INFORMATION EXTRACTION FROM MALAY VEHICLE ADVERTISEMENT TEXT USING NATURAL LANGUAGE PROCESSING TECHNIQUES

NORFADILA BINTI MAHROM

UNIVERSITI UTARA MALAYSIA (2007)

PUSAT PENGAJIAN SISWAZAH
(*Centre For Graduate Studies*)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

NORFADILA MAHROM

calon untuk Ijazah
(candidate for the degree of) **MSc. (Intelligent System)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/ her project paper of the following title)

INFORMATION EXTRACTION FROM MALAY VEHICLE ADVERTISEMENT TEXT USING NATURAL LANGUAGE PROCESSING TECHNIQUES

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory knowledge of the field is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **DR. SHAIDAH JUSOH**

Tandatangan
(Signature) : Shaidah

Tarikh
(Date) : 10 December 2007

INFORMATION EXTRACTION FROM MALAY VEHICLE ADVERTISEMENT TEXT USING NATURAL LANGUAGE PROCESSING TECHNIQUES

A thesis submitted to the Faculty of Information Technology

in partial fulfillment of the requirements for the degree

Master of Science (Intelligent System)

Universiti Utara Malaysia

By

Norfadila Binti Mahrom

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor, in her absence, by the Dean of the Faculty of Information Technology. It is understood that any copying or publication or use of this thesis or parts there of for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of material in this thesis, in whole or in part, should be addressed to:

Dean of Faculty of Information Technology

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Darul Aman

ABSTRAK

Sebahagian iklan kenderaan disiarkan dalam bentuk dokumen teks dan ianya memerlukan pembaca membaca keseluruhan dokumen, dan seterusnya memahami isi kandungan dokumen tersebut sebelum maklumat penting dapat dikeluarkan. Proses ini semestinya mengambil masa yang lebih lama berbanding jika mempunyai sebuah sistem yang dapat mengeluarkan maklumat penting daripada dokumen tersebut secara automatik tanpa memerlukan pembaca membaca keseluruhan dokumen. Dalam kajian ini, sebuah sistem prototaip telah dibangunkan untuk membantu pembaca mengeluarkan isi-isi penting dari iklan kenderaan berbahasa Melayu dengan mengaplikasikan teknik pemprosesan bahasa natural (NLP). Teknik pemprosesan bahasa natural yang telah digunakan fokus kepada proses sintaksis.

ABSTRACT

Some of the vehicle advertisements are represented in textual documents, and it requires a reader to read the entire document and understands its content before the important information can be extracted. This process consumes more time instead of having a system that can extract the important information from the document automatically without the reader needs to read the whole document. In this study, a prototype system was developed to assist a reader to extract important information from the Malay vehicle advertisement by applying natural language processing (NLP) techniques. The NLP techniques that have been used are focused on the syntactic processing.

ACKNOWLEDGEMENTS

I would like to express my thanks and gratitude to Allah, the Most Beneficent, the Most Merciful whom granted me the ability and willing to start and complete this project. I pray to his greatness to inspire and enable me to continue the work for the benefits of my country, specifically for educational institutions.

I also would like to show my appreciation to my supervisor, Dr. Shaidah Jusoh whose help, giving suggestion and encouragement to me at all time during the development of this project and writing the proceeding paper and report of this project. The valuable guidance from her has made this project come true and success.

My special thanks to my family, especially to my sister, Nor 'Aini Mahrom who has helped me to understand about the Malay grammar and always encourage me, giving advice and love. Finally, I would like to thanks to all my friends who always beside me, understand and supported me towards the completion of this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Acronym	Meaning
BLOB	Binary Large Objects
CFG	Context-Free Grammar
Det	Determinant
DISCOTEX	Discovery from Text Extraction
GUI	Graphical User Interface
IE	Information Extraction
IS	Imperative Sentence
IT	Information Technology
KDD	Knowledge Discovery from Databases
KDT	Knowledge Discovery from Texts
LR	Left-Right
MITA	MetLife's Intelligent Text Analyzer
NC	Noun Complement
NCL	Noun Complement List
NLP	Natural Language Processing
NP	Noun Phrase
POS	Part-of-Speech
PP	Prepositional Phrase
Prep	Preposition
RAM	Read Access Memory
S	Sentence
TAGs	Tree Adjoining Grammars
VB	Visual Basic
VP	Verb Phrase

LIST OF APPENDICES

CHAPTER ONE

INTRODUCTION

This chapter briefly presents the main idea of this study which is information extraction using natural language processing technique. Natural language processing technique that is focused in this study is syntactic processing. In addition, this chapter discusses the problem statement, objectives, scope and the significance of the study.

1.1 Overview of the study

With the rapid progress of computer and other science technologies, a demand for finding and discovering interesting and relevant knowledge from any kind of information become high. However, most of the valuable information is stored in text form.

Usually, information is presented in two different manners, which are structured and unstructured manners. Structured information refers to data that is stored in fixed fields

The contents of the thesis is for internal user only

REFERENCES

Abd Rahman, S. (2006). *Automated Document Preprocessing for Text Categorization System (TCS).* Thesis for the degree Master of Science (Information Technology), Universiti Utara Malaysia, Kedah.

Ahmad, F., Yusoff, M., & Sembok, T. M. T. (1996). Experiments with a Stemming Algorithmm for Malay Words. *Journal of the American Society for Information Science*, 47(12), pp. 909-918.

Alonso, M. A., Cabrero, D., & Vilares, M. (1999). *Generalized LR Parsing for Extensions of Context-Free Grammars*, Universidade da Coruna.

Arimura, H., Abe, J., Fujino, R., Sakamoto, H., Shomozono, S., & Arikawa, S. (2001). *Text Data Mining: Discovery of Important Keywords in the Cyberspace*, pp.220-226.

