An Information Theoretic Approach to Language Relatedness

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Information Systems
at Massey University.

Anand Venkt Raman
1997

This documentation was prepared from a camera ready copy made by the author using LaTeX. The International Phonetic Alphabet characters were produced using the TeX font (wsuipa) created by Washington State University. The graphs were produced using `gnuplot`. PFSA diagrams were produced using a modified version of Jos Van Einjdhoven's `graphplace` program. Other figures were produced using the `xfig` program. The final Postscript file was generated using `dvips`, printed on a Hewlett-Packard Laserjet printer and photoreproduced and bound by Wills Bookbinding and Printing, 6 Dahlia Street, Palmerston North.

The software used to produce the results for this work are available through anonymous ftp from the URL ftp://fims-ftp.massey.ac.nz/pub/ARaman. Bug reports may be sent to A.Raman@massey.ac.nz. Work done using this software must cite either Raman and Patrick (1997d) or this dissertation.

No Microsoft products were used in the preparation of this work.

Abstract

This dissertation examines the prospect of applying information theoretic principles to help solve problems in historical linguistics. The Minimum Message Length principle attributed to Chris Wallace (similar to the Minimum Description Length principle of Jorma Rissanen) is used to judge the goodness of hypotheses in the field of historical linguistics. The idea is that theories that require a shorter message to describe with their data are better than those that require long messages.

Work in collecting the linguistic data tracing the derivation of some 2714 words in Modern Cantonese and Modern Beijing from their forms in a reconstruction of Middle Chinese is described as also is the work in transforming this data into a format suitable for use with software developed for this project.

Heuristics for inferring Probabilistic Finite State Automata (PFSA[1]) from such data are reviewed and some new heuristics are introduced. These are then applied to training data and benchmark results presented.

Finally, the inference process is applied to the actual linguistic data which allows a conjecture regarding a relative closeness of the Chinese dialects to their reconstructed ancestor to be formed.

[1] In this dissertation, the abbreviation PFSA has been used to denote both the singular and plural of these machines, the "A" in PFSA being understood to represent both *Automaton* and *Automata*.

Contents

List of Tables

List of Figures

Preface

In late 1993, Professor Jon Patrick, who was then just about to leave Deakin University in Melbourne for Massey gave a seminar here on some computational work he was doing in Basque linguistics. I had at that time just joined the Computer Science department as an assistant lecturer and was avidly looking for a PhD project to begin. Like almost everybody else I have met, I also had a compelling interest in historical linguistics and had my own opinion of how it should be done. I sent a brief CV to Jon asking if he would like to take on a doctoral student in his Basque project after his arrival and the answer soon came back in the affirmative. In the beginning, though, our intention was quite different, and much more grandiose and ambitious — we were hoping to assess whether or not the linguistic isolate Basque was related to the Dravidian language Tamil, which by the way, I happen to be very fluent in.

As time went on, however, the focus of the project changed rather dramatically and eventually became much more realistic. It narrowed down into developing a technique for finding distance measures between related natural languages to aid linguists in their task of subgrouping. We decided, wisely perhaps, that bold proposals in the linguistic domain are best made by linguists. We also decided, with the advice of my second supervisor, Dr John Newman in the Department of Linguistics and Second Language Teaching, to use Chinese data for the project. This data had been provided to John courtesy of Professor William S-Y. Wang at the University of California, Berkeley.

The field of computational linguistics is such that almost all Computer Scientists seem to have their own unique notions of how one should apply computational methods to historical data. While there may be merits and demerits to each such approach, this dissertation is not an evaluation of their relative goodness. It is a detailed account of one particular approach — the one described here — and how it can be used to provide linguists with a tool for effective subgrouping of languages.

Overview of contents

Chapter 1 gives a brief introduction to the problem and describes some linguistic terms and concepts for the benefit of the reader unacquainted with them. Chapter 2 is a brief survey of previous work in the field of linguistics which has been pursued along similar lines. An example of the methodology applied to a toy problem is then provided in Chapter 3. Chapter 4 describes the procedure by which the data for this project was collected and the motivation for using it. The MML criterion, which is central to this project, is introduced and described in Chapter 5, which is followed by Chapters 6–8 where methods to infer structure from the data using the MML criterion are looked at. Finally, Chapter 9 gives the results obtained, discusses them and describes the prospects for future work in this area.

