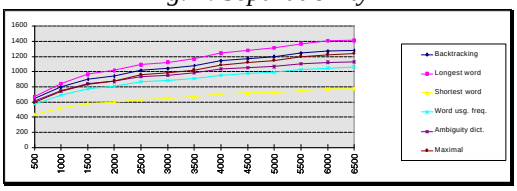

Fig. 2.Word validity

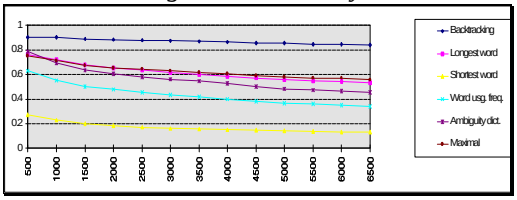

Fig. 3. Valid ratio/ words separated

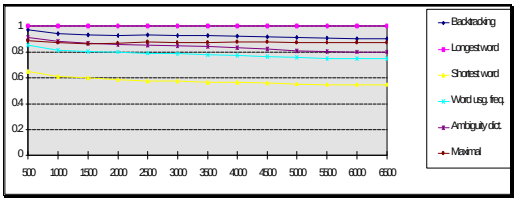

Fig. 4. Valid ratio compared to Longest

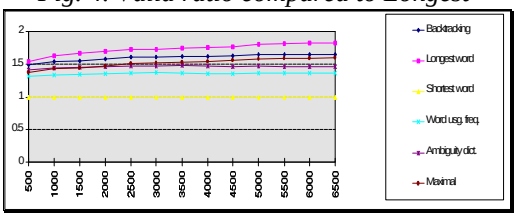

Fig. 5. Valid ratio compared to Shortest

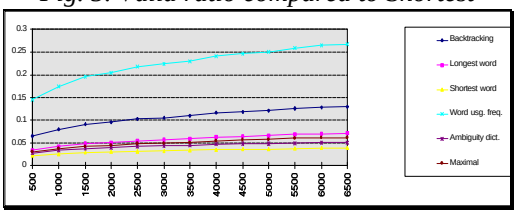

Fig 6. Valid ratio/ number of words in its dictionary

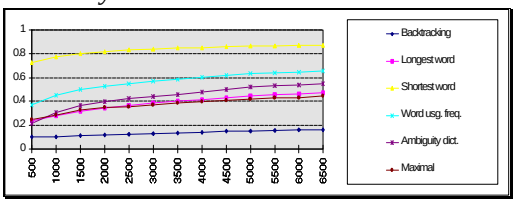

Fig 7. Error ratio

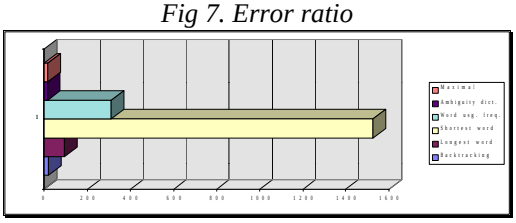

Fig 8. Ambiguous cases error

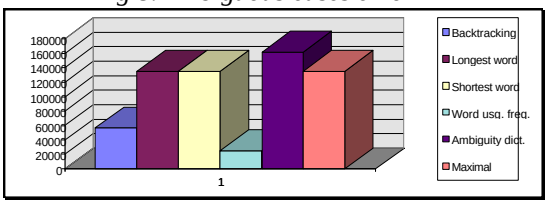

Fig 9. Resource utilization

CONCLUSION

The usage of word separation algorithms differs depending on applications. Some types of language processing require high degree of accuracy, such as text-to-speech program and spell checker, while some others do not, such as document formatting. From the test result, the shortest-word pattern-matching algorithm is not suitable for any kind of processing due to its poor performance. The word-usage-frequency algorithm needed to be enhanced for some type of processing. The maximal-matching and the longest-word pattern-matching algorithm are adequate for document formatting and indexing but do not meet the requirements of complexed tasks like text-to-speech and spell checker due to ambiguous problem. The ambiguity-dictionary approach is the only one adequate for those types of complexed processing. The further study of ambiguous cases in Thai language is recommended in order to improve the ambiguity dictionary.

REFERENCES

[1] R. Rattayanontr, *Data Structure for Electronics Dictionary.* Master Thesis, Chulalongkorn University, 1992.

[2] B. Thanasunthornpaisarn, *Design and Development of Thai Word Separation Interfaces.* Master Thesis, Chulalongkorn University 1990

[3] Y. Puvorawan and V. Imarrom, *Thai Spelling Checker.* Paper, Electrical Engineering Conference, 1987

[4] R. Varakulsiripan, J Ngamvivit, S. Junvan, S. Jivatayakul and S. Tipjuksurat, *Thai Word Separation Using Longest Word Pattern Matching.* Papers on Natural Language Processing, Compiled by V. Sornlertlamvanich, 1995

[5] R. Varakulsiripan, W. Suchaichit, S. Junvan and S. Tipjaksurat, *Word Usage Frequency Algorithm.* Papers on Natural Language Processing, Compiled by V. Sornlertlamvanich, 1995

[6] T. Koanantakul, *Basic Standard for Thai Information Technology.* National Electronics and Computer Technology Center (NECTEC), 1991

[7] P. Zuill, P. Laverick, B. Kirk, D. Ramos, *Thai S370 Assembly Routines in Computerized Customer Services System (CCSS).* Summarized by P. Promchan, ISS, TelecomAsia, 1995

[8] N. Thongpumpurksar *A Thai Text Retrieval System Using Digital Search Trees and SQL.* Computer Science Master Thesis, Asian Institute of Technology, 1993

[9] M. Allen Weiss, *Data Structure and Algorithm Analysis in C.* The Benjamin/ Cummings Publishing Company, Inc., 1993

[10] A. V. Aho, J. E. Hopcroft, Jeffery D. Ullman, *The Design and Analysis of Computer Algorithms.* Addison Wesley Publishing Company, 1974

[11] S. Meknawin, P. Charoenpornsawat, Boonserm Kijsirikul *Feature-based Thai Word Segmentation.* NLPRS' 97 Proceedings of the National Language Processing Pacific Rim Symposium 1997

[12] W. Kanlayanawat, S. Prasitjutrakul, *Autometic Indexing for Thai Text with Unknown Words using Trie Structure.* NLPRS' 97 Proceedings of the National Language Processing Pacific Rim Symposium 1997

[13] AI Research and Development Center, *Thai Data Bank.* Machine Translation System Laboratory Center of the International Cooperation for Computerization, KMITT, Technical Report 1995

[14] S. Rareunrom, *Dictionary-based Thai Word Separation.* Senior project, Dept. of Computer Engineering, Chulalongkorn University, 1991

[15] P. Promchan, *Analysis of Guidelines for Performance Comparison of Thai Word Separation Programs.* Master Thesis, Dept. of Computer Engineering, Chulalongkorn University, 1997