Non-Word Identification or Spell Checking Without a Dictionary

Donald C. Comeau and W. John Wilbur
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room N611S, 8600 Rockville Pike, Bethesda, MD 20894.
E-mail: comeau@ncbi.nlm.nih.gov

MEDLINE® is a collection of more than 12 million references and abstracts covering recent life science literature. With its continued growth and cutting-edge terminology, spell-checking with a traditional lexicon based approach requires significant additional manual follow-up. In this work, an internal corpus based context quality rating α, frequency, and simple misspelling transformations are used to rank words from most likely to be misspellings to least likely. Eleven-point average precisions of 0.891 have been achieved within a class of 42,340 all alphabetic words having an α score less than 10. Our models predict that 16,274 or 38% of these words are misspellings. Based on test data, this result has a recall of 79% and a precision of 86%. In other words, spell checking can be done by statistics instead of with a dictionary. As an application we examine the time history of low α words in MEDLINE® titles and abstracts.

Introduction

Spell checking is a routine feature of modern word processing programs. Before this, stand-alone programs were used to verify the spelling of words in a document. The basic algorithms are fairly simple and involve the use of a dictionary or list of accepted words. The challenge is to store this dictionary in the least amount of space and to access it in the least amount of time. Often, optimizations to this process avoid explicitly constructing the entire dictionary. An example would be storing a root word and indications of legal prefixes and suffixes instead of explicitly listing the root word and its variants. Nonetheless, the dictionary is implicitly present. The suffix handling is simply an optimization to reduce space requirements (McIlroy, 1982). Bentley (1985) provides an easily accessible introduction to the challenges of spell checking and creating useful word lists. An extensive review of more recent developments is given by Kukich (1992).

MEDLINE® is a large corpus of life science research abstracts. The vocabulary is large, technical, and continually expanding. Maintaining a dictionary with all properly spelled words would not be realistic. Even if a dictionary was developed covering all references through a given time period, the advance of research and the creativity of the researchers would continue to expand the vocabulary. Given this changing environment, it is challenging to detect misspelled words. Ideally, one would detect misspelled words using internal features and characteristics of the corpus and this has been our goal.

Our particular interest is in context and how it can be used to detect spelling errors. Context has been used for spelling correction, to indicate which of several possible correct words the author most likely intended. Bowden (1995) used a lexicon, single-error misspellings, tag checking, and partial parsing of subsentential context to interactively suggest corrections while text is entered. Golding and Schabes (1996) used part of speech trigrams and a Bayesian hybrid classifier, with context of up to 10 words on either side, for correcting valid word errors. Elmi and Evens (1998) use the part of speech, syntax analysis, and semantic analysis of the immediate sentence to eliminate impossible words from a list of putative spelling corrections. Nylander (1999) looked for OCR errors without a lexicon. Character trigrams and phonotax were used to identify likely misspellings. A significant amount of text is needed to generate rules that are general enough to be useful (Nylander, 1999). Our approach differs from these methods in that we make use of the larger context of a word and do not use any form of syntactic analysis or limit our interest to the immediately adjacent words.

Our method is based on the measure of a word's context developed by Kim and Wilbur (2001) and therein referred to as strength of context. For convenience, we refer to this measure as α. The measure a is α nonnegative real number the size of which reflects how strongly the word associates

Received December 1, 2002; revised June 27, 2003; accepted June 27, 2003

with the other words appearing in the documents in which it occurs. Observation of low α words revealed that many of them are misspellings. There are simple reasons that misspellings have low α scores. Because a misspelling occurs less frequently than the correct variant, it does not have the same opportunity to build a consistent, reliable context. Since mistakes are infrequent and haphazard, their occurrences appear random. Even worse, the words that would provide the context of the misspelled word appear more frequently with the correct word, not the misspelled one. All of these factors contribute to the low α score of the misspelled word.

