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:

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.

2.2 Table Lookup

Under this approach, the stemming process is performed manually, wherein the stems are defined for each word and stored in some kind of structured form. The advantage is that it generates perfect stems. However, the approach is limited to retrieving only those words that have been previously stored. What is more, the space occupied for storage tends to grow as the corpus expands, which can make the search process inefficient.

2.3 Statistical Stemmers

Here, the stemming process involves statistical methods whereby, through a process of inference and based on a corpus, rules are formulated regarding word formation. Some of the methodologies adopted are: frequency counts, n-gram [7], link analysis [1], and Hidden Markov Models (HMM) [6]. This approach does not require any linguistic knowledge whatsoever, being totally independent of the morphological structure of the target language.

3 Evaluation Methods

In this section, we present the methodologies utilized to evaluate the performance of the stemmers, namely: the manual method, vocabulary reduction and Paice's method.

In the manual method, a human being, who decides the correct stem for each word, performs the stemming process. Three evaluation measurements are obtained in this manner: the number of correct results; the number of errors due to over stemming; and the number of errors due to under stemming.

One of the purposes of stemmers is to reduce the size of the vocabulary for indexing purposes. The vocabulary reduction is obtained by dividing the number of words in the corpus by the number of stems generated, excluding repetitions.

In Paice's method [8], three measurements are implemented: the over stemming index (OI); the under stemming index (UI); and the stemming weight (SW). This method requires a word sampling, with no repetitions, separated into conceptual groups in which the words are semantically and morphologically related. The over stemming and under stemming errors are counted for each group and the OI and UI are calculated for all the groups. The SW is given by the ratio OI/UI.

4 STEMBR

Our algorithm composes he classical steps and the affix stripping is adequate to Portuguese language. The rationale, affix treatments and the STEMBR model are presented below.

4.1 The Rationale for the STEMBR

4.1.1 Corpus

Our stemmer is based on a statistical study (the Evaluator module in figure 1) of the LexWeb corpus [5]. The LexWeb is a lexical generator for the Portuguese language, constructed with tools that visit Brazilian web pages and select the most frequently used words. The size of the corpus is approximately 130,000 words.

4.1.2 The Most Common Last Letters

From the statistical study, a list was obtained showing the frequency, in descending order, of the last letter of the words surveyed. The nine most common letters on the list (see figure 2) represent approximately 85% of the total sampling.

4.1.3 Suffix Size Ordering

An important factor in the construction of the stemmer is the way in which each suffix is removed. For example, consider the word "cantávamos (we sang)", and the suffixes "mos" and "ávamos". Either of these suffixes could be stripped from the word, as both represent its substrings. The best choice would be to remove the largest substring, generating the stem "cant".

