A Frequency-based Technique to Improve the Spelling Suggestion Rank in Medical Queries

Article *in* Journal of the American Medical Informatics Association · May 2004

DOI: 10.1197/jamia.M1474 · Source: PubMed

4 authors, including:

Qing Zeng-Treitler
George Washington University
144 PUBLICATIONS **2,772** CITATIONS

SEE PROFILE

Research Paper ■

A Frequency-based Technique to Improve the Spelling Suggestion Rank in Medical Queries

Jonathan Crowell, MS, Qing Zeng, PhD, Long Ngo, PhD, Eve-Marie Lacroix, MS

Abstract **Objective**: There is an abundance of health-related information online, and millions of consumers search for such information. Spell checking is of crucial importance in returning pertinent results, so the authors propose a technique for increasing the effectiveness of spell-checking tools used for health-related information retrieval.

Design: A sample of incorrectly spelled medical terms was submitted to two different spell-checking tools, and the resulting suggestions, derived under two different dictionary configurations, were re-sorted according to how frequently each term appeared in log data from a medical search engine.

Measurements: Univariable analysis was carried out to assess the effect of each factor (spell-checking tool, dictionary type, re-sort, or no re-sort) on the probability of success. The factors that were statistically significant in the univariable analysis were then used in multivariable analysis to evaluate the independent effect of each of the factors.

Results: The re-sorted suggestions proved to be significantly more accurate than the original list returned by the spell-checking tool. The odds of finding the correct suggestion in the number one rank were increased by 63% after re-sorting using the authors' method. This effect was independent of both the dictionary and the spell-checking tools that were used.

Conclusion: Using knowledge about the frequency of a given word's occurrence in the medical domain can significantly improve spelling correction for medical queries.

■ **J Am Med Inform Assoc.** 2004;11:179–185. DOI 10.1197/jamia.M1474.

Millions of people search for health-related information online,[1] but medical terminology is alien to most people and contains many words that are difficult to spell for the average English speaker.[2] General-purpose programs ranging from word processors to search engines now routinely provide spelling correction tools to the user. Consumers performing health information retrieval (HIR) searches, however, are in particular need of good spelling correction tools due to the complex spellings of many medical terms and the high error rate of medical queries (our analysis suggests that as many as 14% of all queries submitted for health information retrieval contain a misspelled term [Table 1]). In response to this need, health information Web sites such as MedlinePlus[3,4] are now equipped with spelling correction functionality. Spelling correction has not yet been perfected, however, and if a term is misspelled, the health information retrieval process will almost certainly fail. There is, therefore, clearly a benefit in improving the spelling correction of HIR queries, as this would increase the efficiency of the HIR process.

Much research has been conducted on spelling correction in the computer science domain.[5] Within medical informatics, however, relatively few groups have worked on spelling correction. Worth noting is the work on GSpell, an open-source spell-checking Java application–programmer interface (API) being built by the National Library of Medicine.[6,7] There also has been little reporting on how well current spelling correction tools perform in the context of HIR.

As part of our work on consumer HIR, we have developed a new frequency-based method for improving the rank of the correct terms suggested by spelling-correction tools. For testing, two spelling-correction tools, ASpell and GSpell, were used. Like GSpell, ASpell is open-source and freely available. In addition, ASpell is widely used in the Linux community and self-reports that "it does a much better job of coming up with possible suggestions than just about any other spell checker out there for the English language, including Ispell and Microsoft Word."[8] In testing we also used two different dictionary configurations: a medical dictionary and a much larger comprehensive dictionary.

In this study, we first analyzed the baseline performance of both ASpell and GSpell for spelling correction of consumer health information queries and then compared it with the performance after applying the new ranking method. A one-page abstract that contained early findings from this work appeared in the 2003 AMIA proceedings.[9]

Affiliations of the authors: Decision Systems Group, Brigham & Women's Hospital, Harvard Medical School, Boston, MA (JC, QZ); Department of Biostatistics, Harvard School of Public Health, Boston, MA (LN); Public Services Division, National Library of Medicine, Bethesda, MD (E-ML).

A one-page abstract that contained early findings from this work appeared in the 2003 AMIA Proceedings.

Supported in part by grant R01 LM07222 from the National Library of Medicine.

Correspondence and reprints: Jonathan Crowell, MS, Decision Systems Group, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115; e-mail: <jcrowell@dsg.bwh.harvard.edu>.

Received for publication: 10/28/03; accepted for publication: 12/21/03.