As one might expect, simply a low value of α is not sufficient to guarantee that a word is a misspelling. There are several reasons a word may have a low α. Words may be used and spelled correctly; but their use is sporadic, and their meanings are only indirectly related to their context. Examples include "quincunx," "tantalized," and "vilification." Names can refer to different people or things. The same abbreviation may have wildly different expansions. As an example, "npba" can mean "N-phenyl-beta-alanine" or "nonpuerperal breast abscess." Because α is not alone sufficient, we must rely on other information as well. Once α scores and frequency counts are obtained for each word, we create a list of alternatives. These alternatives are all the other words that differ from the word in question by one of the single-error misspellings observed by Damerau (1964): deletion, insertion, substitution, and transposition. Using the α scores and frequency counts of both the potential misspelling and its possible correct alternatives as features, a classifier can be trained to predict which words are most likely to be misspellings.

With a list of probably misspelled words, we looked at the time history of misspellings in MEDLINE®. We found a rise in misspellings when abstracts where first included in MEDLINE® references. We also saw a decrease in misspelled words as automatic spell checking became available both to authors and the Library of Medicine.

Method Overview

We began with a group of about 40,000 low α words from MEDLINE®. Human judges looked at 2,000 words randomly selected from this set and indicated which words were misspellings and which were correct. Half of these words were used to train the classifiers and the rest were held back for testing.

Several features were identified as possible indications of a misspelling. These included the number of times the word occurs in MEDLINE®, the strength of the word's context, a ratio of the frequency in MEDLINE® compared to a *Wall Street Journal* (*WSJ*) corpus, and the frequency of the word's letter trigrams in MEDLINE®. We also looked at the number of words that differed by only one single-error and the same list of features for those words. Details regarding these features appear in the Discussion.

TABLE 1. Categorization methods and 11-point average precisions.

Categorization method	All features	No *WSJ* features
Mahalanobis	0.866	0.847
LogLinear	0.885	0.873
CMLS	0.891	0.877
Boosting	0.850	0.836

Among the humanly judged set, each word with at least one occurrence marked "Simple Misspelling" was flagged as a misspelling. The other words were considered correct. Several categorization methods were applied to the half of the judged words used as the training set. The methods used the words from the training set and their feature values to produce a model that describes which words are misspelled and which are correct. This model is then used to rate the words in the test set, the remaining half of the judged words. The words in the test set are ordered from the most likely to be a misspelling to the least likely to be a misspelling.

Once the classifiers ranked the words from most likely to be a misspelling to most likely to be correct, several measures were used to evaluate the results. *Recall* is the proportion of misspellings identified. *Precision* is the proportion of actual misspellings among the words claimed to be misspellings. *F-score* combines precision and recall into a single meaningful number. *Eleven-point average precision (11-pap)* measures the quality of the ranking over a range of recall values.

More details on these evaluation measures and the other methods used are in Appendix A.

Results

Table 1 gives the results of each categorization method distinguishing misspellings from the remaining words. The results are quite high, which suggest we can do a good job of distinguishing simply misspelled words from other words. The best results were obtained using the C modified least squares (CMLS) method. The LogLinear results are almost as good. The Mahalanobis results are a little off the pace with Boosting appearing a small bit lower still.

While the *WSJ* corpus is not a true dictionary of correctly spelled words, it is a resource outside of the MEDLINE® corpus. Table 1 also reports results without the two features based on the *WSJ* corpus. As one would expect, the 11-pap scores are reduced, but only a small amount. Even the relative order of the methods remained unchanged.

While the difference is not large, CMLS produced better results than the other methods. This method will be used in our more detailed results and discussion. It will also be used in the applications that follow. Table 2 reports the precision, F score, and raw score obtained at 11 different recall values using the CMLS method and all features.

In order to judge the relative importance and value of the features, we applied the CMLS method using the features

TABLE 2. Eleven precision points for simple misspellings using all features and the CMLS method.

