A Syllabification Algorithm for Spanish

Heriberto Cuayáhuitl

Universidad Autónoma de Tlaxcala,
Department of Engineering and Technology,
Intelligent Systems Research Group,
Apartado Postal #140, Apizaco, Tlaxcala, Mexico, 90300.
hcuayahu@ingenieria.uatx.mx

Abstract. This paper presents an algorithm for dividing Spanish words into syllables. This algorithm is based on grammatical rules which were translated into a simple algorithm, easy to implement and with low computational cost. Experimental results in an evaluation text corpus show an overall error rate of 1.6%. Most of the error is attributed to words with diphthongs and to confusion in the use of prefixes where grammatical rules are not always absolute. Syllabification is an essential component of many speech and language processing systems, and this algorithm might be very useful to researchers working with the Spanish language.

1 Introduction

Currently, the development of speech synthesizers and speech recognizers, frequently requires working with subword units such as syllables [1-3]. For instance, robust speech recognition often makes use of word spotters based on syllables for detecting Out-of-Vocabulary (OOV) speech [2] and for modeling unknown words in spontaneous speech [3]. Today, every new system being developed requires the implementation of a new algorithm for dividing words into syllables, due to the fact that a formal algorithm shared among the linguistic community does not exist, at least for Spanish. In the linguistic literature, we can find grammatical rules or attempts to explain the division of words into syllables step-by-step, but nothing beyond that. In the past, syllabification algorithms have been proposed for different languages, including English and German, among others [4], implemented as a weighted finite state transducer, but this is not the case for Spanish, where few research efforts have been documented. Thus, the purpose of this work is to formulate an algorithm for dividing Spanish words into syllables, and to share this algorithm so that other researchers in the area of speech and language processing will not have to duplicate the work.

In this research, is proposed an algorithm to divide Spanish words into syllables. Our experiments are based on a text corpus containing representative words for each grammatical rule. Results are given in terms of a simple division of correctly syllabified words by the total number of words. In the remainder of this paper we first provide an overview of Spanish syllabification in section 2. In section 3 we describe the syllabification algorithm itself. In section 4 we present experimental results. Finally, we provide some conclusions and future directions.