Table 1 ■ Statistics from Log Data of Queries Submitted to the MedlinePlus Search Engine

Query Log Data from MedlinePlus	
Total number of queries	2,179,181
Number of unique queries	629,915
Average query length	1.9 words
Average word length	7.0 characters
Misspelled words	4.9%*–7.6%†
Queries containing a misspelling	9.1%*–13.9%†
Average length of misspelled words	8.1†–8.8*

*Checked against the comprehensive dictionary.
†Checked against the medical dictionary.

Background

Spelling Correction

Computer-aided spelling correction has been the subject of investigation since the 1960s.[10] Research has focused on several different areas, from pattern matching algorithms and dictionary searching techniques to optical character recognition (OCR), used in the automated scanning of printed text, and context-dependent text analysis that seeks to identify and correct even real-word spelling mistakes (such as the error in "striving for world piece").

The emphasis of the research has generally been on the English language taken as a whole, which results in a very broad context in which to conduct spelling correction. In contrast, this study focused only on words submitted to search engines attached to repositories of medical information. This limitation of the domain allowed us to exploit trends in search patterns to markedly improve the efficacy of the spelling correction tools.

Frequency-based Approaches

The idea of considering the likelihood of a suggestion based on the frequency of its occurrence in language is not completely original with us. Kernighan, Church, and Gale have previously used frequency analysis to rank a list of suggestions.[11,12] Their scoring method, which is somewhat different from ours, is described as follows:

> Each candidate correction, c, is scored by $\Pr(c)\ \Pr(t|c)$, and then normalized by the sum of the scores for all proposed candidates. The prior, $\Pr(c)$, is estimated by $(freq(c) + 1) / N$, where $freq(c)$ is the number of times that the word c appears in the 1988 AP corpus (N = 44 million words).[11]

This analysis makes use of both the absolute likelihood of a word's occurrence in the corpus, $\Pr(c)$, and the likelihood of the typographical error given the candidate, $\Pr(t|c)$, which they calculate using confusion matrices created from a training set based on AP corpus (which contains typos). This method assumes that all the suggestions are initially equally likely to be correct, and they are then scored according to the method just described. Also, in their analysis, they differentiate only between words that are one edit distance away from the incorrect word.

Our method re-sorts suggestions that are often more than one edit distance away and that already have been sorted with sophisticated algorithms. The goal of this study is to improve upon any existing ranking method using word-frequency knowledge, not to establish a new ranking system from scratch, so our method differs from the method used by Kernighan, Church, and Gale. Due to the nature of our study, we are also able to quantify the benefit of considering word frequency, which has not previously been reported.

MedlinePlus and Query Log Data

A flagship project of the National Library of Medicine is the MedlinePlus Web site, located at www.MedlinePlus.gov. This Web site is updated daily, contains information on more than 650 medical conditions, and is intended to be both authoritative and usable by the layperson. The site relies on no commercial funding and receives more than 20 million page views monthly.[13]

This study is specifically concerned with spelling errors in queries submitted for medical information retrieval, so a corpus of words from log data of queries submitted to the MedlinePlus Web site was used to generate word-frequency statistics. The log data consisted of all queries submitted to the MedlinePlus search engine between August 1 and November 30, 2001. During that time 2,179,181 queries were submitted consisting of 4,249,606 words. Table 1 summarizes the log data set.

Hypothesis

The underlying hypothesis of the study is that rates of occurrence of words used for health information retrieval are distinctive enough that they can be exploited to significantly improve the suggestion list returned by a spell-checking tool. Thus, by measuring the likelihood that a word will be submitted to a medical search engine, the ranking of the correct term in a suggestion list returned by a spell-checking tool for a medical search engine can be improved.

Methods

Obtaining Misspelled Words

We chose to automatically generate misspellings because our initial attempt at using actual spelling mistakes found in the query log data proved unsatisfactory. Human perusal of the log data tends to be biased in favor of easily recognizable spelling mistakes, and for other mistakes it is often impossible to ascertain what the intended word is.

To generate misspellings, an initial list of correctly spelled words is required. We obtained this list by intersecting the words in the log data of MedlinePlus queries with the words contained in a searchable index of MedlinePlus generated by a Lucene search engine (Lucene is a high-quality open-source search engine API developed by the Apache Jakarta project[14,15]). Stop words ("the," "and," etc.) and short words (≤3 letters) were removed from the list. The reason for excluding short words is explained in the Discussion section.

