STEMBR: A Stemming Algorithm for the Brazilian Portuguese Language

Reinaldo Viana Alvares, Ana Cristina Bicharra Garcia, and Inhaúma Ferraz

UFF – Universidade Federal Fluminense, Instituto de Computação,
Rua Passo da Pátria, 156 Bloco E - 3º Andar,
São Domingos, Niterói, RJ 24210-240
{ralvares, bicharra, ferraz}@ic.uff.br

Abstract. Stemming algorithms have traditionally been utilized in information retrieval systems as they generate a more concise word representation. However, the efficiency of these algorithms varies according to the language they are used with. This paper presents STEMBR, a stemmer for Brazilian Portuguese whereby the suffix treatment is based on a statistical study of the frequency of the last letter for words found in Brazilian web pages. The proposed stemmer is compared with another algorithm specifically developed for Portuguese. The results show the efficiency of our stemmer.

1 Introduction

A word is a string of letters organized in such a way that they can represent the meaning of language expression about objects and ideas. It is common in texts to find affixal variations of words. For example, the words "cantores (singers)", "cantora (female singer)" and "canto (song or singing)" represent, in generic terms, the meaning of "cantar (to sing)". Research in the area of information retrieval (IR) aims to find a single representation for such words, thus providing the user with a broader search result. Stemming algorithms are an option for this task.

Stemming is the process of converting variations of a word into a concise and accurate representation. The concept adheres to the annotation principle [10]. The aim of the stemming process is to merge words, that have a common meaning, into a single representation known as a stem. The stem is the result of the process of stemming. In the stemming process, two kinds of error may occur:

- Over stemming: removing too many letters, so that words with different meanings are merged to a single stem. See Table1:

Table 1. Example of over stemming errors

Meaning	Word	Stem
comportamento (behavior)	comportado (well behaved)	comp
comparar (compare)	comparou (compared)	comp

- Under stemming: leaving too many letters, so that words with the same core meaning are merged into different stems. See Table 2:

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 693 – 701, 2005.

Table 2. Example of under stemming errors

Meaning	Word	Stem
movimento (movement)	movimentação (moving)	movimentaç
movimento (movement)	movimentar (to move)	movimenta

Various different stemming algorithms have been proposed for the English language [3]. However, few solutions have been put forward regarding the Portuguese language.

Below, we present the approaches most commonly used in the stemming process, the methods for evaluating these algorithms, the STEMBR, a case study, and overall comments regarding the work.

2 Stemming Algorithms

In this section, we present the approaches most commonly used in the stemming process, which are: affix stripping, table lookup and statistical methods.

2.1 Affix Stripping

This approach is dependent on the morphology of the target language. The stem is obtained by stripping some elements (morphemes) from the beginning/end of the word.

We find here the most traditional method of extracting suffixes: the Porter algorithm [9]. Originally developed for the English language, this stemmer is made up of five steps, during which certain rules are applied to the words and the most common suffixes are removed. Based on a specific measurement, relating to the number of vowels/consonants in a word, the algorithm attempts to avoid removing letters when the stem is very short. As it was a pioneering work in this area, the method has been adapted for a variety of languages, including Brazilian Portuguese.

A stemmer specially developed for the Portuguese language is presented in [4], and has shown itself to be more efficient than the Brazilian version of the Porter algorithm. This stemmer, hereon referred to as STEMP, comprises 8 steps (plural, feminine, augmentative, adverb, noun and verb reduction, remove vowel and remove accents) that are performed in a predetermined order. Each step is made up of a set of rules that are sequentially applied, but with only one rule being applied in each instance. Each rule has four elements:

- The suffix to be removed;
- The minimum size of the stem;
- An alternative suffix, if necessary, and;
- A list of exceptions.

To illustrate a rule: *{"ura",4,"",{"acupuntura", "costura"}}*, where "ura" is the suffix to be removed, 4 is the minimum stem size, and the words in inverted commas represent the list of exceptions, in this case because there is no alternative suffix.