Word frequency and spell-checker accuracy

Péter Halácsy[1], András Kornai[2], László Németh[1], András Rung[3], István Szakadát[1], and Viktor Trón[4]

[1] Centre of Media Research and Education, Budapest University of Technology and Economics, {halacsy,szakadat,rung}@mokk.bme.hu
[2] MetaCarta Inc., andras@kornai.com
[3] Center of Cognitive Science, Budapest University of Technology and Economics, rung@itm.bme.hu
[4] School of Informatics, University of Edinburgh, v.tron@ed.ac.uk

The Szószablya (Wordsword) project[5] aims at the creation of open source Hungarian language resources. We have already collected large (over billion words) corpora from the web, and we are in the process of distilling these into more usable word frequency lists and dictionaries. We are also developing a family of low-level analytic tools, including the Hunspell spellchecker, the Hunstem stemmer, and the Hunmorph morphological analyzer, which share the same morphological analysis core and the same base dictionary, currently containing about 80k stems and their morphological subclassification.

This paper focuses on formally characterizing the level of synergy between these two efforts: to what extent can frequency-ordered lists of word-forms be exploited for improving the quality of a stemmer, morphological analyzer, or spellchecker? We concentrate on this last case (since Hunspell is already available on SourceForge) but note here that our analysis carries over without significant changes to the pure stemming/morphological analysis task as well.

First we define the error of a spellchecker given a non-interactive scenario, when the analysis of each word can result in acceptance, rejection (with or without suggesting alternatives), or the spellchecker explicitly noting that the word is outside its scope. Under realistic assumptions about the inherent error rates of the morphological analysis component and the morphological information contained in the stemmer, it turns out that the driving factor of decreasing the error of a spellchecker is the amount we can increase its scope. Next we assume a greedy algorithm, whereby the spellchecker dictionary is gradually increased to include the stems for the first r word forms in frequency order. We use "frequency of frequencies" statistics obtained from some of our larger corpora (670m and 113m words) to demonstrate that Zipf's Law $p_r \sim 1/r^B$ offers a reasonable statistical characterization of Hungarian with $B = 1.25$. Finally, we compute the frequency of word forms left out of scope by a spellchecker based on the first r stems, and conclude that for Hungarian this decreases only with the fourth root of rank which suggests a practical limit to the corpus-sampling technique of boosting stem-dictionary coverage.

[5] managed by the Centre of Media Research and Education of Budapest University of Technology and Economics, supported by Axelero Internet and Ministry of Informatics