The intersection of MedlinePlus terms and MedlinePlus queries was used as an alternative to a medical dictionary to identify correctly spelled words for four reasons:

1. We assumed that MedlinePlus, being a respectable Web site, generally contains correctly spelled words. Moreover, it is unlikely that both MedlinePlus and an arbitrary user would make the same spelling mistake.
2. Words in the queries were used to search the content of MedlinePlus, so from a practical point of view the

searchable index of MedlinePlus words defines what it means for a word to be correctly spelled for the restricted domain of this study.
3. By not relying on an actual dictionary, the list avoids being biased toward a particular dictionary being used by the spell-checking tool.
4. By including in our test list only words that appear in the query log, we limit ourselves to words that are known to have been used for consumer HIR.

The list of correct words for our study was selected randomly by iterating through the full list of words from the intersection described above, generating a pseudo-random number between 0 and 1 for each one, and selecting only those words with a score below 0.1. This method was used to make the list size more manageable while avoiding any selection bias.

Next, misspelled words were generated according to how typical spelling mistakes are made. Damerau and Mays[10] have shown that 80% of all spelling mistakes involve a single insertion, a single deletion, a single substitution, or a transposition of two letters in a word. The automatic error generator for this study created mistakes according to these four rules and also generated mistakes by applying the rules twice and creating phonetic mistakes if possible. In applying the rules twice, two of the single-mistake transformations were randomly chosen, possibly resulting in the same transformation being applied twice (e.g., two letters dropped). Phonetic mistakes were generated according to rules derived from Lawrence Philips' double metaphone algorithm.[16] The list below enumerates the different methods that were used and gives the result obtained upon application of each method to the word "hallucinating":

1. Dropping a letter → hallucinaing
2. Adding a letter → hkallucinating
3. Transposing adjacent letters → halluicnating
4. Replacing a letter → halmucinating
5. Performing two of 1–4 → hallucinanig
6. Phonetic alteration (if possible) → halusinating

Configuring the Dictionaries

Dictionary configurations have a significant impact on spelling correction.[5,17–19] For this study, two different dictionary configurations were used to test the hypothesis independently of the dictionary used. The two dictionaries were:

1. A medical dictionary created from the 2003 Specialist Lexicon[20] provided with the UMLS knowledge base that is maintained by the National Library of Medicine. This dictionary "is intended to be a general English lexicon that includes many biomedical terms."[21]
2. A comprehensive dictionary created by combining: (a) the medical dictionary from 1; (b) terms selected from the mrxns.eng file provided with the UMLS knowledge base[22] (this file contains English normalized strings from all the entries in the UMLS meta-thesaurus—strings containing digits were not considered); and (c) a large English dictionary created using the SCOWL (Spell Checker Oriented Word Lists) package.[23] Note that the comprehensive dictionary is a superset of the medical dictionary.

Configuring the Spell-checking Tools

ASpell and GSpell both return a suggestion list of correct words for a misspelling. They also both have configuration parameters that affect their performance. ASpell was run in its default configuration, which includes the normal mode (as opposed to ultra mode, fast mode, or bad-spellers mode). In its default configuration, ASpell will sometimes return up to 100 suggestions for a given misspelling. More precise tailoring of ASpell's behavior is not possible without rewriting the source code. GSpell was configured to consider 75,000 candidates, return up to 100 suggestions, and consider candidates up to an edit distance of five. These are more lenient settings than GSpell's default, which is 3,000 candidates for consideration, a maximum of ten suggestions returned, and a maximum edit distance of four.

Assigning Frequency Scores to Words

To re-sort the list of suggestions returned by GSpell and ASpell, a quantification of the likelihood that a given word will be used is needed. The following formula was used to obtain a frequency score for each word:

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

Some words are far more frequently used than others. For instance, the MedlinePlus log data contained a few dozen outliers such as "syndrome" and "disease" that appeared several thousand times each, whereas the majority of words appeared less than 50 times each. This skewing of the data lends itself to the use of the log transformation.

A 1 was added to the log of the frequency to ensure that a word appearing only one time would not receive a score of 0. (Adding 1 in this scenario is a common practice, although Gale and Church have proposed an alternative to it.[24]) If the word did not appear in the log data at all, then the score was assigned the value 0.5.

The frequency score used here is based on the frequency of a word's occurrence in the log data of queries submitted to the MedlinePlus Web site. The resulting frequency scores ranged from 0.5 for absent words to 11.549 for the word "disease," which appeared 38,133 times.

