CHAPTER ONE

INTRODUCTION

1.1 HLT IN THE DEVELOPING WORLD

Human language technologies (HLT) hold much promise for the developing world, especially for user communities that have a low literacy rate, speak a minority language, or reside in areas where access to conventional information infrastructure is limited. For example, information systems that provide speech-enabled services via a telephone can serve a user in his or her language of choice in a remote location, without requiring additional expertise from the user, or a sophisticated Internet infrastructure. In South Africa, while Internet penetration is still low, more than 90% of the population has access to a telephone; and the potential of voice services for improving access to information is receiving increasing attention [1].

The development of various forms of human language technology - such as speech recognition, speech synthesis or multilingual information retrieval systems - requires the availability of extensive language resources. The development of these resources involves significant effort, and can be a prohibitively expensive task when such technologies are developed for a new language. The development of an accurate automatic speech recognition system, for example, requires access to an electronic pronunciation dictionary, a large annotated speech corpus from various speakers, and an extensive textual corpus. Such resources are freely available for only a small subset of the world's languages. This presents a significant obstacle to the development of HLT in the developing world, where few electronic resources are available for local languages, skilled computational linguists are scarce, and linguistic diversity is high. (India, for example, recognises nineteen official languages and South Africa eleven. In countries such as Indonesia and Nigeria, several hundred languages are widely spoken [2].)

In order to realise the potential benefit of HLT in the developing world, the language resource barrier must first be overcome: techniques are required that support the fast and cost-effective development of language resources in a new language, without extensive reliance on assistance from skilled computational linguists or access to prior language resources. Techniques that can support or accelerate the language resource development effort include the cross-language re-use of information [3, 4], statistical approaches to automated resource generation [5, 6], and bootstrapping, the focus of this thesis.

1.2 BOOTSTRAPPING OF HLT RESOURCES

The popular saying "to pull oneself up by one's bootstraps" is typically used to describe the process of "improving one's position by one's own efforts" [7]. In computer terminology this term was originally used to describe the process of iteratively loading a computer operating system from a few initial instructions, but soon came to describe any process where "a simple system activates a more complicated system" [8]. We use the term to describe an iterative process whereby a model of some form is improved via a controlled series of increments, at each stage utilising the previous model to generate the next one[1].

This generic technique has been applied successfully to the language resource development problem previously, especially in the creation of automatic speech recognition systems [3, 9–11]. When acoustic models are developed for a new target language, an automatic speech recognition system can be initialised with pre-developed models from an acoustically similar source language, and these initial models improved through an iterative process whereby audio data in the target language is automatically segmented using the current set of acoustic models, the models retrained and the target data re-segmented via a set of incremental updates. The potential saving in resource requirements achieved through such a process was well demonstrated by Schultz and Waibel [12], among others.

When considering resource bootstrapping approaches in more detail (as discussed in Chapter 3) it becomes apparent that these approaches rely on an automated mechanism that converts between various representations of the data considered. Each representation provides some specific advantage – making the data more amenable to a particular form of analysis – which can be utilised in improving or increasing the resource itself. In the above example, two representations are utilised: annotated audio data and acoustic models; and the mechanisms to move from one representation to the other are well defined through the phone alignment and acoustic modelling tasks, respectively.

The bootstrapping process has been applied successfully to a variety of additional language resource development tasks, including the development of parallel corpora [13], morphological dictionaries [14], morphological analysers [15] and linguistically tagged corpora [16]. We are specifically interested in the use of bootstrapping for the development of pronunciation models in new languages.

[1] The term 'bootstrapping' is discussed in more detail in Section 2.3.

1.3 PRONUNCIATION MODELLING WITHIN A BOOTSTRAPPING FRAMEWORK

A pronunciation model for a specific language describes the process of letter-to-sound conversion: given the orthography of a word, it provides a prediction of the phonemic realisation of that word. This is a component required by many speech processing tasks – including general domain speech synthesis and large vocabulary speech recognition – and is often one of the first resources required when developing speech technology in a new language.

The letter-to-sound relationship is typically modelled through explicit pronunciation dictionaries [17–19], but can also be represented according to various abstract letter-to-sound formalisms. Grapheme-to-phoneme (g-to-p) rule sets can either be hand-crafted, or a letter-to-sound representation can be obtained from a given training dictionary, using approaches such as neural networks, instance-based learning, decision trees, or pronunciation by analogy models [20–25]. In effect, these data-driven letter-to-sound formalisms provide a second representation of the training dictionary, by converting the training dictionary to a set of base elements and operators of some form, which we will refer to in general as g-to-p rule sets. These letter-to-sound formalisms have been studied over the past twenty years (as discussed in more detail in Section 2.2), resulting in a number of efficient representation techniques. Since efficient techniques exist to analyse the same pronunciation data according to more than one representation, it should be possible to utilise these representations during bootstrapping.

A letter-to-sound conversion mechanism is valuable, not only in the absence of explicit pronunciation dictionaries, but also in order to accommodate speech technology in memory constrained environments, or to deal with out-of-vocabulary words in speech systems. Such applications require a balance between the need for small rule sets, fast computation and optimal accuracy, and various approaches to pronunciation modelling have been defined to meet these requirements. Bootstrapping introduces an additional requirement: the ability to obtain a high level of generalisation given a very small training set. If such a g-to-p mechanism can be obtained, it seems probable that a bootstrapping approach will be beneficial in improving the speed and accuracy with which pronunciation models can be developed in a new language.

1.4 OVERVIEW OF THESIS

The aim of this thesis is to analyse the pronunciation modelling task within a bootstrapping framework. The goals are two-fold: (a) to obtain a mechanism for pronunciation modelling that is well suited to bootstrapping; and (b) to analyse the bootstrapping of pronunciation models from a theoretical and a practical perspective, as a case study in the bootstrapping of HLT resources.

The thesis is structured as follows:

- In Chapter 2 we provide background information with regard to the pronunciation modelling task and the use of bootstrapping for the development of HLT resources in general.
- In Chapter 3 we sketch a framework for analysing the bootstrapping process. This framework provides the context for subsequent chapters, and describes the requirements for a conversion algorithm suitable to bootstrapping.
- In Chapter 4 we analyse the grapheme-to-phoneme conversion task in the search for an appropriate conversion algorithm. This leads to the definition of *Default & Refine*, a novel algorithm for grapheme-to-phoneme rule extraction that is well suited to bootstrapping.
- In Chapter 5 we utilise the characteristics of the pronunciation modelling task analysed in the prior chapter in order to define a new framework for grapheme-to-phoneme prediction. We define the concept of *minimal representation graphs*, and demonstrate the utility of these graphs in obtaining a minimal rule set describing a given set of training data.
- In Chapter 6 we apply the new grapheme-to-phoneme algorithms in the bootstrapping of pronunciation models. We experiment with a number of options, and analyse the efficiency of this process according to the framework defined in Chapter 3. We develop bootstrapped pronunciation models in three languages (isiZulu, Afrikaans and Sepedi) and integrate the bootstrapped dictionaries in speech technology systems.
- In Chapter 7 we summarise the contribution of this thesis, and discuss further applications and future work.