A Study on the use of Stemming for Monolingual Ad-Hoc Portuguese Information Retrieval

Viviane Moreira Orengo
Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
vmorengo@inf.ufrgs.br

Abstract

For UFRGS's first participation on CLEF our goal was to compare the performance of heavier and lighter stemming strategies using the Portuguese data collections for Monolingual Ad-hoc retrieval. The results show that the safest strategy was to use the lighter alternative (reducing plural forms only). On a query-by-query analysis, full stemming achieved the highest improvement but also the biggest decrease in performance when compared to no stemming. In addition, statistical tests showed that the only significant improvement both in terms of mean average precision and precision at ten was achieved by our lighter stemmer.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Linguistic processing. H.3.4 [Systems and Software]: Performance evaluation

General Terms

Experimentation, Performance, Measurement, Algorithm

Additional Keywords and Phrases

Stemming algorithms, Portuguese language

1 Introduction

This paper reports on monolingual information retrieval experiments that we have performed for CLEF2006. We took part on the ad-hoc monolingual track, focusing on the Portuguese test collections.

Our aim was to compare the performance of lighter and heavier stemming alternatives. We compared two different algorithms: a Portuguese version of the Porter stemmer[1] and the "Removedor de Sufixos da Língua Portuguesa (RSLP)" (Orengo & Huyck, 2001).

The remainder of this paper is organised as follows: Section 2 presents RSLP stemmer; Section 3 discusses the experiments and results; and Section 4 presents the conclusions.

2 The Stemming Algorithm

We have used the RSLP algorithm, proposed in our earlier work (Orengo & Huyck, 2001). This section introduces the algorithm. The RSLP is based solely on a set of rules (not using any dictionaries) and is composed by 8 steps that need to be executed in a certain order. Figure 1 shows the sequence those steps must obey:

[1] Available from http://www.snowball.tartarus.org/algorithms/portuguese/stemmer.htm