Re-sorting the Suggestion Lists

For each word in the suggestion list, we computed a ranking score based on the word's original score given by the spell-checking tool and the word's frequency score as computed from the log data. Each word was assigned a re-sort score according to the formula:

$$ReSortScore = \frac{SpellScore^C}{FrequencyScore}$$

where C is a constant used to regulate the strength of the original score. Empirically, we used $C = 3$ for this study. A higher C will lower the impact of the frequency, and a lower C will magnify the impact of the frequency.

ASpell assigns each word a score between about 30 for a very closely matching word and about 600 for a very long shot. GSpell assigns each word a score between 0 and 1, with scores nearer to 1 being closer matches. ASpell's score was plugged directly into the formula, but because GSpell scored at

a different order of magnitude than ASpell and also scored close matches with a high number as opposed to a low number, GSpell's score was first transformed according to the formula:

$$SpellScore = (1 - GSpellScore) \times 100$$

The adjusted scores from the two spell-checking tools were not equivalent, but this does not affect the outcome (because cross-comparisons are not made or needed). This adjustment was made purely for programming convenience, for if the GSpell score is not transformed, then a different C would be needed in the formula when dealing with GSpell. The list of suggestions was then re-sorted according to the re-sort score. Words not appearing in the log data are assigned a frequency score of 0.5, thereby disadvantaging them as suggestions.

Evaluation

First, both ASpell and GSpell were benchmarked by testing them on the list of misspelled words with each of the two different dictionaries. The results generated by ASpell and GSpell then were re-sorted based on the frequencies of the words in the log data and the formulas given above.

Three different types of outcomes were measured: (1) whether the correct word was ranked number one (first outcome), (2) whether the correct word was ranked in the top ten (second outcome), and (3) whether the correct word was found at all (third outcome). Each of these is a binary variable. The observation's response (each row of the dataset is a misspelled word and is called an observation) was treated as independent because the response variable (rank number) is generated by the search method without taking into account the information from the other observations' ranks.

There are eight different possible search combinations resulting from the three different parameters in the study: each search method uses either the GSpell tool or the ASpell tool, is configured with either the medical dictionary or the comprehensive dictionary, and either does or does not employ re-sorting.

There are some observations that have the misspelled word identified as a correct word because that particular arrangement of letters appears in the given dictionary. These results are known as real-word errors, and were not considered in our analysis. Note that this issue is not a deep-seated problem because the spell-checking tools are displaying ideal behavior in assessing words that appear in the dictionary as being correctly spelled. For the sake of uniform statistical analysis, however, an observation was dropped if either spell-checking tool identified it as correct under either of the two dictionary configurations. (Although, because the comprehensive dictionary is a superset of the medical dictionary, any word mistakenly identified as correct by the medical dictionary suffers the same fate at the hands of the comprehensive dictionary). Of the 12,304 observations (misspelled words), 715 (5.8%) were eliminated due to this issue, 328 of which were contributed by only the comprehensive dictionary configuration. This left 11,589 independent observations for analysis.

The statistical significance level was set at 5%. All statistical tests were evaluated as two-tailed tests. Probability of success was defined as having a successful outcome (i.e., the correct word was found in the number one slot for the first outcome, the correct word was found in the top ten slots for the second outcome, and the correct word was found at all for the third outcome).

Chi-square tests were used to evaluate whether there is a significant difference overall in the probabilities of success for each of the three outcomes of interest from the eight different combinations of the search method. Univariable analysis was carried out to assess the unadjusted effect of each factor (spell checker, dictionary type, re-sort, or no re-sort) on the probability of success. The possible confounding effect of the length of the word was also investigated with the use of an unpaired t-test. The factors that were statistically significant in the univariable analysis were then used in the multivariable analysis to evaluate the independent effect of each of the factors of the search method.

Three multivariable logistic regressions[25] were set up to look at each of the three outcomes. The length of the word was included in all the models to adjust for the effects of the search factors (i.e., spell-checking tool, dictionary configuration, and whether re-sorting is applied). The odds ratios and 95% confidence intervals were computed.

Results

Obtaining Misspelled Words

The list of correctly spelled search terms generated by intersecting the MedlinePlus terms with the log data contained 25,922 distinct words from which 2,375 words were (pseudo) randomly chosen to form the list of words from which misspellings would be generated. The resulting list contains a range of words that were used in a HIR context (e.g., "borborygmus," "daughter," "fundal," "path," and "phosphoribosyltransferase"), but many of the words, as one might expect, are not exclusively medical. The list of 2,375 correctly spelled words yielded 12,304 misspellings (phonetic transformations were not possible in every case).

