Learning to pronounce written words : a study in inductive language learning

Summary

Learning to pronounce written words means learning the intricate relations between the speech sounds of a language and its spelling. For languages with alphabetic writing systems, such as English, the relations can be captured to a large extent by *induction* (reasoning by analogy). After all, a pervasive phenomenon in alphabetic writing systems is that similarly-spelled words have similar pronunciations. However, mainstream (Chomskyan) linguistic theories have put forward the claim that pronouncing known and unknown words cannot be performed without the assumption of several levels of abstraction between spelling and pronunciation. Since general-purpose inductive-learning methods cannot discover such abstraction levels autonomously, linguistic theorists claim that inductive-learning methods cannot learn to pronounce words as well as generalise this knowledge to previously unseen words.

The present study challenges this claim. The study is embedded in both (i) the tradition of *structural* linguistics, building quite directly on ideas expressed a century ago by De Saussure, and (ii) the recent developments in machine learning (a subdomain of artificial intelligence). De Saussure claimed that language processing can be performed by assuming only two basic operations: segmentation and classification. In the machine-learning domain, it is claimed, and occasionally demonstrated, that inductive learning methods can learn complex real-world segmentation and classification tasks, attaining a high level of generalisation accuracy when given a sufficient amount of examples. The apparent and intriguing contrast between the claims from mainstream (Chomskyan) linguistics on the one hand, and those of structural linguistics and machine learning on the other hand, prompted us to perform an empirical study of the inductive learning of word pronunciation. The results of this empirical study allow us to claim that *inductive-learning algorithms can learn to pronounce written words with adequate generalisation accuracy, even*

when the task definition (and thus the system architecture) does not reflect explicitly any of the levels of abstraction assumed necessary by linguistic theories.

Chapter 1 introduces the historical background mentioned above on arguments and counterarguments concerning the feasibility of inductive language learning. On the basis of claims from structural linguistics and machine learning, the problem statement is formulated.

Chapter 2 provides the reader with an overview of relevant background knowledge. It describes three groups of inductive-learning algorithms: connectionist learning, instance-based learning, and decision-tree learning. They are suited, in principle, for learning word pronunciation. The chapter reviews mainstream linguistic views on the domains of morphology and phonology, highlighting levels of abstraction assumed present in word pronunciation. It then introduces the resource of word-pronunciation examples used in our study, i.e., the CELEX English lexical data base, and describes the general experimental methodology employed throughout the study.

Chapter 3 presents the application of the selected learning algorithms to five subtasks of the word-phonemisation task: (i) morphological segmentation, (ii) graphemic parsing, (iii) grapheme-phoneme conversion, (iv) syllabification, and (v) stress assignment. They represent five linguistically-motivated abstraction levels of word pronunciation. The results obtained with the five algorithms trained on each of the five subtasks in isolation indicate that the learning algorithms attain reasonable to excellent generalisation accuracy. Moreover, the results indicate that the less a learning algorithm abstracts from the learning material by data compression, the better its generalisation accuracy is on any of the subtasks.

In Chapter 4, modular word-pronunciation systems are constructed, learned, and tested. The architecture of the modular systems is inspired by two existing text-to-speech systems. Both modular systems perform the five subtasks investigated in Chapter 3. Each subtask is assigned to a single module; the five modules perform their subtasks in sequence. Generalisation-accuracy results indicate that cascading errors passed on between the modules seriously impede the overall accuracy of the systems. To prevent some of the errors we abandon the assumption that a five-modular decomposition is necessary, and investigate two three-modular systems, in which two pairs of subtasks from the five-modular systems are integrated into single tasks. The systems distinguish between (i) morphological segmentation, (ii) grapheme-phoneme conversion, and (iii) stress assignment. Their generalisation accuracy is significantly better than that of their five-modular counterparts.