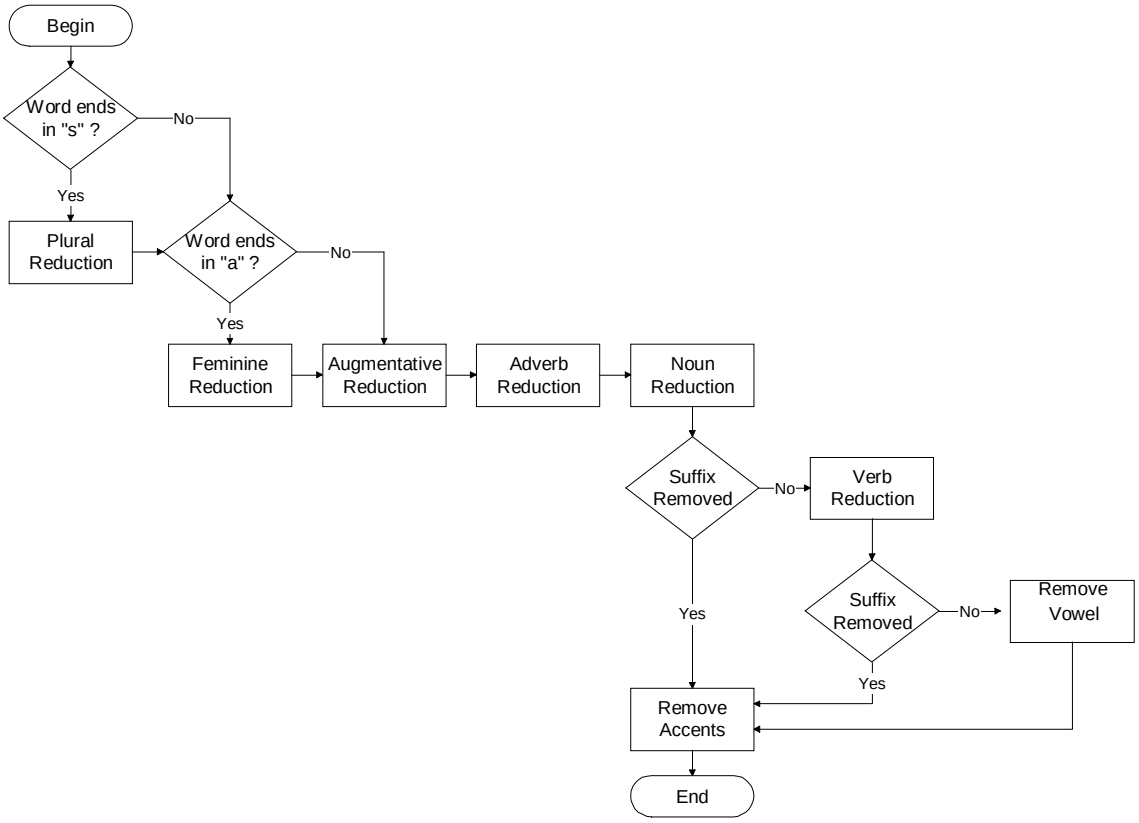

Figure 1. Sequence of steps for the Portuguese Stemmer

Each step has a set of rules, the rules in the steps are examined in sequence and only one rule in a step can apply. The longest possible suffix is always removed first because of the ordering of the rules within a step, e.g. the plural suffix *–es* should be tested before the suffix *–s*. At the moment, the Portuguese Stemmer contains 199 rules. please refer to (Orengo & Huyck, 2001) for the complete list.

Each rule states:

- The suffix to be removed;
- The minimum length of the stem: this is to avoid removing a suffix when the stem is too short. This measure varies for each suffix, and the values were set by observing lists of words ending in the given suffix. Although there is no linguistic support for this procedure it reduces overstemming errors. Overstemming is the removal of a sequence of characters that is part of the stem and not a suffix.
- A replacement suffix to be appended to the stem, if applicable;
- A list of exceptions: for nearly all rules we defined, there were exceptions, so we added exception lists for each rule. Such lists were constructed with the aid of a vocabulary of 32,000 Portuguese words freely available from (Snowball). Tests with the stemmer have shown that exceptions list reduce overstemming errors by 5%.

An example of a rule is:

```
"inho", 3, "", {"caminho", "carinho", "cominho",
"golfinho", "padrinho", "sobrinho", "vizinho"}
```

Where "*inho*" is a suffix that denotes diminutive, 3 is the minimum size for the stem, which prevents words like "*linho*" (linen) from being stemmed and the words between brackets are the exceptions for this rule, that is, they end in the suffix but they are not diminutives. All other words that end in *–inho* and that are longer than 6 characters will be stemmed. There is no replacement suffix in this rule.

Below we explain the eight steps involved in our stemming procedure.

Step 1: Plural Reduction

With rare exceptions, the plural forms in Portuguese end in *–s*. However, not all words ending in *–s* denote plural, e.g. *lápis*, (pencil). This step consists basically in removing the final "*s*" of the words that are not listed as exceptions. Yet sometimes a few extra modifications are needed e.g. words ending in *–ns* should have that suffix replaced by "*m*" like in *bons → bom*.

Step 2: Feminine Reduction

All nouns and adjectives in Portuguese have a gender. This step consists in transforming feminine forms to their corresponding masculine. Only words ending in *–a* are tested in this step but not all of them are converted, just the ones ending in the most common suffixes, e.g. *chinesa → chinês.*

Step 3: Adverb Reduction

This is the shortest step of all, as there is just one suffix that denotes adverbs *–mente*. Again not all words with that ending are adverbs so an exception list is needed.

Step 4: Augmentative/Diminutive Reduction

Portuguese nouns and adjectives present far more variant forms than their English counterparts. Words have augmentative, diminutive and superlative forms e.g. "small house" = *casinha*, where *–inha* is the suffix that indicates a diminutive. Those cases are treated by this step. According to (Cunha & Lindley-Cintra, 1985) there are 38 of these suffixes, however some of them are obsolete therefore, in order to avoid overstemming, our algorithm uses only the most common ones that are still in common usage.

Step 5: Noun Suffix Reduction

This step tests words against 61 noun (and adjective) endings. If a suffix is removed here, steps 6 and 7 are not executed.

Step 6: Verb Suffix Reduction

Portuguese is a very rich language in terms of verbal forms, while the regular verbs in English have just 4 variations (e.g. talk, talks, talked, talking), the Portuguese regular verbs have over 50 different forms (Macambira, 1999). Each one has its specific suffix. The verbs can vary according to tense, person, number and mode. The structure of the verbal forms can be represented as: root + thematic vowel[2] + tense + person, e.g. *and* + *a* + *ra* + *m* (they walked). Verbal forms are reduced to their root.

Step 7: Vowel Removal

This task consists in removing the last vowel ("a", "e" or "o") of the words which have not been stemmed by steps 5 and 6, e.g. the word *menino* (boy) would not suffer any modifications by the previous steps, therefore this step will remove its final *–o*, so that it can be conflated with other variant forms such as *menina*, *meninice*, *meninão*, *menininho*, which will also be converted to the stem *menin*.

Step 8: Accents Removal

Removing accents is necessary because there are cases in which some variant forms of the word are accented and some are not, like in *psic**ó**logo* (psychologist) and *psic**o**logia* (psychology), after this step both forms would be conflated to *psicolog*. It is important that this step is done at this point and not right at the beginning of the algorithm because the presence of accents is significant for some rules e.g. *ói**s** → ol* transforming *sóis* (suns) to *sol* (sun). If the rule was o*is → ol* instead, it would make mistakes like stemming dois (two) to *dol*.