Configuring the Dictionaries

The spell-checking tools were run using two different dictionary configurations: The medical dictionary obtained from the Specialist Lexicon contained 191,474 total distinct words, including the 10,000 most common words in English.

The comprehensive dictionary was formed using (a) the medical dictionary, containing 191,474 words; (b) the list selected from the UMLS mrxns.eng file, containing 364,742 unique strings; and (c) the large English word list from the SCOWL package, containing 100,057 distinct words. The resulting comprehensive dictionary contained 509,665 words (there was some overlap among the three lists).

Statistical Analysis Results

The three outcomes we measured were: (1) whether the correct suggestion was ranked number one; (2) whether the correct suggestion was ranked in the top ten; and (3) whether the correct suggestion was found at all. Table 2 shows the probability of success among the eight different search combinations for each of the three outcomes. Chi-square tests with $p < 0.0001$ indicate that there was an overall statistically significant difference among the search methods. The reported percentages also indicate an informative order showing that re-sorting improves the probability of success. The ASpell/Comp/Re-sort combination has the highest percentages, 76.2%, 91.2%, and 92.3%, for the three outcomes,

Table 2 ■ Percent of Misspelled Words (*N* = 11,589) for Which the Correct Word Was Found

Search Method*	% with Correct Word Ranked #1	% with Correct Word Ranked #1–10	% with Correct Word Found at All
Aspell/Med/No Re-sort	64.1	82.8	84.9
Aspell/Med/ Re-sort	70.8	83.9	84.9
Aspell/Comp/No Re-sort	59.5	85.9	92.3
Aspell/Comp/ Re-sort	76.2	91.2	92.3
Gspell/Med/No Re-sort	53.7	69.8	73.3
Gspell/Med/ Re-sort	59.4	70.8	73.3
Gspell/Comp/No Re-sort	53.7	73.7	79.4
Gspell/Comp/ Re-sort	64.3	77.2	79.4

*Chi-square test shows dependence between the search method and each of the three binary outcomes (p-value < 0.0001).

with ASpell/Med/Re-sort coming in second best. The impact of re-sorting appears to be highest for the first outcome, less for the second outcome, and ineffective, of course, for the third outcome.

Table 3 shows the unadjusted univariable analysis for each of the search method's factors. ASpell appears to dominate GSpell for all three outcomes. The comprehensive dictionary appears to be superior to the medical dictionary especially for the second and third outcomes. For the first outcome, the superiority of the comprehensive dictionary is slight and is also affected by the dropped observations, many of which were contributed by only the comprehensive dictionary. Re-sorting shows a clear benefit for the first outcome (about a 10-percentage-point difference), less benefit (2% difference) for the second outcome, and none for the third outcome (with p-value of 1.000).

The length of the word is taken into account as a possible confounder in later analysis. The results indicate a statistically significant difference in the length of the word between the

Table 3 ■ Univariable Analysis for Each Factor of the Search Method

Search Factor	% with Correct Suggestion Ranked #1	% with Correct Suggestion Ranked #1–10	% with Correct Suggestion Anywhere in List
Spell-checking tool*			
ASpell	67.7	85.9	88.6
GSpell	57.8	72.9	76.3
Dictionary configuration*			
Comprehensive	63.4	82.0	85.8
Medical	62.0	76.8	79.1
Re-sorting*			(p-value = 1.000)
Yes	67.7	80.8	82.5
No	57.8	78.0	82.5

*p-value from chi-square and t-test < 0.0001.

Table 4 ■ Univariable Analysis of the Length of the Correct Word as a Factor in Searching

	Mean Word Length* of Correctly Spelled Word		
	Ranked #1	Ranked #1–10	Anywhere in List
Correct suggestion found	9.33	8.90	8.80
Correct suggestion not found	7.34	7.41	7.71

*p-value from chi-square and t-test < 0.0001.

two groups (success and failure of the outcome). It appears that the longer the word, the higher the probability of finding the correct suggestion ranked number one, ranked in the top ten, or ranked at all. Table 4 summarizes the effect of the word length on spelling correction. For each of the three outcomes, the average length of words for which the correct suggestion was found exceeded the average length of words for which the correct suggestion was not found.