Recall	Precision	F score	Raw score
0.0	1.000	0.000	1.535
0.1	1.000	0.182	1.214
0.2	1.000	0.333	1.093
0.3	0.993	0.461	0.890
0.4	0.984	0.569	0.773
0.5	0.968	0.659	0.673
0.6	0.958	0.738	0.508
0.7	0.915	0.793	0.255
0.8	0.860	0.829	−0.019
0.9	0.706	0.791	−0.505
1.0	0.413	0.585	−3.097

individually and leaving out one feature at a time. Table 3 reports the 11-point average precisions from these experiments. Since Alternative Frequency is the most important single feature, we also applied the method using one other feature and the Frequency of Alternatives (Alt Freq). Finally, Table 3 reports the coefficients obtained by the CMLS method using the full model with all the features.

The feature set was selected to help identify simple spelling errors. But we also generated models for identifying complex spelling errors and abbreviations. These results appear in the figures below, and details are in Appendix B.

In our work comparing different methods for detecting misspelled words, the model orders the words from most likely to be misspelled to least likely to be misspelled. There are applications where a concrete misspelled/not-misspelled decision must be made. The ranking provided by the models allows many possible decisions points. The choice would depend on the application's balance between the value of finding more misspelled words and the cost of claiming correct words are misspelled. When using CMLS for classification, the score of zero is a reasonable, balanced cutoff. It is the natural cutoff for a method that tries to assign values of +1 and –1 to the items being classified. In our case, these are misspelled and correctly spelled words. Looking at Tables 2, 6, and 7, we see that the recall with the highest F scores had raw scores close to zero. A large F score suggests a good balance between precision and recall. Depending upon the needs of the application, a larger cutoff would give better precision while a lower cutoff would improve recall.

Application

By the end of 2000, the MEDLINE® database contained nearly 11 million article entries. These entries contain 417,395,580 word occurrences, of which 84,321,664 appear in article titles. Figure 1 shows the number of words each year in total, and only in titles. Both curves show strong, steady growth. In 1975, MEDLINE® began including abstracts in the entries. This explains the dramatic increase in total words that year.

The temporal distribution of misspellings and abbreviations in words with an α score less than 10 was investigated. We used the CMLS method and the models developed from the test case words scored by hand. Words with a score larger than zero were considered misspellings, or abbreviations depending on the model being investigated. Results for both the simple and complex misspelling models are reported. From the 42,340 alphabetic $\alpha < 10$ words, the simple spelling mistake model predicted 16,274 misspelled words, while the complex spelling model predicted 20,939 mistakes. The abbreviation model predicted 3,463 abbreviations. The precisions and recalls for these models appear in Table 4.

It is important to note that these are not all the misspellings and abbreviations in MEDLINE®. First, there are the observed recall limits derived from the test data. Then there are the unknown numbers of misspellings and abbreviations that have α scores larger than 10. These have not been included in this investigation.

To track the occurrences of misspellings by year, each appearance of an $\alpha < 10$ word in MEDLINE® was located and the year recorded. Then the number of $\alpha < 10$ words, spelling mistakes, and abbreviations each year can be estimated. Figure 2 shows the time distribution of tokens in titles with $\alpha < 10$ scaled by the total number of tokens in MEDLINE® titles. Scaling removes effects resulting from the increased growth in MEDLINE® each year. Limiting the words to alphabetic characters shows a small reduction. The plot also includes the number of spelling mistakes predicted by the complex and simple misspelling models. Finally, the distribution of abbreviations is shown. Figure 3 shows the same information for all text in MEDLINE®, both titles and abstracts.

From 1965 to 1975, the proportion of misspelled $\alpha < 10$ tokens in titles fluctuates around 0.0006. Beginning in 1975 until 1980, this proportion falls to less than 0.0001. Looking at all words, there is a sharp rise in $\alpha < 10$ words and misspelled words in 1975. The proportion of misspellings is about 0.0015 and falls to about 0.0002 by the early 1980s.

