Evaluation of Stemming Errors: Towards a Qualitative Analysis

Reinaldo Viana Alvares, Rubem Mondaini
Federal University of Rio de Janeiro – UFRJ – CT/COPPE,
Ilha do Fundão, 21.941-972, P. O. Box 68511, Rio de Janeiro, Brazil
E-mail: reinaldoviana@cos.ufrj.br, mondaini@cos.ufrj.br

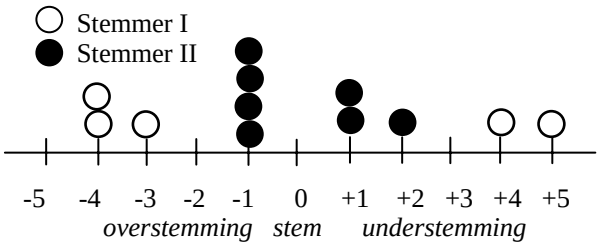
ABSTRACT

In this work we stress the importance of the qualitative evaluation approach of stemming algorithms. The evaluation methods of stemming errors are usually depicted in a quantitative analysis [2].

The application of stemming algorithms to a previously chosen word leads to the isolation of its stem. The stem is considered a concise representation of a word and should be seen as its smallest and unambiguous root [1]. It should also be sufficiently broad in order to capture its meaning as well as its multiple variations [3]. A basic example could be represented by the words *biology*, *biologist* and *biological*, all of which can be represented by the correspondent stem *biolog*.

The results of algorithm applications may be classified as understemming and overstemming. The first group corresponds to the prediction of a larger stem as the correct one. The prediction of the second group is a smaller stem. This will also imply a generation of different or equal stems for synonymous words, respectively [4].

Our aim is the proposal of a quantitative method for the assessment of stemming algorithms. The figure below hopes to convey an example of the proposed method.



We note that Stemmer II, though performing poorly in the quantitative method, displays a better qualitative performance since its predictions share more proximity to the assumed optimum.

References

[1] Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley, Boston, MA, 1999.

[2] Paice, C.D.: An Evaluation Method for Stemming Algorithms, in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press. Pages 42-50, Dublin, Ireland, 1994.

[3] Fuller, M., and Zobel, J. Conflation-based Comparison of Stemming Algorithms. In Proceedings of the Third Australian Document Computing Symposium, Pages 8-13 Sydney, Australia, 1998.

[4] Alvares, R. V.; Garcia, A. C. B.; Ferraz, I. N. STEMBR: A Stemming Algorithm for the Brazilian Portuguese Language. In: 12th Portuguese Conference on Artificial Intelligence (EPIA 2005), Covilhã, Portugal. Lecture Notes in Artificial Intelligence. v. 3808. Pages 693-701, 2005.