

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	vii
List of Tables	viii
Acknowledgements	ix
Abstract	x
Note on access to contents	xi
1. Introduction.....	1
1.1. Overview of data modelling.....	2
1.2. Difficulty in ER Modelling.....	3
1.2.1. Analysis on the difficulty of Database Modelling subject.....	6
1.3. Objectives of research.....	8
1.4. Thesis structure.....	9
2. Literature review.....	11
2.1. Application of natural language processing (NLP) to database design.....	11
2.1.1. ANNAPURNA.....	12
2.1.2. VCS (View Creation System).....	14
2.1.3. Tseng et al. (1992).....	15
2.1.4. DMG (Data Model Generator).....	16
2.1.5. FORSEN.....	17
2.1.6. RADD (Rapid Application and Database Development).....	19
2.1.7. COLOR-X.....	21
2.1.8. E-R Generator.....	23
2.1.9. CM-Builder.....	26
2.2. Summary of systems that apply natural language processing to database design.....	27
2.3. Heuristics in database design.....	29
2.4. WordNet.....	31
2.5. Intelligent tutoring systems (ITSs).....	32
2.5.1. Components of an ITS.....	33
2.6. ITS for database design.....	37
2.6.1 DB-Tutor.....	37
2.6.2 Canavan (1996).....	38

	iii
2.6.3 SQL-Tutor.....	38
2.6.4 COLER.....	40
2.6.5 Kermit.....	42
2.7. Summary.....	43
3. Natural language processing in database design.....	44
3.1. Brief overview of database systems analysis.....	44
3.2. Natural language processing in database design.....	46
3.2.1. Ambiguities in natural language specifications.....	49
3.2.2. Solutions to the problem of ambiguity in natural language requirements' specifications.....	52
3.3. Memory-based Shallow Parser (MBSP).....	53
3.3.1. Tokenizer.....	54
3.3.2. Memory-based tagger.....	54
3.3.3. Chunker.....	57
3.4. Summary.....	58
4. Heuristics in database design.....	59
4.1. Existing heuristics.....	59
4.1.1. Heuristics to determine entity types.....	60
4.1.2. Heuristics to determine attribute types.....	61
4.1.3. Heuristics to determine relationship types.....	65
4.1.4. Heuristics to determine cardinality types.....	66
4.2. Proposed new heuristics.....	66
4.2.1. Heuristics to determine entity types.....	66
4.2.2. Heuristics to determine non-entity types.....	68
4.2.3. Heuristics to determine attribute types.....	69
4.2.4. Heuristics to determine relationship types.....	70
4.2.5. Heuristics to determine cardinality types.....	71
4.3. Development of the heuristics.....	73
4.4. Heuristic weights.....	74
4.5. Results on training datasets.....	77
4.6. Justification on heuristics' selection.....	79
4.7. Summary.....	81

5. ER-Converter.....	83
5.1. The Entity-relationship Converter tool.....	83
5.1.1. Part of speech tagging using Memory-Based Shallow Parser.....	85
5.1.2. Identification of attributes and entities from tagged text file.....	86
5.1.3. Human intervention.....	91
5.1.4. Attachment of attributes to their corresponding entities.....	92
5.1.5. Attachment of entities to their corresponding relationships.....	94
5.1.6. Attachment of entities to their corresponding cardinality.....	94
5.1.7. Output final result.....	95
5.2. Summary.....	96
6. Experimental results.....	97
6.1. Evaluation types.....	97
6.1.1. Adequacy evaluation.....	97
6.1.2. Diagnostic evaluation.....	97
6.1.3. Performance evaluation.....	98
6.2. Evaluation metrics.....	98
6.2.1. Criterion.....	99
6.2.2. Measure.....	99
6.2.3. Method.....	103
6.3. Evaluation results.....	104
6.3.1. Overall result.....	104
6.3.2. Contribution of individual heuristics.....	110
6.3.3. Weight applications.....	111
6.3.4. Rejected heuristics' results.....	113
6.3.5. Problems identified as result of evaluation.....	114
6.4. Summary.....	116
7. Conclusions and future work.....	118
7.1. Comparison with related work.....	120
7.2. Future work.....	124
7.2.1. Semantic analysis.....	125
7.2.2. WordNet.....	127
7.2.3. Heuristics' weights	129
7.2.4. Part of a domain model in an ITS.....	130

