中国学生英语作文自动评分模型的构建

Constructing a Model for
the Computer-Assisted Scoring of
Chinese EFL Learners' Argumentative Essays

梁茂成 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING
AND RESEARCH PRESS

中国学生英语作文自动评分模型的构建

Constructing a Model for
the Computer-Assisted Scoring of
Chinese EFL Learners' Argumentative Essays

梁茂成 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

北京
BEIJING

图书在版编目（CIP）数据

中国学生英语作文自动评分模型的构建＝Constructing a Model for the Computer-Assisted Scoring of Chinese EFL Learners' Argumentative Essays：英文/梁茂成著.— 北京：外语教学与研究出版社，2011.1

ISBN 978-7-5135-0499-7

Ⅰ. ①中… Ⅱ. ①梁… Ⅲ. ①英语—写作—教学研究—英文 Ⅳ. ①H315

中国版本图书馆 CIP 数据核字（2011）第002605号

出 版 人：于春迟
责任编辑：郑建萍
封面设计：覃一彪
版式设计：涂 俐
出版发行：外语教学与研究出版社
社　　址：北京市西三环北路19号（100089）
网　　址：http://www.fltrp.com
印　　刷：北京传奇佳彩数码印刷有限公司
开　　本：650×980 1/16
印　　张：19.75
版　　次：2011年1月第1版 2011年4月第2次印刷
书　　号：ISBN 978-7-5135-0499-7
定　　价：69.90元

* * *

购书咨询：(010)88819929 电子邮箱：club@fltrp.com
如有印刷、装订质量问题，请与出版社联系
联系电话：(010)61207896 电子邮箱：zhijian@fltrp.com

物料号：204990001

Preface

This study reports an attempt to construct a statistical model for the computer-assisted scoring of Chinese EFL writers' essays as well as to explore the extent to which the model can predict EFL writers' argumentative essay scores with extractable essay features.

The data used in the study were 220 timed essays written by English-major students at Nanjing University across four years. The collected written compositions were first random-sampled into two sets, respectively the training set (120 essays) and the validation set (100 essays). Three rating experts were recruited to rate the compositions using an analytic rating scale, with sub-scores for three major aspects of writing quality: language, content, and organization.

Data analysis consisted of a model-training stage and a model-validation stage. In the training stage, techniques in Natural Language Processing, Information Retrieval, and Corpus Linguistics were employed to extract a number of features extracted from the training essays. These features were then correlated with human-assigned essay scores so as to identify predictors of EFL writing quality. The 15 predictors thus identified were then taken as independent variables and human-assigned essay scores were taken as the dependent variable. The multiple regression analysis performed resulted in a 13-predictor model.

In the validation stage, the model constructed during the training stage was employed to predict the scores for the validation essays. In addition, double cross-validation was also conducted so that computer

scores were also assigned for the training essays using a model constructed on the validation essays. In both cases, computer-predicted scores were compared with human-assigned scores in order to test the reliability of computer scores.

Findings of the study indicate that the 13-predictor model has strong predicting power over human-assigned essay scores. The model has a multiple R of 0.837, accounting for 70% of the variation in the dependent variable. Reliability analysis shows that computer-human correlation was r= 0.739, considerably higher than human-human correlation (r = 0.675). Besides, on the six-point scale, the computer-human percent exact agreement was 59.67% while the human-human percent exact agreement was 55.33%; the computer-human percent exact-plus-adjacent agreement was 99.33% while the human-human percent exact-plus-adjacent agreement was 98.89%.

In sum, scores generated by the model proposed in this study are as reliable as or even more reliable than human-assigned scores. The model is comparable to the existing computer-assisted essay scoring systems abroad, which have been programmed to score essays written by native speakers of English.

Acknowledgements

Even though a book has only one author, a book is seldom the work of just one individual. This work was influenced by countless other individuals whom I was fortunate enough to meet during the process of writing. While space does not permit me to acknowledge them all, I would be remiss if I did not acknowledge the following individuals whose guidance, support, and wisdom so greatly influenced the whole body of this work.

First and foremost, this book would not have started at all if it weren't for the unique insights of my supervisor, Professor Wen Qiufang, who probably had been wondering for quite some time whether computer-assisted assessment of EFL writing is possible before she gave me her nudge. She was the catalyst for this project. During the process of my work, she has always been an unwavering source of support and a genuine inspiration, and she cheerfully provided solutions to many problems that had me in a panic. In addition to offering me a helping hand, perhaps more importantly, she believed in me, even when I didn't believe in myself. Another source of my satisfaction was Professor Ting Yenren, my co-supervisor and a recognized expert of EFL writing assessment at Nanjing University and beyond. Since the essence of computer-assisted assessment of writing is to simulate human rating, whenever I was wondering whether a certain factor contributes to human-assigned scores, I would simply resort to Professor Ting, who would most unreservedly tell me what he was affected by as a rater. Not surprisingly, in many cases, my findings

would simply prove his intuitions. I later came to see that the nature of my project is to find out to what extent the computer can simulate Professor Ting as an essay rater.