Publications

Parts of this dissertation have been published in various places during the course of working on this project. The following is a list of them, with brief notes on which chapters they refer to.

Raman, A. V. and J. D. Patrick (1997). Linguistic similarity measures using the Minimum Message Length principle. In R. D. Blench and M. Spriggs (Eds.), *Archaeology and language I: Theoretical and methodological orientations*, pp.260–277, London: Routledge. This paper was also read at the WAC-3 Conference, New Delhi, December 1994. Material from this paper can be found in Chapters 2 and 3.

Raman, A. V. and J. D. Patrick (1997). The sk-strings method for inferring PFSA. In *Proceedings of the workshop on automata induction, grammatical inference and language acquisition at the 14th international conference on machine learning — ICML-97*, Nashville, Tennessee, (in press). Material from this paper can be found in Chapter 6.

Raman, A. V. and J. D. Patrick (1997). A heuristic using cross-entropy for sk-strings In *Technical Report TR 1/97*, Information Systems Department, Massey University. Material from this report can be found in Chapter 7.

Raman, A. V. and J. D. Patrick (1997). Beam search and simba search for PFSA inference, In *Technical Report TR 2/97*, Information Systems Department, Massey University. Material from this report can be found in Chapter 8.

Raman, A. V., J. Newman and J. D. Patrick (1997). A complexity measure for diachronic Chinese phonology. In *Proceedings of the SIGPHON97 workshop on computational phonology*, Madrid (in press). Material from this paper can be found in Chapter 9.

An explanation for some typesetting decisions

I feel compelled to say a few words about the typesetting process this dissertation has been through. I initially started writing the thesis using Microsoft Word on my Apple Macintosh, but found its performance unsatisfactory. Moreover, all my programs to manipulate the data were being run under Unix. What I wanted was a single command that could automatically take the ASCII outputs from my various experiments, turn them into tables, graphs and pictures and insert them in the right places in the thesis and also generate my bibliography neatly.

I thus deemed it best to migrate the work to LaTeX and embarked on this bold venture. Not one bit of this effort was wasted as LaTeX, and more particularly the Unix environment have paid it back several times and over. The amazing variety of utilities and high-quality programs available freely under Unix enabled me to get precisely what I wanted. Unix scripts using `awk`, `sed` and `grep` did most of the initial formatting of my program's ASCII outputs. All the formatted tables were generated automatically by scripts that ran the experiments themselves. So were the various graphs which were produced using the excellent `gnuplot` program. The various files which made up the chapters were edited under TeX mode in GNU `emacs`. `Emacs` macros also helped immensely in migrating my bibliography database from `EndNote` on the Macintosh into BibTeX. Finally, I was able to put the various commands into a `Makefile` and my dream of generating the entire thesis with a single command was realised. "`Make dvi`" generated the dvi version of my thesis, which I then printed off using `dvips`. The dvi and postscript versions of the thesis are available from me on request. My thanks to Donald Knuth for TeX, Leslie Lamport for LaTeX, Oren Patashnik for BibTeX and to the authors of the various free software packages, especially GNU `emacs` and family, which I have found to be of utmost use in typesetting this work.

BibTEX explicitly recommends against the use of the Chicago style A for bibliographies.[2] Since I did use that style in the end, I feel a need to justify my decision. When I read a paper, I like to get an idea of when a certain result being cited was achieved. The plain style doesn't allow for this easily, as it is annoying having to flip to the reference section each time you have passed a citation. Also, I am told that although real scientists don't admit to it, most of us do unofficially build up knowledge about the reputation of authors and would like to know if a result being cited is attributed to a reputable author or not. Again, the plain style doesn't allow for this easily. Perhaps it is alright for a short paper as there aren't too many references to wade through, but in a dissertation such as this, I felt Chicago A would be most appropriate.

Acknowledgments

From the time this project began in 1994, several people have had signal input to it. It would be futile task to list them all, but I try to do a modest job of it here.

First and foremost among those I should thank are Jon Patrick and John Newman, my two supervisors. Jon's infectious enthusiasm for MML and comparative linguistics from which this project originated has been the source of much inspiration for many of the results achieved. In many ways, Jon was my ideal supervisor. His constantly supportive role played a more than important part in this project and his eagerness to discuss and try out new ideas was extremely refreshing during times of burn-out. Jon was always willing to listen, not only to my research problems but also to all my ideas. I also thank him for financial assistance he has provided in travelling to conferences so as to present this work.