Table 5 shows the results of the multivariable logistic regression in which four parameters were entered into the model as independent variables. We are interested in estimating the adjusted or independent effect of the three factors of the search method controlling for the confounder (length of the word). For the first outcome, if ASpell is used in the search instead of GSpell, then the odds of finding the correct word in the number one rank is increased by 62% (with a 95% confidence interval of between 57% and 67%), a substantial difference. The difference between the comprehensive and medical dictionaries is smaller, with an odds difference of 7% (which, as noted above, has been mildly skewed in the comprehensive dictionary's favor by the pattern of dropped observations).

The difference between re-sorting and not re-sorting is large: re-sorting increases the odds of finding the correct word in the number one rank by 63% over not re-sorting. The effect of re-sorting is reduced (20% difference in odds) but is still significant for the second outcome. For the third outcome, there is no difference (odds ratio of 1) between re-sorting and not re-sorting, as expected.

Of interest is the independent effect of the length of the word. For all three outcomes, the longer the length of the word, the higher the probability of having a successful outcome. Notice also for the second and third outcome, the effect of the spell checking tool and the dictionary type increases substantially.

Table 5 ■ Multivariable Logistic Regression Analysis for Each Factor of the Search Method*

	% with Correct Word Rank #1	% with Correct Word Rank #1–10	% with Correct Word Found
ASpell versus GSpell	1.62 (1.58–1.67)	2.38 (2.29–2.46)	2.48 (2.39–2.57)
Comprehensive versus medical	1.07 (1.04–1.10)	1.41 (1.37–1.46)	1.64 (1.58–1.69)
Re-sort versus No Re-sort	1.63 (1.58–1.68)	1.20 (1.16–1.24)	1.00 (0.97–1.04)
Per letter increase in word length	1.34 (1.33–1.35)	1.23 (1.22–1.24)	1.15 (1.14–1.16)

*Odds ratio and 95% confidence interval from the statistical model.

Discussion

Significance

Our results show a clear improvement in spelling correction for health information retrieval by applying frequency-based re-sorting. The improvement is statistically significant and results in markedly improved suggestion lists for misspelled words.

Our analysis showed that ASpell outperformed GSpell, which is not surprising since the GSpell developers acknowledge a weakness in GSpell's algorithm on its Web site:

> There is a consequence to the bathtub heuristic. If the spelling error occurs in the first or last character of the string, the suggestions, if any, will more than likely be wrong. In a future release, two additional techniques may be employed to ameliorate this: a left and right truncation retrieval and a conservative stemming heuristic.[7]

Since our automatic spelling mistake generator might have produced more first- and last-character mistakes than would be expected in human errors, this weakness in GSpell's algorithm may have been exaggerated by this study. Our study should not, therefore, be viewed as a conclusive comparison between GSpell and ASpell.

Implication

The clear implication of our research is that in the specific domain of health information retrieval, spelling correction tools can be improved by considering the frequency score of each word in the suggestion list. Our method may well be transferable to other domains for which good word usage statistics are available.

The results also indicate that having an appropriate dictionary improves the effectiveness of the spelling suggestion tool. While the multivariable logistic regression analysis indicates that the comprehensive dictionary provides a mild advantage of 7% over the medical dictionary for the first outcome, this result is somewhat skewed due to the fact that the comprehensive dictionary contributed many more dropped observations than the medical dictionary did. Prior to the observations being dropped for the sake of uniform statistical analysis, the medical dictionary had placed more correct suggestions in the number one slot than the comprehensive dictionary had.

With the help of the re-sort algorithm, however, our results suggest that using a larger dictionary is preferable to using a smaller and more precisely tailored one. A larger dictionary is able to find more words, but at the expense of a larger number of overlooked misspellings (due to a coincidental match with a word in the dictionary), and a significant degradation in the ranking of the correct suggestion. After re-sorting the suggestion list in our study, however, the comprehensive dictionary yielded notably better results than the list from the re-sorted medical dictionary. The larger number of overlooked misspellings is not addressed by the re-sorting method, but this deficiency is more than offset by a decrease in the number of words for which the correct suggestion is not found at all, and a marked improvement in the ranking of those words that are found.

