Modified Makagonov's Method for Testing Word Similarity and its Application to Constructing Word Frequency Lists

Xavier Blanco,[1] Mikhail Alexandrov,[1,2] and Alexander Gelbukh [2]

[1] Department of French and Romance Philology, Autonomous University of Barcelona
dyner1950@mail.ru, Xavier.Blanco@uab.es

[2] Center for Computing Research, National Polytechnic Institute (IPN), Mexico
dyner@cic.ipn.mx; www.Gelbukh.com

Abstract. By (morphologically) similar wordforms we understand wordforms (strings) that have the same base meaning (roughly, the same root), such as *sadly* and *sadden*. The task of deciding whether two given strings are similar (in this sense) has numerous applications in text processing, e.g., in information retrieval, for which usually stemming is employed as an intermediate step. Makagonov has suggested a weakly supervised approach for testing word similarity, based on empirical formulae comparing the number of equal and different letters in the two strings. This method gives good results on English, Russian, and a number of Romance languages. However, his approach does not deal well with slight morphological alterations in the stem, such as Spanish *pensar* vs. *pienso*. We propose a simple modification of the method using n-grams instead of letters. We also consider four algorithms for compiling a word frequency list relying on these formulae. Examples from Spanish and English are presented.

1 Introduction

Given a large text or text corpus and a pair of wordforms (strings), we consider the task of guessing whether these two words have the same root and thus the same base meaning. We call this (morphological) *word similarity*: two words are *similar* if they have the same root. This relation permits grouping together the words having the same root, e.g., *sad*, *sadly*, *sadness*, *sadden*, *saddened*, etc. This task has numerous applications, such as constructing word frequency lists. Our motivation is to improve information retrieval and similar practical applications. Consequently, our goal is to provide a reasonably accurate statistical-based algorithm (tolerating certain error rate) and not a precise linguistic analysis.

For grouping together the words with the same root, two morphology-based methods are usually used: lemmatization and stemming. Lemmatization reduces words to the base form: *having* → *have*; stemming truncates words to their stems: *having* → *hav-* (often lemmatization task is also referred to as stemming).

Stemming or lemmatization can be used for testing the (morphological) similarity between two words: both words are first reduced to lemmas or stems; if the resulting strings are equal then the two given words are declared similar. This gives a symmet-

Advances in Natural Language Proccesing
Research in Computing Science 18, 2006, pp. 27-36