In Chapter 5 the concept of sequential modularisation is abandoned and the alternative of parallel modularisation is tested in three new modular systems. In the first system, word pronunciation is performed by a single module converting spelling to phonemic transcriptions with stress markers in a single classification pass. The second system performs two tasks in parallel, viz. the conversion of letters to phonemes and the conversion of letters to stress markers. In the third system, the letter-phoneme conversion task is split further into 25 partial subtasks: each of these subtasks represents the detection of one articulatory feature of the phoneme to be classified. The results indicate that the single-module and two-module parallel systems perform better, and the articulatory-feature-detection system performs worse than the best three-module sequential system described in Chapter 4, when trained with the same algorithms.

Chapter 6 deals with three linguistically-uninformed gating systems for word pronunciation. In these systems, the word-pronunciation task is split in two parallel-processed partial word-pronunciation tasks. Rather than decomposing the task on the output level, the task is decomposed by applying a gating criterion at the input level, viz. on the spelling of (parts of) words. Three gating systems are tested: randomised gating, typicality-based gating, and occurrence-based gating. Randomised gating is demonstrated to be learned with lower accuracy than the word-pronunciation task as a whole; the typicality-based and occurrence-based systems are found to perform as accurate as the undecomposed system. Thus, gating does not lead to improvements in generalisation accuracy, but is capable of automatically decomposing the word-pronunciation data in essentially different subsets.

A summary of results reported in Chapters 3 to 6 is given in Chapter 7. Additional attention is paid to measures of computational efficiency of the learning algorithms tested. Instance-based learning attains the best (and adequate) generalisation accuracy on all (sub)tasks; the systems induced by decision-tree learning provide the best trade-off between generalisation accuracy and computational efficiency. Furthermore, analyses on the word-pronunciation data are performed searching for the cause of the success of instance-based learning: the analyses indicate that instances of word pronunciation tend to come in families containing small amounts of identically-classified members. Instance-based learning performs favourably with data containing this type of instance families (small disjuncts). Subsequently, the chapter describes two modular systems that combine apparently successful attributes of different systems investigated throughout the study, showing that better systems can

be built on the basis of both linguistic and empirical findings. The chapter then summarises related research, gives an overview of the limitations of the present approach, and indicates topics of future research.

In Chapter 8 the conclusion is drawn that inductive-learning methods, specifically instance-based learning algorithms, can learn the task of word phonemisation, attaining an adequately high level of generalisation accuracy. Linguistic bias in the task definition (and in the system architecture) can be reduced to an absolute minimum, thus be left *implicit*, while the system still attains accurate generalisation.

Samenvatting

Het leren van de uitspraak van woorden is het ontdekken van de ingewikkelde relaties tussen de klanken van een taal en haar spelling. In het geval van talen met een alfabetisch schrijfsysteem, zoals het Engels en het Nederlands, kunnen deze relaties voor een groot gedeelte worden gevonden door middel van *inductie* (redeneren door analogie). Voor alfabetische schrijfsystemen geldt immers dat woorden die in hun spelling op elkaar lijken, ook in hun uitspraak op elkaar lijken. De belangrijkste taalkundige theorieën (die voortbouwen op ideeën van Chomsky) beweren echter dat het uitspreken van bekende en onbekende woorden niet mogelijk is zonder aan te nemen dat er verschillende abstractieniveaus bestaan tussen spelling en uitspraak. Taaltheoretici stellen dat inductief-lerende methoden niet in staat zijn om woorduitspraak te leren en om de geleerde kennis toe te passen op nieuwe, onbekende woorden, omdat inductief-lerende methoden niet in staat zijn om uit zichzelf dergelijke abstractieniveaus te ontdekken.

Dit proefschrift zet een vraagteken bij deze stelling. De studie is ingebed in (i) de traditie van de *structurele* taalkunde, die vrij rechtstreeks voortbouwt op de ideeën van De Saussure van een eeuw geleden, en in recente ontwikkelingen binnen (ii) automatisch leren (*machine learning*), een deelgebied van de kunstmatige intelligentie. De Saussure stelde dat het verwerken van taal mogelijk is onder de aanname van slechts twee operaties: segmentatie en classificatie. Binnen het automatisch leren geldt de opvatting dat inductieve leermethoden in staat zijn om complexe, reële segmentatie- en classificatietaken te leren, mits er voldoende leervoorbeelden voorhanden zijn; er bestaan verschillende voorbeelden van toepassingen die dit aantonen. De in het oog lopende en intrigerende tegenstelling tussen de stellingen van de Chomskyaanse taalkunde aan de ene kant, en die van de structurele taalkunde en het automatisch leren aan de andere kant, bracht ons tot het uitvoeren van een empirische studie van het inductief leren van woorduit-