The Portuguese version of the Porter Stemmer and the RSLP are based solely on rules that need to be applied in a certain order. However there are some differences between the two stemmers:

---

[2] There are 3 classes of verbs in Portuguese according to the ending of their infinitive form: "ar", "er", "ir". Thematic Vowel the letter ("-a", "-e" and "-i") that groups verbs into categories.

- The number of rules – RSLP has many more rules than the Portuguese Porter because it was designed specifically for Portuguese. There are some morphological changes such as augmentatives and feminine forms that are not treated by the Portuguese Porter.
- The use of exceptions lists – RSLP includes a list of exceptions for each rule as they help reducing overstemming errors.
- The steps composing the two algorithms are different.

3 Experiments

This section describes our experiments submitted to the CLEF-2006 campaign. Section 3.1 details the resources used, and Section 3.2 presents the results.

3.1 Description of Runs and Resources

The Portuguese data collections were indexed using SMART[3]. We used the title and description fields of the query topics. Query terms were automatically extracted from the topics. Stop words were removed from both documents and topics. In addition, terms such as "find documents" were removed from the topics. The processing time was less than 4 minutes for all runs. This includes indexing the 210,734 documents and running all 50 queries.

Four runs were tested:

- NoStem – No stemming was applied, this run was used as the baseline
- Porter – Full stemming using the Portuguese version of the Porter stemmer
- RSLP – Full stemming using the RSLP stemmer
- RSLP-S – applying only the first step of RSLP to deal with plural reduction only

3.2 Results

Table 1 shows the number of terms indexed in each run. Full stemming with RSLP achieved the highest reduction on the number of entries, followed by the Portuguese version of the Porter stemmer. The lighter stemming strategy reduced the number of entries by 15%.

Table 1 – Number of Terms in the Dictionary for all runs. The percentages indicate the reduction attained by each stemming procedure in relation to the baseline

Run	Number of Terms
NoStem	425996
Porter	248121 (-41.75%)
RSLP	225356 (-47.10%)
RSLP-S	358299 (-15.89%)

Table 2 – Results in terms of MAP and Pr@10. The asterisk denotes a statistically significant improvement in relation to the baseline

Run	Mean Average Precision	Precision at 10
NoStem	0.2590	0.3880
Porter	0.2790 (+7.72%)	0,4260 (+9.79%)
RSLP	0.2790 (+7.72%)	0,4320 (+11.34%)
RSLP-S	0.2821 (+8.91%)*	0,4300 (+10.82%)*

The results show that the best performance, in terms of mean average precision (MAP), was achieved by RSLP-S. Both runs in which full stemming was performed achieved identical results in terms of MAP. However, the RSLP outperformed the Portuguese version of the Porter stemmer in terms of Pr@10, but the difference was only marginal.

[3] Available from ftp://ftp.cs.cornell.edu/pub/smart/

In order to tell whether the performance improvements shown in Table 1 are statistically significant, a paired T-test was performed. Although our data is not perfectly normally distributed, Hull (Hull, 1993) argues that the T-test performs well even in such cases. The standard threshold for statistical significance (α) of 0.05 was used. When the calculated p value is less than α, there is a significant difference between the experimental runs. The results of the statistical tests show that full stemming does not produce a statistically significant improvement (in terms of both MAP and Pr@10) for either algorithm (p values of 0.25 for RSLP and 0.22 for Porter considering MAP and p values of 0.14 for RSLP and 0.18 for Porter when analysing Pr@10). RSLP-S, however, has achieved a statistically significant improvement compared to baseline for both MAP and Pr@10 (p values of 0.003 for MAP and 0.01 for Pr@10). Figure 2 shows recall-precision curves for all runs.