7.3. Summary.....	130
A. Questionnaire.....	131
B. Penn Treebank II tagset.....	136
C. Training dataset.....	137
C.1.Articles.....	138
C.2.Building.....	138
C.3.Department_project.....	138
C.4.Document.....	138
C.5.Hospital.....	139
C.6.Instructor_course.....	139
C.7.Supplier.....	139
C.8.Training course.....	140
C.9.Vehicle_driver.....	140
C.10.Vehicle_registration.....	140
C.11.Worked example of Department_project.....	141
D. Test dataset.....	146
D.1. Airplane.....	147
D.2. Bank.....	149
D.3. Boat hire.....	152
D.4. Bus.....	154
D.5. Cars.....	156
D.6. Client.....	159
D.7.Company.....	161
D.8. Computer.....	164
D.9. Doctor.....	166
D.10.Dreamhome.....	168
D.11.Electronic supplier.....	170
D.12.Employee.....	173
D.13.Fault.....	175
D.14.Hospital.....	178
D.15.Invoice.....	180
D.16.Library.....	182
D.17.Library books.....	185

	vi
D.18.Machine.....	188
D.19.Musician.....	190
D.20.Order.....	192
D.21.Painter.....	195
D.22.Photograph.....	197
D.23.Professor.....	200
D.24.Project.....	203
D.25.Reliable Rentals.....	206
D.26.Sales representatives.....	208
D.27.Student hall.....	210
D.28.Student.....	212
D.29. Travel.....	214
D.30.University database.....	217
E. List of heuristics.....	219
F. REFERENCES.....	223

LIST OF FIGURES

Figure 1.1 Example of an ER diagram using the Chen notation (Chen, 1976).....	3
Figure 1.2 Representation of subtypes and supertypes.....	6
Figure 2.1 An example of an S-diagram (Eick and Lockemann, 1985).....	13
Figure 2.2 An example of the logical form (Tseng et al., 1992).....	16
Figure 2.3 FORSEN Approach (Meziane and Vadera, 2004).....	19
Figure 2.4 Example of a COLOR-X Event Model (Burg and van de Riet, 1995)....	22
Figure 2.5 The E-R Generator approach (Gomez et al., 1999).....	24
Figure 2.6 The Components of an ITS.....	33
Figure 2.7 Outline of the initial prototype structure (adapted from Canavan, 1996).....	39
Figure 2.8 The architecture of SQL-Tutor (Mitrovic and Ohlsson, 1999).....	40
Figure 3.1 The stages in database systems analysis.....	45
Figure 3.2 Parse tree for the sentence “The man likes the car”.....	47
Figure 4.1 Abbreviations used for categories of heuristics.....	60
Figure 4.2 Processes involved in the development of heuristics.....	75
Figure 5.1 Architecture of ER-Converter tool.....	84
Figure 5.2 An Entity-Relationship model of the ‘Company’ scenario.....	85
Figure 5.3 Extract from the algorithm for heuristics to determine entities and attributes.....	87
Figure 5.4 Extract from algorithm for heuristics to determine relationships.....	88
Figure 5.5 Extract from algorithm to determine cardinalities.....	89
Figure 5.6 Record structure to store words’ information.....	90
Figure 5.7 Extract of output from ‘Company’ scenario’.....	96
Figure 6.1 Venn Diagram to illustrate evaluation measures.....	102
Figure 7.1 Output from WordNet for the Hypernym search category of the word ‘Employee’.....	129

LIST OF TABLES

Table 1.1 Difficulty of Database subject areas.....	7
Table 2.1 Correspondence between English structure and ERM constructs (Chen, 1998).....	12
Table 2.2 General form of CPL specification language.....	22
Table 2.3 Binary rule cases (Gomez et al., 1999).....	26
Table 2.4 Systems that apply NLP to database design.....	28
Table 2.5 Function of COLER's coach sub-modules.....	41
Table 3.1 Accuracy of different taggers.....	56
Table 4.1 Heuristics's weights.....	76
Table 4.3 Training dataset results.....	79
Table 4.3 Contribution of new and old heuristics.....	79
Table 4.4 Frequency count of heuristics applied in the training sets.....	81
Table 5.1 Example using MBSP.....	86
Table 5.2 Relationships between heuristics for determining attribute attachment...	92
Table 6.1 Definition of each of the sets in Venn Diagram (Figure 6.1).....	103
Table 6.2 Results from ER-Converter applied to test dataset.....	105
Table 6.3 Evaluation results.....	106
Table 6.4 Unattached and wrongly attached results.....	109
Table 6.5 Frequency of heuristics applied correctly and incorrectly.....	111
Table 6.6 Rejected heuristics' frequencies in test dataset.....	114
Table 7.1 Comparison of results with related work.....	124
Table 7.2 Semantic roles and their definitions (Jurafsky and Martin, 2000).....	125
Table 7.3 History list of the first sentence in Employee.....	126

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

"I hereby declare that with effect from the date on which the thesis is deposited in the Library of the University of Ulster, I permit the Librarian of the University to allow the thesis to be copied in whole or in part without reference to me on the understanding that such authority applies to the provision of single copies made for study purposes or for inclusion within the stock of another library. *This restriction does not apply to the British Library Thesis Service (which is permitted to copy the thesis on demand for loan or sale under the terms of a separate agreement) nor to the copying or publication of the title and abstract of the thesis.* IT IS A CONDITION OF USE OF THIS THESIS THAT ANYONE WHO CONSULTS IT MUST RECOGNISE THAT THE COPYRIGHT RESTS WITH THE AUTHOR AND THAT NO QUOTATION FROM THE THESIS AND NO INFORMATION DERIVED FROM IT MAY BE PUBLISHED UNLESS THE SOURCE IS PROPERLY ACKNOWLEDGED".