spraak. De resultaten van deze empirische studie staan ons toe om te stellen dat *inductieve leeralgoritmen in staat zijn om de uitspraak van woorden te leren met een bevredigend generaliseringsvermogen, zelfs wanneer de taakdefinitie (en ook de systeemarchitectuur) geen enkele van de abstractieniveaus reflecteert die expliciet als noodzakelijk worden verondersteld door taalkundige theorieën.*

Hoofdstuk 1 introduceert de bovengenoemde historische achtergrond van argumenten en tegenargumenten met betrekking tot de haalbaarheid van inductief leren van natuurlijke taal. Op basis van claims van de structurele taalkunde en het automatisch leren wordt vervolgens de probleemstelling geformuleerd.

In Hoofdstuk 2 wordt een overzicht gegeven van relevante achtergrondkennis. Er worden drie groepen inductieve leeralgoritmen beschreven die (in principe) geschikt zijn om woorduitspraak te leren: connectionistisch leren, instantie-gebaseerd leren, en het leren van beslissingsbomen. Het hoofdstuk biedt een overzicht van heersende taalkundige ideeën over morfologie en fonologie, en legt de nadruk op de abstractieniveaus die aanwezig worden verondersteld bij het uitspreken van woorden. Vervolgens wordt de gegevensbron beschreven waaruit de in de studie gebruikte voorbeelden van de uitspraak van woorden zijn gehaald: de Engelse lexicale data base van CELEX. Tenslotte wordt de methodologie beschreven die door de hele studie heen gevolgd is.

Hoofdstuk 3 beschrijft de toepassing van de geselecteerde leeralgoritmen op vijf deeltaken van de uitspraaktaak. De deeltaken representeren vijf taalkundig gemotiveerde abstractieniveaus binnen woorduitspraak: (i) morfologische segmentatie, (ii) grafemische ontleding, (iii) grafeem-foneem-omzetting, (iv) lettergreepsplitsing, en (v) klemtoontoekenning. De resultaten behaald met het toepassen van de vijf algoritmen op ieder van de deeltaken laten zien dat de algoritmen een redelijk tot excellent generaliseringsvermogen kunnen halen. Daarnaast laten de resultaten zien dat hoe minder een leeralgoritme abstraheert over het leermateriaal door gegevenscompressie, des te beter zijn generaliseringsvermogen is, op iedere deeltaak.

In Hoofdstuk 4 worden modulaire woorduitspraaksystemen geconstrueerd, geleerd en getest. De architectuur van de modulaire systemen is geïnspireerd op twee bestaande tekst-naar-spraak-systemen. Beide modulaire systemen voeren de uitspraaktaak uit in sequentieel verwerkende modules. De behaalde generaliseringsscores geven aan dat opeenstapelingen van fouten, doorgegeven van module naar module, de prestatie van de systemen ernstig hinderen. Om een gedeelte van deze ongewenste fouten

te ondervangen laten we het idee van een systeem met vijf modules varen en onderzoeken we twee systemen met drie modules. In deze drie-module systemen worden twee paren van deeltaken van de vijf-module systemen geïntegreerd tot enkele deeltaken. De precisie van deze systemen, die onderscheid maken tussen (i) morfologische segmentatie, (ii) grafeem-foneem-omzetting en (iii) klemtoontoekenning, is significant beter dan die van hun tegenhangers met vijf modules.