Our method is easily tuned for any body of text so long as it is copious enough to provide good word-frequency counts. The spelling correction tool can be trained first by processing the text and counting the frequency of each correctly spelled word using an appropriate dictionary. The text then can be reprocessed immediately, this time correcting the misspelled words with the aid of the frequency scores learned from the first pass through the text. Using this method on a corpus of 10,000 surgical pathology reports, we were able to improve the suggestions for many misspelled terms. The word "biopsy," for instance, was given a boost within that context, whereas alternative suggestions such as "bipod," "bios," and "boobs" were appropriately downplayed.

Limitations

One of the drawbacks in our study was that the automatic spelling-mistake generator occasionally created mistakes that a human typist would be unlikely to make. For instance, our mistake generator produced "tngivoma" as a misspelling of "angioma." While it is certainly possible that a human being could misspell "angioma" in that manner, it seems unlikely. A more sophisticated mistake generator would take into account such factors as the keyboard layout, which tends to promote certain typos, such as the replacement of "a" with "s," over others, such as the replacement of "a" with "m." Also, it is known that spelling mistakes occur less frequently in the first or last letters of a word,[5,26] whereas our mistake generator treated all positions in a word equally.

Ideally, of course, actual spelling mistakes would be used instead of automatically generated mistakes. The difficulty in that method, particularly in the health domain, is determining which word was intended and avoiding bias toward easily recognizable mistakes and words. This difficulty is surprisingly strong even for native English-speaking physicians. This problem is particularly acute in analysis of Web queries, which are short and do not provide much context. Other data sets containing large numbers of consumer-generated and health-related spelling errors, however, are not easy to acquire.

In calculating the re-sort score, an empirically assigned constant C is used to regulate the impact of the word frequency on the new ranking. The constant C should be adjusted for the particular scoring methods used by a spelling correction tool. Since we have limited knowledge of the intricacies of ASpell and GSpell's scoring algorithms, our calculations likely did not take full advantage of the nuanced information in the original scores. Presumably a much finer formula could be designed for use in conjunction with the specific algorithms employed by those tools. However, this study is not intended to develop a specific method for a particular software package. Rather, we set out to show the general (and intuitive) benefit of taking word frequency into account when performing spelling correction—especially in a constrained domain such as health care. Toward that end, the significant improvements resulting from a simple and rough re-sort formula highlight the potency of employing word frequency.

Another limitation of our study is that we generated mistakes exclusively for words at least four characters in length. Our results do not have much to say, therefore, about very short words. There were three reasons we eliminated very short words in our study:

1. Short words are less likely to be misspelled. In a study of spelling error patterns Yannakoudakis and Fawthorp found an error rate of only 1.5% in short words,[27] and a study by Pollock and Zamora found a 9.2% error rate in words of three or four letters.[26] The list in this study only eliminated words up to three characters. As noted in Table 1, the average word length in the log data was seven characters.
2. Medical terms contain an abundance of acronyms, and many acronyms are very short, such as "aaa" for abdominal aortic aneurysm, "blv" for bovine leucosis virus, and "cjd" for Creutzfeldt-Jakob disease. In fact, MedlinePlus contains 1,587 three-letter words that are entirely alphabetic (hundreds more include digits). Although this accounts for slightly less than 10% of all possible permutations, it is not a complete list of acronyms, and new acronyms are constantly being created. This great number of acronyms makes it ineffective to identify misspellings and make suggestions for very short words.
3. Short words present a greater challenge to the spelling correction tool. Kukich notes that "short words are more difficult to correct, in part because less intraword contextual information is available to the corrector."[5] In addition, Kukich cites a study by Landauer and Streeter that showed that short words have more nearby neighbors in terms of editing distance, which results in a clouded spelling suggestion list.[5,28]

Put plainly, context-independent spell checking of three-letter (and shorter) words in the context of medical information retrieval queries is both difficult, due to the unusually large number of short words in close proximity to each other, and relatively unimportant, due to the rare occurrence of misspellings among short words. Nevertheless, the elimination of short words from our study should be noted.

Future Work

Further testing of our method would be appropriate in light of these limitations, perhaps focusing on the impact of word length on spelling correction. The method would also benefit from being tested in different domains or with a list of actual, human-generated misspellings. In addition, several areas could prove fertile ground for further research. Phrase-based spelling correction would very likely further improve the results. For example, the existence of the term "myocardial" in the query "myocardial imfraction" should promote the term "infarction" over "infraction" in the suggestion list for the misspelled word in spite of the fact that "infraction" is closer in terms of edit distance. Although queries tend to have only two or three words of context, the words are often significant and highly related terms.