Spelling mistakes occur at a higher rate in the text of abstracts than in the text of titles. This may be seen in Figure 3 with its large increase in the proportion of misspelled words in 1975. As noted in Figure 1, this is the same year MEDLINE® began including abstracts in the references. From 1975 to 1980, the proportion of spelling mistakes in

TABLE 3. Properties of individual features using the CMLS Categorization Method.

Feature	Only feature	Without feature	Feature and alt freq	Coefficient
Frequency	0.415	0.886	0.769	−0.00944
Alpha	0.445	0.891	0.770	−0.00516
Number of alternatives	0.437	0.860	0.873	−0.03444
Frequency of alternatives	0.770	0.623		0.181854
Alpha of alternatives	0.549	0.891	0.835	8.42E-05
Wall Street Journal	0.576	0.881	0.800	0.070752
WSJ of alternatives	0.555	0.890	0.771	−0.02666
Trigram	0.536	0.890	0.801	3.32E-05

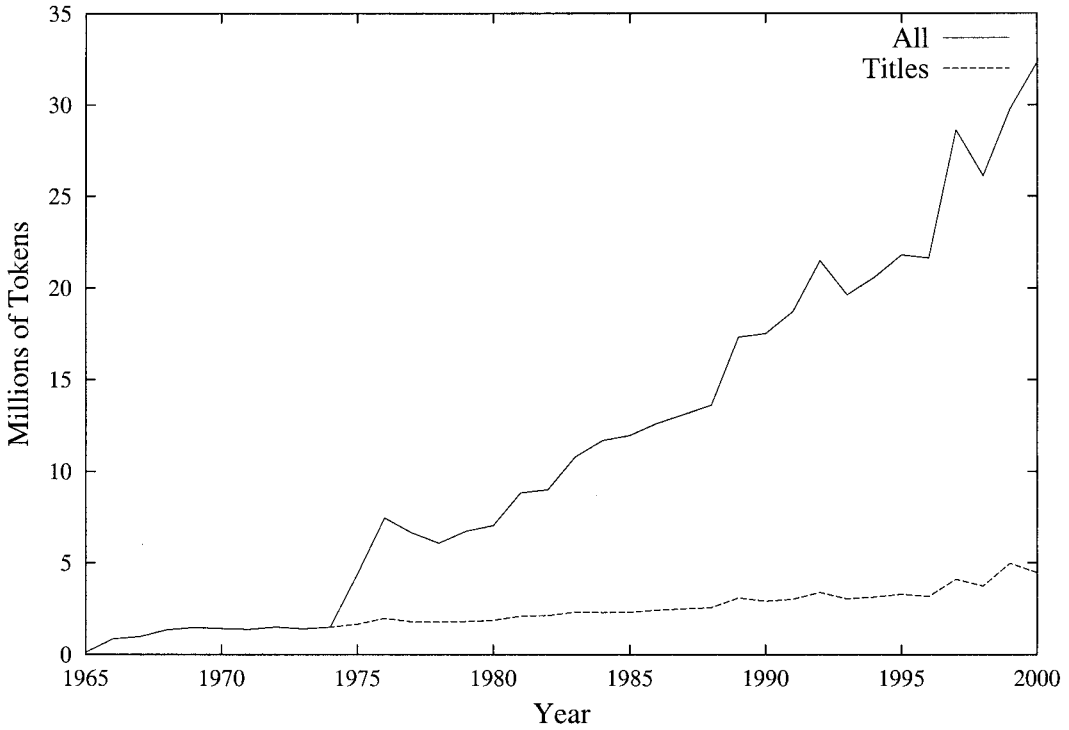

FIG. 1. Total number of tokens per year.

abstracts is about 0.0015 while in titles it is near 0.0006. From 1980 on, the proportion of mistakes in abstracts is about 0.0002 while in titles it is less than 0.0001. As often as abstracts are reviewed for mistakes, the title of a paper is considered even more.