I would, of course, have found it impossible to do much at all in this project if not for John Newman, under whose watchful eye I learnt all the linguistics necessary[3] and finally how to achieve consistency in the presentation of this dissertation. I owe him tremendously for his countless pieces of invaluable advice and for introducing me to the scientific method in linguistics. Thanks also for his timely injections of encouragement which supplied me with much needed fuel to march on with this work.

In mid-1996, I had the opportunity to spend one week at Monash University, Melbourne, discussing various aspects of this project with researchers in the Computer Science Department there. Most importantly, it gave me a chance to meet and discuss with Chris Wallace, who in 1968 had originated the MML principle that is fundamental to this work. However, things didn't go as planned — I was down with a severe cold just that week and Chris was already semi-retired from the University at that stage. I only got to meet him twice, and that too only briefly. Nevertheless, these two brief meetings made all the difference. Chris had given me several important insights[4] into the use of MML and I had given him my cold. Chris retired from Monash as this dissertation was being written up. I hope this is fitting as a tribute to a man who laid the foundations for what is and will continue be a prolific research area in future. The kindness of the Monash CS Department who made available to me resources during that week is also gratefully acknowledged.

Other people who have also had significant input to various parts of this work by way of stimulating discussions are John Hudson and Peter Kay, Computer Science Department Massey University, Peter Andreae, Computer Science Department, Victoria University, Ian Witten and Craig Neville-Manning, Computer Science Department,

[2] See BibTEXing, by Oren Patashnik, p.11

[3] This included auditing his phonology lectures in 1994.

[4] The decision to use a prior and directly specify the number of outgoing arcs from a state of the PFSA in Chapter 5 is due to a discussion I had with him on the 26th August.

Waikato University and Lloyd Allison, Rohan Baxter, Jon Oliver and Matthew Collins, Computer Science Department, Monash University. Peter Kay also proof-read pre-final drafts of this thesis and made many useful suggestions to improve its content. I am certain that any errors you may discover in this thesis belong to the few paragraphs that have been added since his careful eyes had scoured the thesis.

The work has also benefited greatly from several reviewers, some anonymous, who read interim reports and portions of the work presented at conferences. Among them, special thanks go to Dr Roger Blench, Cambridge University, Dr Peter Christian, Goldsmith's college, London, and Dr Sheila Embleton, York University, Toronto.

Dr Rosemary Haddon, Department of East Asian Languages, Massey University gave me a primer course in Mandarin and I am grateful to her for letting me audit her lectures in 1995.

Dr Siva Ganesh, Statistics Department, Massey University spent much effort in introducing me to `SAS` for graphing some of my results. I thank him for it, although I decided in the end to use `gnuplot`. Likewise, thanks also to the software support staff of the Faculty, who persevered in helping me with several Microsoft programs, although I decided in the end to not use any of them. I did, however, find much use for Philip Etheridge's timely and prompt assistance with TEX and LATEX and the many new packages he had to install on the system for me.

Ben Patrick and Kirsty How undertook the painstaking jobs of entering into a spreadsheet the derivations for each of the 2714 words in Modern Beijing and Modern Cantonese. I doubt if I could have done a better job in less time. My thanks go to both of them.

The Computer Science Department at Massey, especially Dr Chris Phillips, deserve a big note of thanks for arranging things so that I would have a reduced teaching load during the time I wrote up. Also, thanks to Professor Mark Apperley, now at Waikato University, who encouraged me to apply for an academic position in 1993 and start work on my PhD.

Finally I must record here my heartfelt thanks to my lovely wife Mathangi and beautiful son Panini[5] who made it possible to persevere in the face of apparently blank walls in my research at times.

[5] http://fims-www.massey.ac.nz/~ARaman/images/pdgr.jpg

A Koan about Prior Knowledge[6]

In the days when Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6.

"What are you doing?", asked Minsky.

"I am training a randomly wired neural net to play Tic-Tac-Toe" Sussman replied.

"Why is the net wired randomly?", asked Minsky.

"I do not want it to have any preconceptions of how to play", Sussman said.

Minsky then shut his eyes.

"Why do you close your eyes?", Sussman asked his teacher.

"So that the room will be empty."

At that moment, Sussman was enlightened.

[6]Taken from The New Hackers Dictionary, 2nd ed., Compiled by Eric S Raymond, MIT Press, Cambridge MA, 1994, p.475