References ■

1. Fox S, Raine L, Horrigan J, et al. The Online Health Care Revolution: How the Web Helps Americans Take Better Care of Themselves. Washington, DC: The Pew Internet & American Life Project, 2000.
2. Zeng Q, Kogan S, Ash N, Greenes R, Boxwala A. Characteristics of consumer terminology for health information retrieval. Methods Inf Med. 2002;41:289–98.
3. MedlinePlus. Available at: www.medlineplus.gov. Accessed Oct 28, 2003.
4. Miller N, Lacroix EM, Backus JE. MedlinePlus: building and maintaining the National Library of Medicine's consumer health Web service. Bull Med Libr Assoc. 2000;88:11–7.
5. Kukich K. Techniques for automatically correcting words in text. ACM Comput Surv. 1992;24:377–439.
6. Divita G, Browne A, Tse T, Cheh M, Loane R, Abramson M. A spelling suggestion technique for terminology servers. In: Proc AMIA. 2000:993.
7. Divita G. GSpell. Version 0.27. Produced by National Library of Medicine. Available at http://umlslex.nlm.nih.gov/nlsRepository/gspell/doc/userDoc/. Accessed February 2004.
8. Atkinson K. GNU ASpell. Version 0.50.3. Produced by SourceForge.Net. Available at http://aspell.sourceforge.net/. Accessed February 2004.
9. Crowell J, Zeng Q, Kogan S. A technique to improve the spelling suggestion rank in medical queries. In: Proc AMIA. 2003:823.
10. Damerau F, Mays E. A technique for computer detection and correction of spelling errors. Commun ACM. 1964;7:171–6.
11. Kernighan M, Church K, Gale W. A spelling correction program based on a noisy channel model. In: Karlgren H (ed). Proceedings of COLLING-90, The 13th International Conference on Computational Linguistics, 1990, Helsinki, Finland, pp 205–10.
12. Church K, Gale W. Probability scoring for spelling correction. Stat Comput. 1991;1:93–103.
13. MedlinePlus Web Usage Fiscal Year 1999 to Present. Available at: http://www.nlm.nih.gov/medlineplus/usestatistics.html. Accessed February 2004.
14. Cutting D. Lucene. Version 1.3. Produced by the Apache Jakarta project. Available at: http://jakarta.apache.org/lucene/. Accessed February 2004.
15. Walls C. Search-enable your application with Lucene. Java Developers J. 2002;7:16–24.
16. Philips L. Hanging on the metaphone. Comput Language. 1990; 7(12):39–43.
17. Peterson J. A note on undetected typing errors. Commun ACM. 1986;29:633–7.
18. Damerau FJ, Mays E. An examination of undetected typing errors. Inf Process Manage. 1989;25:659–64.
19. Walker DE, Amsler RA. The use of machine-readable dictionaries in sublanguage analysis. In: Grishman R, Kittredge R (eds). Analyzing Language in Restricted Domains: Sublanguage Description and Processing. Hillsdale, NJ: Lawrence Erlbaum; 1986:69–83.
20. National Library of Medicine. The specialist lexicon: Description. Available at: http://specialist.nlm.nih.gov/LexiconDescription.html. Accessed Jan 8, 2004.
21. National Library of Medicine. Specialist lexicon fact sheet. Available at: http://www.nlm.nih.gov/pubs/factsheets/umlslex.html. Accessed Jan 8, 2004.
22. National Library of Medicine. 2003AB UMLS documentation. Available at: http://www.nlm.nih.gov/research/umls/archive/2003AB/UMLSDOC.html. Accessed Jan 8, 2004.
23. Atkinson K. SCOWL. Version 5. Produced by SourceForge.Net. Available at: http://wordlist.sourceforge.net/. Accessed February 2004.
24. Gale W, Church K. What's wrong with adding one? In: Oostdijk N, De Haan P (eds). Corpus-based Research into Language. Amsterdam, The Netherlands: Rodopi Bv Editions, 1994.
25. Hosmer D, Lemeshow S. Applied Logistic Regression. New York: John Wiley & Sons, 1989.
26. Pollock JJ, Zamora A. Collection and characterization of spelling errors in scientific and scholarly text. J Am Soc Inf Sci. 1984;34:51–8.
27. Yannakoudakis EJ, Fawthrop D. The rules of spelling errors. Inf Process Manage. 1983;19:87–99.
28. Landauer TK, Streeter LA. Structural differences between common and rare words. J Verbal Learn Verbal Behav. 1973; 12:119–31.