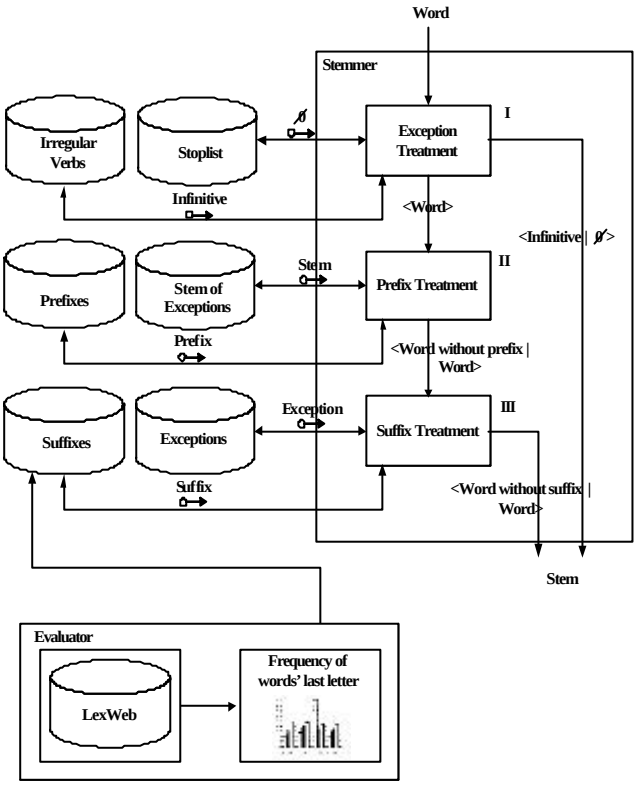

Figure 1. The STEMBR Model

4.2 Prefix Treatment

The treatment of prefixes is simple. Known prefixes are stripped form the words. The only peculiarity occurs when the substring does not represent a prefix. For example, in the word "impossibilidade (impossibility)", the substring "im" performs the function of a prefix. However, in the word "imagem (image)" this is not the case, so its removal would cause the word to lose all meaning. In order to minimize this problem, the algorithm is given a list of exceptions (25 prefixes), which contains the

stems of words whose substrings are not prefixes. Prefixes that generate many exceptions are not treated.

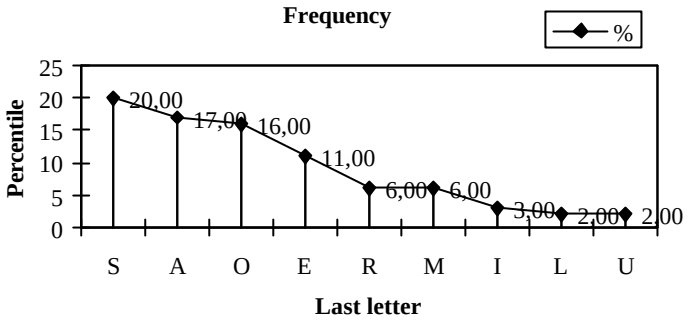

Figure 2. Last letter frequency for words in the LexWeb corpus

4.3 Suffix Treatment

Our challenge is to obtain a simple and efficient stemmer for the Portuguese language. To this end, the suffix treatment is performed for each subset of words from the list generated by the statistical study (words that have the same last letter). The list of suffixes is organized so as to strip the largest suffix from each word. Nevertheless, the key issue relates to the order in which the suffix treatment is performed, in an attempt to generate the fewest possible errors. With the empirical assistance of a Lexicographical Specialist, and following exhaustive testing, we arrived at a configuration that generated the most satisfactory results. The best order for performing the suffix treatment is: “S”, “R”, “M”, “L”, “O”, “A”, “U”, “E” and “I”. The reasoning behind this includes:

- In the Portuguese language, with few exceptions, words with suffixes ending in “s” represent the plural form and are generally longer (number of letters) than the singular form;
- After stripping the suffixes ending in “r”, “m” or “l”, there are cases where the substrings generated still do not represent the correct stem of the word. Some of these cases are resolved after performing suffix treatment for the vowel endings (“a”, “e”, “o”, “i” and “u” – the latter two are considered semi-vowels in Portuguese). Changing the suffix treatment order among the three consonants did not make any significant difference to the results of the process;
- The vowels “o” and “a” represent approximately 33% of the total sampling. Again, no significant differences were noted when their order was changed.

4.4 The STEMBR Model

In the STEMBR model, every word is sequentially submitted to three modules: specific cases; prefix reduction; and suffix reduction.