The drop in the proportion of misspelled words in the early 1980s coincides with the widespread adoption of word processors and personal computers. These tools provide automatic spell checkers, which greatly reduces the number of undetected and unfixed spelling mistakes left by authors. At this time the National Library Of Medicine also started using a dictionary to spell check MEDLINE® references. The resulting drop in misspelled words, for both titles and abstracts, is more than sixfold.

Figure 4 shows the temporal distribution of correct words with $\alpha < 10$. These were determined by subtracting the misspelled words detected by the complex misspelling model from all $\alpha < 10$ words. While not perfectly flat, the number of correct words is fairly smooth. Real correctly spelled words that are nearly independent of their contexts (low α) are used at a steady pace. The proportion of 0.0005 is consistent over the more than 30-year span of MEDLINE® references. There is an increase in "correct" low-context words between 1975 and 1980, a period of increased proportion of misspelled words. These may likely be misspelled words that were not detected by our method. As seen in Table 4, the complex misspelling model has a recall of 0.763 at the threshold used.

TABLE 4. Precision and recall using CMLS with a threshold of zero.

Model	Precision	Recall
Simple misspelling	0.858	0.792
Complex misspelling	0.806	0.763
Abbreviation	0.918	0.761

Over this entire time, and in both titles and abstracts, abbreviations with $\alpha < 10$ also remained a nearly constant proportion of words. True, the total proportion of abbreviations would be much higher. Abbreviations with a consistent meaning often have significant context and high α. Nonetheless, it is interesting to see that, in the realm of low α, abbreviations and misspellings behave differently.

The National Library of Medicine takes the quality of MEDLINE® very seriously. Today, most of the MEDLINE® citations are provided electronically by the publisher. This effectively eliminates the possibility of introducing misspellings in those documents. About 30% are still entered via keyboard or OCR. Regardless of the source of the data, all citations are checked against an 82,000-word dictionary. Words in the citation that don't appear in the dictionary are manually compared against the original source. However the technicians making the comparisons do not have a scientific background and so may not recognize misspellings in the original source.

The library began using the dictionary for spell checking about 1980. This coincides with the dramatic drop in misspellings we see at that time. During parts of 1996–1997 labor problems prevented spell checking of some articles. We see a small rise in misspellings at that time.

