Heuristics-based Entity-Relationship Modelling through Natural Language Processing

by

Nazlia Omar B.Sc.(Hons), M.Sc.

Faculty of Engineering
University of Ulster

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

September 2004

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

I wish to thank both my supervisors, Professor Paul Mc Kevitt and Dr. Paul Hanna for their constant support, guidance and assistance throughout this work. Also, I would like to thank the staff and colleagues at the School of Computing and Mathematics for their help and encouragement. I would also like to extend my gratitude to the helpful comments made by Dr. Ji Ming, Professor Michael McTear, Professor Sally McClean and Professor Terri Anderson. A special thank to my family who have been very supportive throughout my study.

I would also like to acknowledge the financial support from University Kebangsaan Malaysia (UKM), for the funding of this study, without which it would not have been possible.

ABSTRACT

Entity-relationship (ER) modelling, which is a high level conceptual model designed to facilitate database design, can be a daunting task to both designers and students alike due to its abstract nature. Much research has attempted to apply natural language processing (NLP) to extract knowledge from requirements specifications to aid this modelling process. However, research on the formation and use of heuristics to aid the construction of logical databases from natural language has been scarce. In general, human experts draw on their own heuristics to decide whether something should be represented as an entity or a relationship, for instance, in a conceptual model. The main goal of this thesis is to introduce new heuristics to assist this process, and apply them in the automatic processing of natural language and its transformation to Entity-Relationship (ER) models.

This thesis first examines, the current techniques, tools and heuristics in generating conceptual models from natural language. Problems associated with natural language such as ambiguities are also investigated. A parser that is chosen for this research, Memory-Based Shallow Parser (MBSP) is reviewed. What stems from this examination is the formation of new heuristics that can be utilized to assist the database modelling process, through the natural language processing of requirements'specifications and generation of ER models. To realize the utilities of these heuristics, a tool called *ER-Converter* is implemented. ER-Converter has been evaluated in blind trials against a test dataset, which consists of 30 database problems. New measures, in addition to standard measures *recall* and *precision*, are defined. Results generated by ER-Converter are evaluated against human performance and other existing systems' results. ER-Converter has an average of 90% recall and 85% precision and the results compare favourably with other systems. In addition, ER-Converter requires very little user intervention with an average of only 1.6% in the test dataset. The evaluation results are discussed and demonstrate that ER-Converter could be used, for example, within the domain model of a multimedia intelligent tutoring system, designed to assist in the learning and teaching of databases.

Note on access to contents