- Specific cases: at this instance, the word is checked to see if it belongs to some special category, for which prefix and suffix reductions would be inappropriate. Verbs with irregular conjugations are examples of such words. These verbs are merged in their infinitive form. This list was obtained through a Lexographical Specialist's help.
- Prefix reduction: at this point, the prefix treatment is performed. This process strips the prefix from the word, if it is not found on the list of exceptions;
- Suffix reduction: the suffix treatment is the most important process to be performed. The process comprises a set of rules whereby the longest suffix is stripped from the word. Some of these rules were obtained from [4], while others were specially created, through a Lexographical Specialist's help. The stemmer currently has 394 rules. Each rule has three elements: the suffix to be removed; the minimum size of the stem; and a list of exceptions. The minimum stem size is to avoid generating extremely short stems. This technique, utilized in [4], helps to avoid over stemming errors. Meanwhile, the list of exceptions organizes words that, despite ending in a suffix form, should not have this ending removed. In the rule below, the words on the list of exceptions do not have the suffix "mento" removed.

```
{"mento",{"complemento(complement)","instrumento
   (instrument)", "departamento (department)"}}
```

Let us consider the word "irritadíssina" (extremely annoyed, in English). This word is sequentially submitted to three modules:

- Specific cases: If the word is a irregular verb or a stopword it is substituted by the infinitive form or eliminated (stopwords);
  Else the prefix treatment is performed.
- Prefix reduction: If the initial substring of the word is a known prefix the prefix is eliminated;
  Else the suffix treatment is performed.
- Suffix reduction: The suffix treatment begins identifying the final letter of the word. The list of final substrings ending with the encountered letter is classified in length descending order and is used for comparison. The longest suffix is stripped. The "adíssima" substring is eliminated generating the token "irrit" (see figure 3).

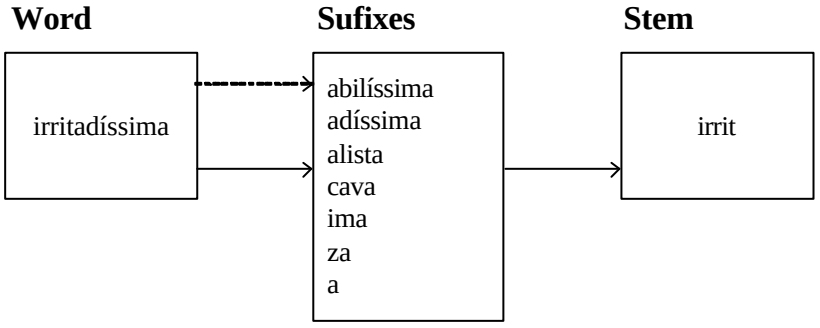

Figure 3. Suffix reduction for the word "irritadíssima"

5 Case Study

In order to evaluate the quality of the STEMBR model, tests were carried out involving the three evaluation methods presented, comparing the model against the stemmer proposed in [4] and the Brazilian version of the Porter algorithm.

5.1 Corpus

We used two samples, of different sizes and origins, as follows:

- 1,000 words, taken from the electronic dictionary "Aurélio – Século XXI (Twenty-first Century)" [2], hereon referred to as "Sample I";
- 5,000 words, taken from the LexWeb, hereon referred to as "Sample II".

5.2 The Manual Method

A Lexicographical Specialist, being a person with considerable Portuguese language experience, performed the manual stemming procedure. The test results are shown in tables 3 and 4:

Table 3. Results of the test using Sample I

Stemmer	Correct	Over stemming	Under stemming
STEMBR	62.20 %	8.90 %	27.10 %
STEMP	55.30 %	4.70 %	37.70 %
PORTER	43.80 %	1.30 %	51.20 %

Table 4. Results of the test using Sample II

Stemmer	Correct	Over stemming	Under stemming
STEMBR	69.02 %	12.05 %	17.96 %
STEMP	67.60 %	8.96 %	22.58 %
PORTER	57.86 %	5.00 %	34.48 %

5.3 Vocabulary Reduction

Tables 5 and 6 show vocabulary reduction of the samples:

Table 5. Vocabulary reduction using Sample I

Stemmer	Reduction to
STEMBR	29.20 %
STEMP	32.70 %
PORTER	40.50 %

Table 6. Vocabulary reduction using Sample II

Stemmer	Reduction to
STEMBR	53.92 %
STEMP	53.90 %
PORTER	60.00 %

5.4 Paice Evaluation

A Lexographical Specialist manually generated a total of 102 and 2,696 semantic groups, respectively, for Sample I and Sample II. The test results are shown in tables 6 and 7:

Table 7. Results of Paice's Method using Sample I

Stemmer	OI	UI	SW
STEMBR	7.30×10^{-4}	0.447	1.60×10^{-3}
STEMP	7.09×10^{-4}	0.492	1.44×10^{-3}
PORTER	3.06×10^{-4}	0.537	0.67×10^{-3}

Table 8. Results of Paice's Method using Sample II

Stemmer	OI	UI	SW
STEMBR	1.01×10^{-4}	0.288	3.50×10^{-4}
STEMP	0.98×10^{-4}	0.295	3.30×10^{-4}
PORTER	0.50×10^{-4}	0.395	1.25×10^{-4}

5.5 Discussion

The test results showed that:

- Using the manual method ("Gold Standard"), with Sample I, the STEMBR model obtained a correct rate approximately 7% higher than the stemmer STEMP and 18.4% higher than the PORTER stemmer. With Sample II, the STEMBR and STEMP were practically equal (the STEMBR having an advantage of 1.42%) and they obtained a correct rate approximately 10% higher than the PORTER stemmer. Our stemmer obtained a lower rate of under stemming and a higher rate of over stemming errors than the STEMP, in both the samples;
- The vocabulary reduction of STEMBR was 3.5% lower than that of STEMP for Sample I and they were practically equal for Sample II;
- Using Paice's method, the rate of under stemming for STEMBR was lower than STEMP and PORTER stemmer, in both the samples;

- Using Paice's method, the rate of over stemming for PORTER stemmer was lower for both samples;

6 Conclusions

This paper presented the development of a stemmer for Brazilian Portuguese, wherein the suffix treatment is performed for each subset of words with a common last letter. The test results show that the STEMBR model is more efficient than the STEMP reference model in terms of under stemming errors, and less so with regard to over stemming errors. In practice, it can be seen that it is a conflicting task to try reducing the two types of error. It is our intention to apply our stemmer on an Information Retrieval system to assess its impact over recall and precision.

References

1. Bacchin, M.., Ferro, N. and Melucci, M.. University of Padua at CLEF 2002: Experiments to evaluate a statistical stemming algorithm. Working Notes for CLEF 2002, pages 161-168: In *Proceedings*, Rome, September 2002.
2. Ferreira, A. B. H.. *Dicionário Aurélio Eletrônico*. CD-ROM (In Portuguese). Nova Fronteira. 1999.
3. Frakes, W. and B. Yates, R.. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, NJ, 1992.
4. Orengo, V. and Huyck, C.. A Stemming Algorithm for The Portuguese Language. In *Proceedings of Eighth Symposium on String Processing and Information Retrieval (SPIRE 2001)*, pages 186-193, Laguna de San Raphael, Chile, November 2001.
5. Junior, A. M.. *LexWeb: um léxico da língua portuguesa extraído automaticamente da internet*. Master Thesis (in Portuguese). Programa de Pós-Graduação em Engenharia Elétrica. UFPA, November 2004.
6. Melucci, M. and Orio, N. A Novel Method for Stemmer Generation Based on Hidden Markov Models. In *Proceedings of Conference on Information and Knowledge Management (CIKM03)*, pages 131-138, New Orleans, LA, November 2003. ACM Press.
7. Mayfield, J. and McNamee, P.. Single N-gram Stemming. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 415-416, Toronto, Canada, July 2003. ACM Press.
8. Paice, C.: An evaluation method for stemming algorithms, in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42-50, Dublin, Ireland, July 1994. ACM Press.
9. Porter, M.. An Algorithm for Suffix Stripping. *Program*, 14(3), 130-137, July 1980.
10. Stefik, M. *Introduction to Knowledge systems*. Morgan Kaufmann Publishers. 1995.