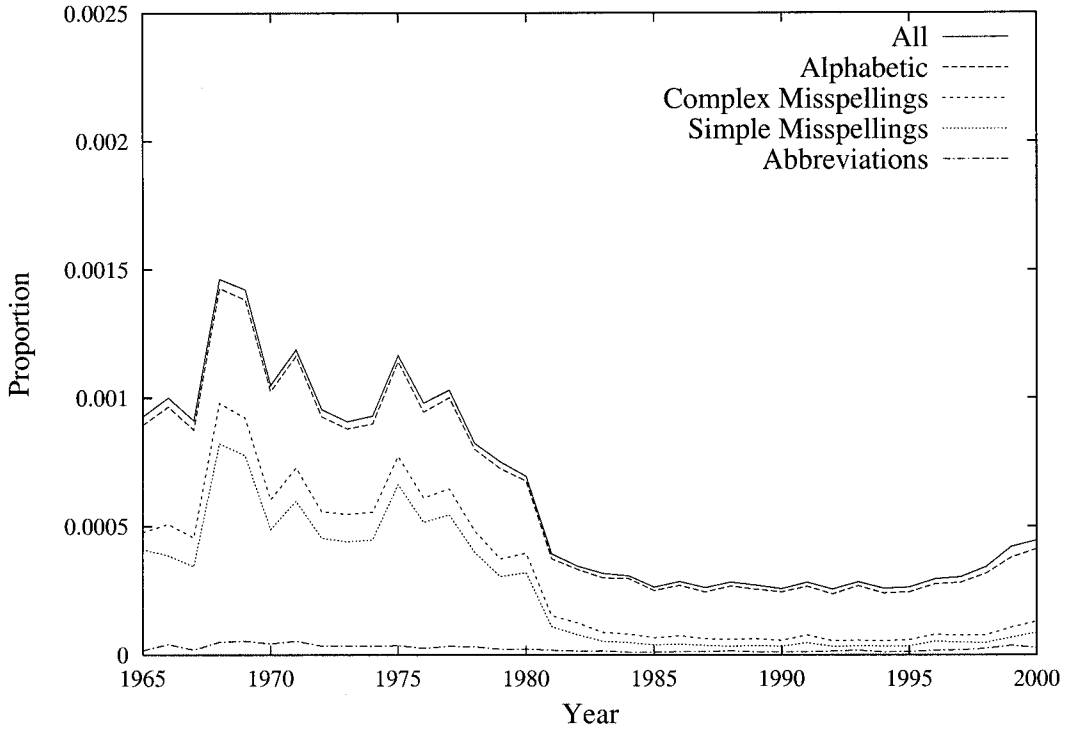

FIG. 2. Proportion of tokens in titles with $\alpha < 10$ by classification.

In order to address the question of the origin of the spelling mistakes, we examined a sample of mistakes in detail. A random sampling of 500 possible spelling mistake occurrences was reviewed to see which ones were actually mistakes. Of the 500 occurrences, 437 were confirmed to be misspellings. This rate of 87.4% is in the neighborhood of the predicted 86%.

We examined the 437 confirmed misspellings in an effort to determine their potential impact on information retrieval.

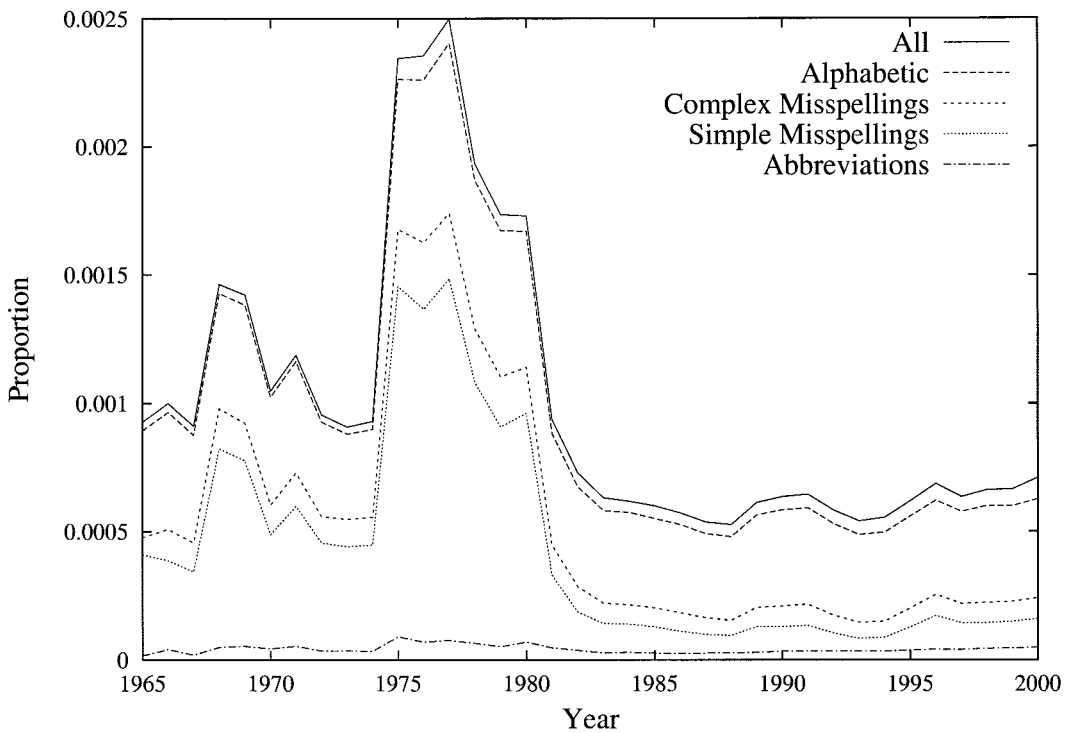

FIG. 3. Proportion of tokens with $\alpha < 10$ by classification.

