Morpheme-Based Automatic Speech Recognition of Basque

Víctor G. Guijarrubia, M. Inés Torres, and Raquel Justo*

Departamento de Electricidad y Electrónica
Universidad del País Vasco, Apartado 644, 48080 Bilbao, Spain
{vgga,manes}@we.lc.ehu.es, raquel.justo@ehu.es

Abstract. In this work, we focus on studying a morpheme-based speech recognition system for Basque, an highly inflected language that is official language in the Basque Country (northern Spain). Two different techniques are presented to decompose the words into their morphological units. The morphological units are then integrated into an Automatic Speech Recognition System, and those systems are then compared to a word-based approach in terms of accuracy and processing speed. Results show that whereas the morpheme-based approaches perform similarly from an accuracy point of view, they can be significantly faster than the word-based system when applied to a weather-forecast task.

Keywords: Speech recognition, Morphological operations.

1 Introduction

Highly inflected languages present a large number of word forms per lemma. Therefore, the vocabulary of any application of natural language processing includes a vast variety of word-forms and is higher than in morphologically poor languages. So, it is usual to resort to morphological units to deal with the problems of higher vocabulary sizes, higher perplexities and higher out-of-vocabulary rates.

Basque is a pre-Indo-European language with unknown origin that is, along with Spanish, official language for the 2.5 million inhabitants of the Basque Country. Basque is a highly inflected language in both nouns and verbs.

The novelty of this work lies in the use of morphological units in Automatic Speech Recognition (ASR) of Basque. The use of morphological units has been studied to cope with the peculiarities of other highly inflected languages in ASR [1,2,3,4,5]. In this work we want to use them for Basque, studying how to integrate them in an ASR system and analysing several options that arise during the process, like the smoothing of the morpheme-based language models or the way to generate the word sequences. In Basque there is no freely available linguistic

* This work has been partially supported by the University of the Basque Country under grant GIU07/57, by the Spanish CICYT under grant TIN2005-08660-C04-03 and by the Spanish program Consolider-Ingenio 2010 under grant CSD2007-00018.

H. Araujo et al. (Eds.): IbPRIA 2009, LNCS 5524, pp. 386–393, 2009.
© Springer-Verlag Berlin Heidelberg 2009

tool that splits the words into proper morphemes either. So, for this reason, we have developed some techniques to split words into morpheme-like units by means of data-driven and linguistically motivated techniques. Overall, Basque can be characterised as a suffixing language, meaning that there is a tendency to modify words by appending one or more suffixes. However, it is convenient to keep a low morpheme-to-word ratio. There are many reasons for this. The higher the ratio the bigger the language models (LMs) we need in order to not lose information. The acoustic separability, the word generation or the chances of making incorrect splittings are also highly affected. Therefore, we decided to explore the approximation of decomposing the words, if possible, into two morpheme-like units: a *root* and an *ending* reassembling the suffixes. For that, two different approaches were studied to split the words into that form.

The rest of the paper is organised as follow: Section 2 describes all the information regarding the morphological units: the approximations explored to get the morphological units, the way they are integrated into an ASR system, analysing several ways to smooth the morpheme-based language models, and how to reconstruct the word sequence from the morphemes; Section 3 contains a description of the evaluation database used in the experiments; Section 4 presents the results obtained the different approximations; and Section 5 discusses the conclusions.

2 Morphological Units

2.1 Getting the Morphological Units

Two different approaches were studied to split the words, if possible, into the desired *root* - *ending* scheme.

Lemma-based approach (LEMMA). This approach has a linguistic motivation since it requires the data to be in not only word form, but also in lemma form. The lemmas are used to get the morphemes, but not directly to get the splittings. The words that make up the vocabulary are grouped according to the lemma. The words belonging to the same group are then analysed to find a common root. Those words are then split into a *common_root* - *different_ending* form. However there are three particular cases where the word is not split and is left as a full word: a) when the ending is *null* b) when for a specific lemma only one word appears and c) when there are some words for the same lemma, but they do not have a common root, like irregular verbs. So, after applying this procedure, the result is a list of units composed of words, roots and endings. Note also that if a word corresponds to more than one lemma, because it has several different meanings, for each of the cases the word appears in the data, it is split according to the lemma in that particular case.

Morfessor-based approach (MORF). This is a data-driven approach based on unsupervised learning of morphological word segmentation. This approximation requires the use of the Morfessor Categories-MAP software [6]. The tool is capable of segmenting words in an unannotated text corpus into morpheme-like units. It