I also owe a lot of thanks to Professor Gui Shichun, Professor He Anping, Professor Yan Chensong, and Professor Wang Chuming, who, when I reported to them what I was working on, kindly offered their advice and encouragement.

My thanks also go to Professor Wang Haixiao, Professor Wang Lifei and Professor Ma Guanghui, who would unreservedly offer me whatever assistance they could offer. I would also like to acknowledge Dr. Miao Zhengke, who, when I was on the brink of giving up, generously gave me the much-needed technical assistance in mathematics. Part of the wisdom needed to complete this book came from Professor Yu Hongliang, who often treated me with beer and cigarettes, which in turn brought about the birth of many of my ideas!

I was able to save a lot of work as I chose the right people to help with the rating. These include Professors Wen Qiufang, Yu Hongliang, Wang Yu, Dr. Bao Gui and Dr. Heng Renquan. Their work has become part of the model proposed in this book, and will be put into operation the time a software program is written on the basis of the model. If the model does the right thing, it is because it was trained by the right raters.

I also need to acknowledge my fellow researchers and classmates at the Ph.D Program in Applied Linguistics at Nanjing University. These, most notably, include Jennifer Chen, Zhou Dandan, Wang Yu and Xu Hongliang, with whom I would often share ideas, miseries, happiness, and meals at the restaurants around the corner. The mutual encouragement we received from each other provided me with renewed energies. The completion of this book is partly attributable to their support and encouragement.

Most of all, I must acknowledge my wife, Xu Wenling. She suffered through missed deadlines, rotten moods (mine, not hers), and many long hours of being apart, and still managed to love and support me in this endeavor. I would not have completed this book without her. She took nearly all the responsibilities of looking after our son, Liang Yuan, who would often push me forward with his academic success at school.

In closing, I would also like to acknowledge the following for their help in preparing this book: Kang Shi Fu (instant noodles), Nescafé (instant coffee), and the dumplings sold outside the University gate at midnight.

Contents

List of Abbreviations xiii

List of Tables xv

List of Figures xix

Part One Introduction 1

Introducing the Study 2

0.1 Introductory remarks 2

0.2 Need for this study 3

0.2.1 Theoretical considerations 3

0.2.2 Practical considerations 7

0.3 Description of the study 10

0.4 Organization of the study 11

0.5 Summary 12

Part Two Literature Review 13

Chapter 1 A Review of Existing Computer-Assisted Essay Scoring Systems 14

1.1 Introduction 14

1.2 Key concepts 14

1.2.1 Computer-assisted essay scoring 14

1.2.2 EFL writing assessment 16

1.3 Existing computer-assisted essay scoring systems 17

1.3.1 Project Essay Grade (PEG):

A form-focused system 17

1.3.2 Intelligent Essay Assessor (IEA):
A content-focused system 20
1.3.3 E-rater: A hybrid system with a modular structure..... 22
1.3.4 An appraisal of the three existing systems 25
1.4 Lessons from existing essay scoring systems..... 28
1.5 Summary 31

Chapter 2 Studies on Measures of Writing Quality 33
2.1 Introduction 33
2.2 Measures of writing quality in the literature 33
2.2.1 Measures of the quality of language..... 34
2.2.2 Measures of the quality of content and
organization 51
2.3 An overview of the measures in the literature..... 57
2.4 A conceptual model for the computer-assisted scoring
of EFL essays..... 61
2.5 Proposed measures of EFL writing quality 65
2.5.1 Proposed measures of the quality of language
in EFL writing 65
2.5.2 Proposed measures of the quality of content
in EFL writing 69
2.5.3 Proposed measures of the quality of organization
in EFL writing 71
2.6 Summary 75

Part Three Methodology 77

Chapter 3 Research Questions and Data Preparation 78
3.1 Introduction 78
3.2 Research questions 78

3.3 The corpus 80
3.4 The rating scheme 82
3.4.1 Selecting a rating scale 82
3.4.2 The revised rating scale 84
3.4.3 The evaluation of content 87
3.4.4 The weighting scheme 90
3.5 Rating 91
3.5.1 Rater selection 92
3.5.2 Rater training 92
3.5.3 The rating sessions 93
3.6 Score reliability 94
3.7 Summary 96

Chapter 4 Text Analysis and Statistical Analysis 97
4.1 Introduction 97
4.2 Tools 97
4.3 Essay feature extraction 99
4.3.1 Language features 100
4.3.2 Content features 103
4.3.3 Organizational features 110
4.4 Data analysis 111
4.4.1 Correlation analysis 111
4.4.2 Multiple regression analysis 112
4.4.3 Stages of data analysis 113
4.5 Summary 117