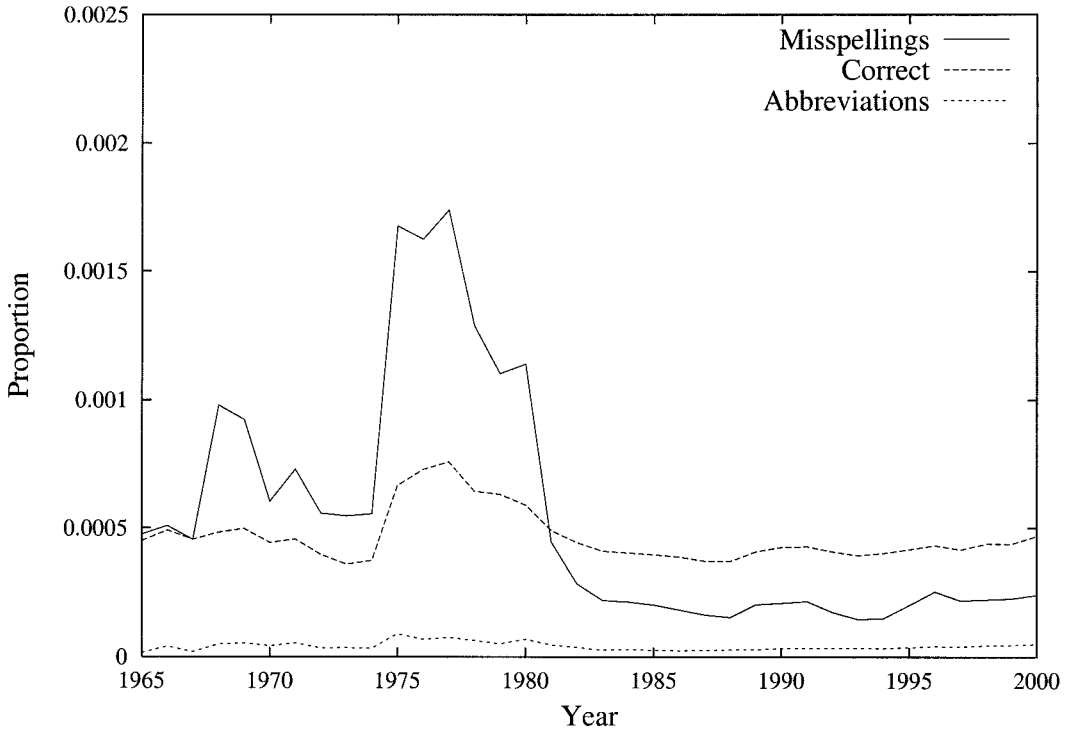

FIG. 4. Proportion of correct and misspelled tokens with $\alpha < 10$ using the Complex Misspelling Model.

In 94 cases, the misspelled word also appeared spelled correctly in the MEDLINE® document. Of the remaining 343 misspellings, we identified 140 cases where the mistake was in a content-bearing word with potential to be used in a query. Thus correction of MEDLINE® could potentially improve retrieval in 32% of misspelling instances.

We further examined the 140 cases of content-bearing word misspellings to better understand their origin. We compared these 140 cases with their corresponding original journal articles. MEDLINE® and the journal were the same 88 times. They were different 47 times. Comparisons were not made five times because the information was translated to English by the library (twice) or the original was not immediately available at the library (three times). When the original and MEDLINE® differed, the original was usually correct. Twice the original had a different misspelling than appeared in MEDLINE®. The majority of the time, 64% (88/137), when a significant spelling mistake is found in MEDLINE®, it originated in the source document.