Arnold, D. (2000). Chart Parsing. Retrieved October 10, 2007 from http://www.cs.ualberta.ca/~lindek/650/papers/chartParsing.pdf

Beckstein, C., & Kim, M. (1989). *A Mixed Top-Down and Bottom-Up Deduction Method and its Correctness*. IBM Research Division, T.J. Watson Research Center, Yorktown Heights, New York.

Brill, E. (1992). A Simple Rule-Based Part-of-Speech Tagger. In *Proceeding of the 3rd Conf. on Applied Natural Language Processing.*

Cutting, D., et al. (1992). A Practical Part-of-Speech Tagger. In *Proceeding of the 3rd Conf. on Applied Natural Language Processing.*

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *Proceeding of the 32nd Annual Meeting of the Association for Computational Linguistics.*

Erbach, G. (1990). *Syntactic Processing of Unknown Words*. In Ed. P. Jorrand and V. Sgurev (Eds), Artificial Intelligence IV – Methodology, Systems, Applications, North-Holland, Amsterdam, 1990.

Feldman, R., & Hirsh, H. (1997). Finding Associations in Collections of Text. In Michalski R.S., Bratko I. and Kubat M. (edts); *Machine Learning, data Mining and Knowledge Discovery: Methods and Application* (John Wiley and sons Ltd).

Gao, X., Murugesan, S., & Lo, B. (2005). Extraction of Keyterms by Simple Text Mining for Business Information Retrieval. In *Proceeding of the 2005 IEEE International Conference on e-Business Engineering (ICEBE'05).*

Glasgow, B., Mandell, A., Binney, D., Ghemri, L. & Fisher, D. (1997). *MITA: An Information Extraction Approach to Analysis of Free-form Text in Life Insurance Applications*. American Association for Artificial Intelligence (www.aaai.org).

Goodman, J. T. (1998). *Parsing Inside-Out*., Harvard University, Cambridge, Massachusetts.

Idris, N., & Syed Mustapha, S. M. F. D. (2001). *Stemming for Term Conflation in Malay Texts*. Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur.

Ishikawa, H., Kubota, K., Noguchi, Y., Kato, K., Ono, M., Yoshizawa, N., & Kanaya, A. (1998). *A document warehouse: a multimedia database approach.*

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing*. Prentice Hall, United States of America.

Jusoh, S., Wang, F., & Yang, S. X. (2004). Integrating Fuzzy Approach and History Knowledge to Create an Intelligent Processor for A Human-Robot Interface. *The International Journal of Artificial Intelligence and Machine Learning*, ICGST, No. 12, pp. 7-14, 2004.

Karim, N. S., Onn, F. M., Musa, H., & Mahmod, A. H. (2004). *Tatabahasa Dewan Edisi Baharu*.Dewan Bahasa dan Pustaka, Kuala Lumpur.

Kushmerick, N., Johnston, E., & McGuinness, S. (2000). *Information Extraction by Text Classification*. Smart Media Institute, Computer Science Departmennt, University College Dublin.

Lecture for Natural Language Processing. Retrieved September 3, 2007 from www.cs.jhu.edu/~jason/papers/eisner.earley-anim.ppt

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., & Soderland, S. (1992). *Evaluating an Information Extraction System*. Computer Science Department,

LGRC, University of Massachusetts and Department of Computer Science, Berkeley, CA.

McCallum, A. (2005). Information Extraction : Distilling Structured Data from Unstructured Text. *ACM QUEUE* November 2005.

Mooney, R. J., & Nahm, U. Y. (2003). Text Mining with Information Extraction. Multilingualism and Electronic Language Management: *Proceedings of the 4th International MIDP Colloquium*, September 2003, Bloemfontein, South Africa. Pp. 141-160.

Othman, A. (1993). Pengakar perkataan Melayu untuk system capaian dokumen. Thesis for the degree Master of Science, Universiti Kebangsaan Malaysia, Bangi.

Parsing More Efficiently and Accurately. Retrieved September 3, 2007 from www1.cs.columbia.edu/~julia/cs4705/earley.ppt

Parsing: Chart Parsing, Earley-Algorithm, top down/ bottom up left-right. Retrieved September 3, 2007 from http://www.coli.uni-saarland.de/~hansu/EarleyAlgorithm.pdf

Rajman, M., & Besancon, R. (1997). *Text Mining: Natural Language techniques and Text Mining applications*. Computer Science Department, Swiss Federal Institute of Technology.

Rao, R. (2003). From unstructured data to actionable intelligence. *IEEE Computer Society.*

Saian, R. (2004). *Stemming Algorithm in Searching Malay Text*. Thesis for the degree Master of Science (Information Technology), University Utara Malaysia, Kedah.

Schabes, Y., & Joshi, A. K. (1988). *An Earley-Type Parsing Algorithm for Tree Adjoining Grammars*, Department of Computer and Information Science, University of Pennsylvania, Philadelphia. Pp.256-269.

Sehgal, A. K. (2000). *Text Mining: The Search For Novelty In Text*. Thesis for the degree of Doctor of Philosophy.

Soubeh, M., & Al-Laban, M. J. (1996). Context-Free Relations and Their Characteristics. *Applied Mathematics And Computation*, Elsevier Science Inc., North-Holland, pp. 163-172, 1996.

WIKIPEDIA The Free Encyclopedia. Earley Parser. Retrieved September 3, 2007 from http://en.wikipedia.org/wiki/Earley_parser

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with Java implementations.* Morgan Kaufmann, San Francisco, CA.

Witten, I. H. (2004). *Text mining*. University of Waikato, Hamilton, New Zealand.