In Hoofdstuk 5 wordt het idee van sequentiële modulariteit vervangen door dat van parallelle modulariteit. Drie parallel-modulaire systemen worden getest. In het eerste systeem word woorduitspraak uitgevoerd door een enkele module, die letters omzet naar fonemen met klemtoonmarkeringen in een enkele omzettingsslag. Het tweede systeem voert twee deeltaken parallel uit, namelijk de omzetting van letters naar fonemen, en de omzetting van letters naar klemtoonmarkeringen. In het derde systeem wordt de letter-foneem-omzettingsdeeltaak verder uitgebreid tot 25 partiële deeltaken: ieder van deze partiële deeltaken representeert de herkenning van één articulatorisch kenmerk van het te classificeren foneem. De resultaten wijzen uit dat het systeem met de enkele module en het systeem met de twee modules beter presteren, en dat het articulatorisch-kenmerk-detectiesysteen slechter presteert dan het best presterende drie-modulesysteem beschreven in Hoofdstuk 4 dat getraind is met hetzelfde leeralgoritme.

Hoofdstuk 6 introduceert drie 'poortwachtersystemen' (*gating systems*) voor woorduitspraak waarin geen taalkundige kennis verwerkt is. De woorduitspraaktaak wordt in deze systemen opgesplitst in twee parallel-verwerkte partiële woorduitspraaktaken. In plaats van de taak op te splitsen op het uitvoerniveau wordt de taak opgesplitst op basis van het toepassen van een poortwachtercriterium op het invoerniveau: de spelling van (delen van) woorden. Drie poortwachtersystemen worden getest, te weten die met een toevalsgebaseerde poortwachter, een typicaliteits-gebaseerde poortwachter, en een voorkomen-gebaseerde poortwachter. Uit de resultaten blijkt dat in het toevalsgebaseerde poortwachtersysteem de woorduitspraaktaak slechter wordt geleerd dan de ongesplitste taak in zijn geheel. In de typicaliteits-gebaseerde en voorkomen-gebaseerde systemen wordt de taak met dezelfde precisie geleerd als de ongesplitste uitspraaktaak. Het aanbrengen van een poortwachter leidt niet tot verbetering van de prestatie, maar het is mogelijk om met een poortwachtersysteem de woorduitspraakgegevens automatisch op te delen in essentieel verschillende deelverzamelingen van gegevens.

Een samenvatting van de resultaten uit de Hoofdstukken 3 tot en met 6

wordt gegeven in Hoofdstuk 7. Speciale aandacht wordt besteed aan de mate van computationele efficiëntie van de gebruikte leeralgoritmen. Met instantie-gebaseerd leren worden de beste resultaten geboekt; de systemen die door het leren van beslissingsbomen worden gegenereerd bieden het beste evenwicht tussen generaliseringsvermogen en computationele efficiëntie. Hierop volgend worden analyses op de woorduitspraak-gegevens uitgevoerd om de oorzaken te zoeken voor het succes van instantie-gebaseerd leren: de analyses wijzen uit dat instanties van woorduitspraak in families voorkomen, die bestaan uit kleine aantallen leden met dezelfde klasse. Instantie-gebaseerd leren is in staat om goed te presteren op verzamelingen gegevens die dit soort kleine families (*small disjuncts*) bevatten. Vervolgens beschrijft het hoofdstuk een modulair woorduitspraak-systeem dat een aantal empirische bevindingen aangaande verschillende in het proefschrift onderzochte systemen combineert. Een analyse van het systeem toont aan dat het mogelijk is om betere woorduitspraaksystemen te bouwen door taalkundige en empirische kennis te combineren. Tenslotte biedt het hoofdstuk een overzicht van verwant onderzoek, en geeft een aantal indicaties voor verbetering van de huidige aanpak en onderwerpen van toekomstig onderzoek.

In Hoofdstuk 8 wordt de conclusie getrokken dat inductieve leermethoden, in het bijzonder instantie-gebaseerde leermethoden, in staat zijn om woorduitspraak te leren en daarbij een adequaat generaliseringsvermogen te bereiken. Taalkundige voorkennis in de taakdefinitie (en in de systeemarchitectuur) kan beperkt worden tot een absoluut minimum, kan met andere woorden *impliciet* gelaten worden, zonder dat het generaliseringsvermogen daar onder lijdt.