A Technique to Improve the Spelling Suggestion Rank in Medical Queries

Jonathan Crowell M.S., Qing T. Zeng, Ph.D., Sandra Kogan M.S.
Decision Systems Group
Brigham and Women's Hospital
Harvard Medical School
Boston, Massachusetts, 02115

Abstract

Correct spelling is crucial for online search engines to function well, and health information is highly sought after online. We propose a technique for increasing the effectiveness of spell-checking tools for use with medical queries. Our results show a marked improvement in the ranking of the correct term within the suggestion list returned by the spelling correction tool, as well as a lessening of the drawbacks associated with using larger dictionaries.

Introduction

Millions of consumers search for health information online, but the medical field contains many words which are difficult to spell, and spelling mistakes can significantly impede the retrieval of relevant results. The goal of this study is to use information about consumer querying patterns to improve the ranking of spelling suggestions.

Method

A list of 1795 correctly-spelled words was obtained from the Medlineplus web site. Several different misspellings for each word were generated by creating the same kinds of mistakes that users often make [1]: dropping a letter, adding a letter, transposing adjacent letters, substituting a letter, and making a minor phonetic alteration.

We used the ASpell [2] spell checker for our study and considered two different dictionary configurations, a medical dictionary, and a comprehensive dictionary formed by merging the medical dictionary with a large English dictionary.

Each word in the dictionaries was assigned a score based on the frequency of its occurrence in log data from the Medlineplus search engine. The following formula was used:

$$FrequencyScore = 1 + \ln(Frequency)$$

The log transformation was used because the data in the query log was skewed with some high outliers. In order to insure that no word received a score of 0, a 1 was added to the log of the frequency. If the word did no appear in the log data at all, then the score was assigned the value 0.5.

A rank was assigned to each word in ASpell's suggestion list according to the formula:

$$NewRank = ASpellRank / FrequencyScore$$

Results

The following table summarizes the results of the spell checking tool both before and after the suggestion list was resorted according to the technique described above.

Rank	Medical Dictionary		Comprehensive Dictionary	
	Original	Resort	Original	Resort
Top 1	73.0%	80.8%	69.8%	80.4%
Top 3	86.5	91.7	83.6	91.4
Top 5	90.6	93.8	88.4	93.6
Top 10	93.9	95.5	93.2	95.5
Total	96.1	96.1	96.0	96.0

These results indicate that having a smaller, well-tailored dictionary initially improves the effectiveness of the spelling suggestion tool. Upon resorting, however, not only is the correct word placed significantly higher in the list, but the difference in effectiveness between the two dictionaries is greatly narrowed. This suggests that the resorting method we have described would not only improve the ranking of the suggestions, but would also enable one to derive all the benefits of using a larger dictionary while foregoing the drawbacks.

Acknowledgements

This work was funded by grant R01 LM07222 from the National Library of Medicine. We thank Dr. Ying Chou Sun for his suggestions and preliminary work.

References

1. Damerau F, Mays E. A technique for computer detection and correction of spelling errors. Communications of the ACM 1964;7(3):171-176.
2. Atkinson K. GNU ASpell version 0.50.3. Released Nov. 23, 2002. http://aspell.sourceforge.net/