PRONUNCIATION MODELLING AND BOOTSTRAPPING

By

Marelie Hattingh Davel

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor (Electronic Engineering)

in the

Faculty of Engineering, the Built Environment and Information Technology

at the

UNIVERSITY OF PRETORIA

Advisor: Professor E. Barnard

August 2005

PRONUNCIATION MODELLING AND BOOTSTRAPPING

Bootstrapping techniques have the potential to accelerate the development of language technology resources. This is of specific importance in the developing world where language technology resources are scarce and linguistic diversity is high. In this thesis we analyse the pronunciation modelling task within a bootstrapping framework, as a case study in the bootstrapping of language technology resources.

We analyse the grapheme-to-phoneme conversion task in the search for a grapheme-to-phoneme conversion algorithm that can be utilised during bootstrapping. We experiment with enhancements to the Dynamically Expanding Context algorithm and develop a new algorithm for grapheme-to-phoneme rule extraction (*Default&Refine*) that utilises the concept of a 'default phoneme' to create a cascade of increasingly specialised rules. This algorithm displays a number of attractive properties including rapid learning, language independence, good asymptotic accuracy, robustness to noise, and the production of a compact rule set. In order to have greater flexibility with regard to the various heuristic choices made during rewrite rule extraction, we define a new theoretical framework for analysing instance-based learning of rewrite rule sets. We define the concept of *minimal representation graphs*, and discuss the utility of these graphs in obtaining the smallest possible rule set describing a given set of discrete training data.

We develop an approach for the interactive creation of pronunciation models via bootstrapping, and implement this approach in a system that integrates various of the analysed grapheme-to-phoneme alignment and conversion algorithms. The focus of this work is on combining machine learning and human intervention in such a way as to minimise the amount of human effort required during bootstrapping, and a generic framework for the analysis of this process is defined. Practical tools that support the bootstrapping process are developed and the efficiency of the process is analysed from both a machine learning and a human factors perspective. We find that even linguistically untrained users can use the system to create electronic pronunciation dictionaries accurately, in a fraction of the time the traditional approach requires. We create new dictionaries in a number of languages (isiZulu, Afrikaans and Sepedi) and demonstrate the utility of these dictionaries by incorporating them in speech technology systems.

Keywords: bootstrapping, grapheme-to-phoneme conversion, grapheme-to-phoneme alignment, letter-to-sound, pronunciation modelling, pronunciation prediction, pronunciation rules, pronunciation dictionary, language technology resource development.

UITSPRAAKMODELLERING EN SELFSTEUN

Selfsteuntegnieke beloof om die ontwikkeling van taalhulpbronne vir tegnologiese toepassings te versnel. Hierdie belofte is veral belangrik in die onwikkelende wêreld, waar sulke hulpbronne skaars is, en beduidende taalverskeidenheid voorkom. In hierdie tesis ontleed ons die uitspraakvoorspellingstaak binne 'n selfsteunraamwerk, as 'n gevallestudie van selfsteunontwikkeling van taalhulpbronne.

Ons ontleed grafeem-na-foneemomskakeling, op soek na 'n algoritme wat vir selfsteundoeleindes gebruik kan word. Ons ondersoek verbeteringe aan die "Dinamiese Konteksuitbreiding" (DEC) algoritme, en ontwikkel 'n nuwe algoritme vir die onttrekking van grafeem-na-foneemreëls (*Verstek&Verfyn*) wat die begrip van 'n 'verstekfoneem' gebruik om 'n rits van toenemend afgestemde reëls te skep. Hierdie algoritme vertoon 'n aantal aantreklike eienskappe, insluitende kort leertye, taalonafhanklikheid, goeie uitloopakkuraatheid, ruisbestandheid, en die skep van klein reëlstelle. Om groter plooibaarheid in 'n aantal heuristiese keuses te verkry, stel ons 'n nuwe teoretiese raamwerk vir die ontleed van geval-gebasseerde leerprosesse van herskryfreëls voor. Ons stel die begrip van *kleinste voorstellende grafieke* voor, en bespreek die nut van sulke grafieke in die onttrek van die kleinste moontlike reëlstel wat gegewe leervoorbeelde beskryf.

Ons ontwikkel 'n benadering tot die wisselwerkende skep van uitspraakmodelle deur selfsteun, en verwerklik hierdie benadering in 'n stelsel wat verskeie van die ontlede algoritmes vir belyning en reëlonttrekking saamvat. Ons gee aandag aan die saamvoeg van masjienleer en menslike ingrype om die hoeveelheid menslike inset tydens selfsteun so klein moontlik te hou, en ontwikkel 'n algemene raamwerk vir die ontleding van hierdie proses. Verder ontwikkel ons praktiese gereedskap ter ondersteuning van selfsteun, en ontleed die doeltreffendheid daarvan uit die oogpunte van masjienleer en menslike bruikbaarheid. Ons bevind dat selfs gebruikers sonder taalkundige opleiding akkurate woordeboeke sodoende kan skep, in 'n breukdeel van die tyd wat die gebruiklike benadering vereis. Ons skep nuwe woordeboeke vir verskeie tale (isiZulu, Afrikaans en Sepedi), en toon die nuttigheid van hierdie woordeboeke in spraaktegnologietoepassings.

Sleutelterme: selfsteun, grafeem-na-foneem omsetting, grafeem-na-foneem belyning, letter-na-klank, uitspraakmodellering, uitspraakvoorspelling, uitspraakwoordeboek, uitspraakreëls, hulpbronontwikkeling vir taaltegnologie.

ACKNOWLEDGEMENTS

This research was performed in the Human Language Technologies (HLT) Research Group of the Meraka Institute. It was guided by Etienne Barnard, for the past three years my PhD advisor, colleague and the ideal co-explorer (a rare privilege!)

The HLT Research group grew in parallel with this thesis, and I am grateful to all the group members who assisted in one way or another: Louis Joubert, Francois Aucamp and others, who assisted with the development of 'System B'; Aby Louw who assisted in integrating some of the newly created dictionaries in Text-to-Speech systems; and the many other HLT researchers, developers and students who provided a supportive research environment.

Much of the collected data relied on the patience of the various dictionary developers. I am especially grateful to Nadia Barnard, who assisted with both kindness and skill.

I would also like to thank:

- Johan Eksteen, my manager at the CSIR at the time, who initially supported my decision to make the jump from project manager to researcher, and who has supported my work ever since.
- Liesbeth Botha, who guided my initial explorations in speech-related research.
- Rich Stern and the Carnegie Mellon Speech Group, who hosted me at Carnegie Mellon in Pittsburgh for a very enjoyable year.

Finally, I would like to thank my friends and family for putting up with me as a mostly absent, part-time PhD student; and of course, MC, without whose support this whole endeavour would have been quite impossible.

TABLE OF CONTENTS