TEPT : A TOOL USING HEURISTIC RULES IN MACHINE LEARNING FOR DOMAIN SPECIFIC RESOURCES ACQUISITION AND TEXT PARSING

KULLAWAT WUTTISUNKORNSAKUL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
1999

ISBN 974-663-230-2

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.

3737096 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)

กุลวัฒน์ วุฒิสารกรสกุล : TEPT : เครื่องมือที่ใช้ Heuristic Rules ในการเรียนรู้ด้วยเครื่องเพื่อการสร้างปัจจัยหลักในการวิเคราะห์ไวยากรณ์ภาษาและใช้งานใน Domain เฉพาะที่ผู้ใช้สนใจ (TEPT : A TOOL USING HEURISTIC RULES IN MACHINE LEARNING FOR DOMAIN SPECIFIC RESOURCES ACQUISITION AND TEXT PARSING). คณะกรรมการควบคุมวิทยานิพนธ์ : เจริญศรี มิตรภานนท์, Ph.D., ศุภชัย ตั้งวงศ์ศานต์, Ph.D. 67 หน้า. ISBN 974-663-230-2

วิทยานิพนธ์ฉบับนี้ได้ศึกษาวิจัยถึงการใช้แนวทางของ NLP ในการทำ Parsing โดยใช้เทคนิคการเรียนรู้ด้วยเครื่องร่วมกับ Heuristic Rules เนื่องจากการที่ข้อมูลสารสนเทศจำนวนมหาศาลมีการไหลเวียนบนเครือข่าย Internet และการขาดแคลนเครื่องมือที่ใช้ในการคัดเลือกเอกสารที่ตรงตาม Domain เฉพาะสาขา ทำให้ Specialists ในสาขาเฉพาะด้านต้องเสียเวลาไปในการคัดเลือกข้อมูลสารสนเทศที่ตรงกับ Domain เฉพาะของตน สิ่งเหล่านี้ได้ทำให้ Specialists ในแต่ละ Domain มีความต้องการเครื่องมือที่เหมาะสมและมีประสิทธิภาพเพื่อช่วยในการค้นหา Domain Keywords ในเอกสารโดยเฉพาะอย่างยิ่ง Domain เฉพาะที่เกี่ยวข้องกับ Keywords ประเภทศัพท์เทคนิคต่างๆซึ่งมิได้ปรากฏอยู่ในพจนานุกรมมาตรฐานทั่วๆไป นอกจากนั้นเครื่องมือนี้ควรมีความสามารถในการนำไปใช้กับ Domains อื่นได้ง่ายด้วย การศึกษาวิจัยนี้ได้นำเสนอเครื่องมือวิเคราะห์ไวยากรณ์ภาษาที่มีประสิทธิภาพในเอกสารต่างๆที่เรียกว่า TEPT หรือ Text Efficient Parsing Tool โดย TEPT ประยุกต์ใช้เทคนิค Heuristic Rules และ Machine Learning ในการวิเคราะห์ไวยากรณ์ภาษาในเอกสาร และสามารถสร้างพจนานุกรมซึ่งจำเป็นต่อการวิเคราะห์ไวยากรณ์ภาษาได้เองโดยอัตโนมัติ ซึ่งถือเป็นคุณลักษณะที่สำคัญของ TEPT การสร้างพจนานุกรมที่มีการหมุนเวียนเคลื่อนไหวของคำศัพท์อยู่ตลอดเวลาและสามารถเรียนรู้คำศัพท์ใหม่ๆได้ด้วยตัวเองนี้มีประโยชน์ต่อการคัดเลือกข้อความสำคัญในเอกสารในแต่ละ Domain ที่ผู้ใช้สนใจ การใช้ Heuristic Rules เพื่อควบคุมขั้นตอนของ Machine Learning ในการสร้างพจนานุกรมก็ดี หรือในการทำ Text Parsing ก็ดี วิธีนี้ทำให้ได้พจนานุกรมในแต่ละหัวข้อเรื่องที่ผู้ใช้สนใจโดยขึ้นอยู่กับเนื้อหาในเอกสารเหล่านั้นเป็นสำคัญ และท้ายที่สุดข้อมูลสารสนเทศที่สำคัญและตรงกับ Domain ในเอกสารก็จะถูกคัดเลือกออกมา จากผลการวิจัยแสดงให้เห็นว่า TEPT สามารถคัดเลือก Keywords ใน Computer Science Domain ได้อย่างมีประสิทธิภาพภายหลังจากผ่านการสร้างพจนานุกรมมาแล้ว