Part Four Results and Discussion 119

Chapter 5 Identifying Predictors of EFL Writing Quality 120
5.1 Introduction 120

5.2 Linguistic features and writing quality 120
5.2.1 Fluency and writing quality 123
5.2.2 Complexity of language and writing quality 126
5.2.3 Measures of linguistic idiomaticity and appropriateness 138
5.3 Results of content analysis 144
5.3.1 Results of Latent Semantic Analysis 145
5.3.2 Procedural vocabulary and essay score 149
5.4 Essay organization and writing quality 151
5.4.1 Paragraphing and writing quality 152
5.4.2 Discourse conjuncts and writing quality 159
5.4.3 Demonstratives, pronouns, connective and writing quality 159
5.5 Power of the predictors proposed in this study 159
5.6 Summary 161

Chapter 6 A Statistical Model for Computer-Assisted Essay Scoring 164

6.1 Introduction 164
6.2 Diagnosing the preliminary model 165
6.3 The refined model 168
6.4 Predictors and aspects of writing quality measured 172
6.4.1 Predictors in the language module 173
6.4.2 Predictors in the content module 178
6.4.3 Predictors in the organization module 181
6.4.4 Interdependence of the modules 183
6.5 Implementing the model 185
6.6 Summary 187

Chapter 7 Validating the Model 188
7.1 Introduction 188
7.2 Cross-validating the model 188
7.3 Reliability of computer scores in cross-validation 191
7.3.1 Aspects of reliability 191
7.3.2 Consistency estimates 193
7.3.3 Consensus estimates 195
7.4 Double cross-validation 198
7.4.1 Constructing the model 198
7.4.2 Model statistics and estimated equation 199
7.5 Reliability of computer scores in double cross-validation 201
7.6 Comparison with existing essay scoring systems 204
7.6.1 Comparison with PEG 205
7.6.2 Comparison with IEA 208
7.6.3 Comparison with E-rater 212
7.7 Summary 214

Part Five Conclusion 215

Chapter 8 Conclusion 216
8.1 Major findings 216
8.1.1 A model for the computer-assisted scoring of EFL essays 216
8.1.2 Predictors of EFL writing quality 220
8.2 Limitations of the study 223
8.2 Future work 224

References 226

Appendices 249

Appendix I PEG's proxes and their beta values (Page 1968) 249
Appendix II Page's (1995) model and variables 251
Appendix III Argument weight 253
Appendix IV Examples of good openings and endings 255
Appendix V Scoring table (Organization & Content) 256
Appendix VI Scoring table (Language) 257
Appendix VII List of stopwords 258
Appendix VIII Lemma list (excerpt) 262
Appendix IX List of content words 266
Appendix X Sample essays 283
Appendix XI POS-tagged samples 286

List of Abbreviations

ASL	Average Sentence Length
AWA	Analytical Writing Assessment
AWL	Average Word Length
BNC	British National Corpus
CET	College English Test
CIA	Contrastive Interlanguage Analysis
CL	Corpus Linguistics
CR	Contrastive Rhetoric
DC	Discourse Conjuncts
DefArt	Use of the Definite Article
EFL	English as a Foreign Language
EL	Essay Length
EL4	Fourth Root of Essay Length
E-rater	Electronic Essay Rater
ETS	Educational Testing Service
GerV	Number of Gerundial Verbs
GMAT	Graduate Management Admissions Test
GRE	Graduate Record Examination
IEA	Intelligent Essay Assessor
IL	Interlanguage
IR	Information Retrieval
L1	First Language
L2	Second Language

LSA	Latent Semantic Analysis
NAEP	National Assessment of Education Progress
NLP	Natural Language Processing
OLS	Ordinary Least Squares
Parag	Paragraphing
PEG	Project Essay Grade
PRAXIS	Professional Assessments for Beginning Teachers
Preps	Use of Prepositions
PV	Procedural Vocabulary
RWC	Recurrent Word Combination
SD	Standard Deviation
SDWL	Standard Deviation of Word Length
SLA	Second Language Acquisition
SVD	Singular Value Decomposition
TOEFL	Test of English as a Foreign Language
TTR	Type-Token Ratio
TWE	Test of Written English
VFP	Vocabulary Frequency Profile
VSM	Vector Space Model

List of Tables

Chapter 1

Table 1.1 Comparison of strengths and weaknesses of existing essay scoring systems 26

Table 1.2 Approaches and measured constructs 28

Chapter 2

Table 2.1 Measures of writing quality in previous studies 58

Chapter 3

Table 3.1 Comparison of holistic and analytic scales (from Weigle 2002) 83

Table 3.2 Jacobs et al.'s (1981) scale: Aspects of quality and their emphasis 85

Table 3.3 Modified scheme: Aspects of writing quality 86

Table 3.4 Aspects of writing quality and their emphasis in the revised scale 91

Table 3.5 Inter-rater correlations (Training set) 95

Table 3.6 Mean and standard deviation of scores (Training set) 95

Table 3.7 Inter-rater correlations (Validation set) 95

Table 3.8 Mean and standard deviation of scores (Validation set) 95

Chapter 4

Table 4.1 Directly extracted language features 100

Table 4.2 Computed language features 100