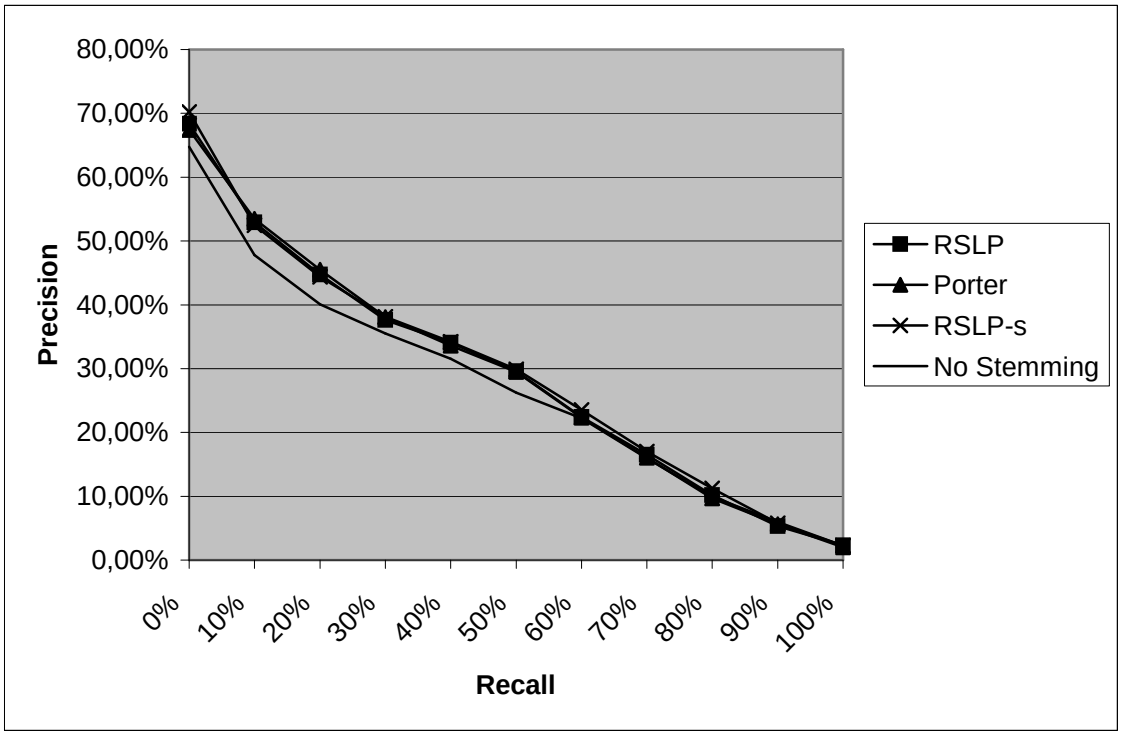

Figure 2 Recall-precision curves

A query-by-query, analysis shown in Table 3, demonstrates that for 12 topics no stemming was the best alternative. Some form of stemming helped 38 out of 50 topics. Confirming the results in terms of MAP and Pr@10, the best performance was achieved by the lighter stemming alternative RSLP-S. Full stemming with RSLP achieved the biggest performance improvement (topic 340 AvP 0.0003 → 0.3039), but also the biggest drop (topic 343 AVP 0.4276 → 0.1243). Stemming also helped finding 221 relevant documents that were not retrieved by the NoStem run.

Table 3 – Runs and the number of topics in which they achieved the best average precision

Run	Number of Topics
NoStem	12
Porter	10
RSLP	12
RSLP-S	16
Total	**50**

It seemed plausible that queries with few relevant documents would benefit more from stemming, resulting in a negative correlation between the number of relevant documents for the topic and the change in performance achieved with stemming. However a weak positive correlation of 0.15 was found. We would like to be able to predict the types of queries that would be benefited from stemming, but that needs further analysis with a larger number of topics.

4 Conclusions

This paper reported on monolingual ad-hoc IR experiments using Portuguese test collections. We evaluated the validity of stemming comparing the Portuguese version of the Porter stemmer and two versions of the RSLP stemmer, one that applies full stemming and one that only reduces plural forms Below we summarise our conclusions:

- The lighter version of the RSLP stemmer yields statistically significant performance improvements both in terms of MAP and Pr@10.
- Full stemming, both with Porter and RSLP, has improved the results in terms of MAP and Pr@10. However the difference was statistically significant.
- On a query-by-query analysis we found that stemming helped 38 out of 50 topics and that it enabled the retrieval of 221 further relevant documents that were missed by the run in which no stemming was used.

Acknowledgements

This work was supported by a CAPES-PRODOC grant from the Brazilian Ministry of Education.

References

Cunha, C., Lindley-Cintra, L., 1985. *Nova Gramática do Português Contemporâneo*. Rio de Janeiro: Nova Fronteira (in Portuguese).

Hull, D.,1993. Using Statistical Testing in the Evaluation of Retrieval experiments. In *ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 329-338). Pittsburgh: ACM.

Macambira, J. R., 1999. *A Estrutura Morfo-Sintática do Português*. São Paulo, Brazil: Ed. Pioneira (in Portuguese).

Orengo, V. M., Huyck, C. R.,2001. A Stemming Algorithm for the Portuguese Language. In *8th International Symposium on String Processing and Information Retrieval (SPIRE)*. Laguna de San Raphael, Chile.

Snowball. Retrieved 29/07/2006, from http://snowball.tartarus.org/portuguese/voc.txt