3737096 SCCS/M : MAJOR : COMPUTER SCIENCE ;
M.Sc. (COMPUTER SCIENCE)
KEY WORDS : HEURISTIC RULE, TEXT PARSING, DOMAIN LEXICON, MACHINE LEARNING

KULLAWAT WUTTISUNKORNSAKUL : TEPT : A TOOL USING HEURISTIC RULES IN MACHINE LEARNING FOR DOMAIN SPECIFIC RESOURCES ACQUISITION AND TEXT PARSING. THESIS ADVISORS : JARERNSRI L. MITRPANONT, Ph.D., SUPACHAI TANGWONGSAN, Ph.D.
67 P. ISBN 974-663-230-2

This thesis studies another approach of Natural Language Processing (NLP) supporting text parsing via Machine Learning controlled by a set of heuristic rules. Information flooding on the Internet and lack of a domain specific parsing tool, require domain specialists to waste time manually selecting domain information. Lexicons are the major resource used to select domain documents. Specialists need a suitable and efficient tool to extract domain keywords such as technical terms that are not in a standard dictionary. In addition, this tool should be easily ported for use in any new domain. This study proposes a Text Efficient Parsing Tool (TEPT) as a new NLP tool to assist domain specialists in extracting keywords from domain documents. The TEPT system in this study applies heuristic rules and machine learning techniques to acquire domain specific resources and lexicons which are essential for parsing. This Dynamic Learnable Lexicon Feature is one of the most outstanding feature of the TEPT. It can be used to self-develop a set of lexicons in any trained domain. Using this TEPT, essential information or domain keywords can be extracted from parsed text via the domain lexicon. Study results demonstrate that after training TEPT it can develop lexicons in Computer Science domain. These results suggest it can be used effectively to extract keywords from other specific domains.