Discussion

Now that we know that misspelled words can be detected using features of the word and the corpus, it is useful to consider the individual features and their importance to the result. The following eight features were used to describe each of the words. They were based on already available statistics and easily computable values. Their contributions to the categorization using the CMLS method are depicted in Table 3. It is interesting that some features, such as Number of Alternatives provide negligible value on their own, but provide a useful addition to Frequency of Alternatives.

- Frequency of Alternatives: Log of the total number of times alternatives appear in MEDLINE®. The more frequently a word (the alternative) appears, the more often it will be misspelled. This is consistent with simple spelling errors in carefully checked work being distributed randomly. This is the most important single feature.
- Frequency: Log of the number of times the word occurred in MEDLINE®. The frequency of the incorrect word has little impact on the model. One would expect the more frequent the word, the less likely it is a misspelling. Because all the words considered have very low α scores, these words tend to have low frequencies. Thus our test set does not provide the broad sampling needed to adequately judge this parameter. The negative coefficient is consistent with "the more frequent a word is, the less likely it is a misspelling."
- Alpha: The strength of a word's context (Kim & Wilbur, 2001). The α score is crucial to our method. We only consider words with a very small α as possible misspellings. As a result, the model never has a chance to learn the relationship between α and misspellings over a large range of α. Nonetheless, the negative coefficient is consistent with words that have higher α values, and hence greater value to their context, are less likely to be misspellings. This parameter would be crucial in any attempt to extend the method to words of higher α.
- *Wall Street Journal* (*WSJ*) Score: Log of the ratio of probability of this word appearing in MEDLINE® versus in the *Wall Street Journal* (Kim & Wilbur, 2001). This feature had a modest effect on the model. The coefficient was positive,

which means that a word seen more often in MEDLINE® than the *Wall Street Journal* is more likely to be a misspelling.

- Trigram: The word is split into overlapping letter trigrams. The counts for each of these trigrams are averaged. The individual trigram counts were obtained by counting all the individual trigrams in each of the words in MEDLINE® which occurred three or more times. This trigram feature played a very small role in the final model and was not very helpful by itself. But combined with Frequency of Alternatives, it was fairly helpful. We suspect it provides similar information to the Number of Alternatives feature. The positive coefficient says that the more common the trigrams in the word, the more likely it is a misspelling. This avoids marking abbreviations as misspellings. The Number of Alternatives does an even better job at avoiding abbreviations.
- Number of Alternatives: Number of words in MEDLINE® that differ by only one single transformation. This is the second most important feature. Including it is more valuable than all the features other than Frequency of Alternatives. This feature avoids calling abbreviations spelling mistakes. Many abbreviations are a single change away from another abbreviation. This does not mean that the abbreviation is incorrect. As discussed elsewhere, an abbreviation can receive a low α score because it can be ambiguous and stand for different phrases. Yes, incorrect abbreviations do occur. Automatically detecting them would require different tools, such as sense disambiguation.
- Alpha of Alternatives: Highest alpha of an alternative word. The alpha score of the possible correct words is not very useful for predicting which words are misspelled. Alpha scores of correct words range all over the map, from quite low to very high. A misspelling of a high content word with a high α is still a misspelling, the same as the misspelling of a low content, low α word.
- *WSJ* of Alternatives: Highest *WSJ* score of an alternative word. This feature has little effect on the model. The negative coefficient suggests if the alternatives are seen more often in the *Wall Street Journal* than MEDLINE®, then the word is more likely a misspelling. But the effect on the model is so small that no firm conclusions can be drawn.

Since comparing the *Wall Street Journal* corpus with MEDLINE® could be seen as using a dictionary, we also categorized misspellings without the features derived from the *WSJ*. The *WSJ* made a positive, although modest, contribution to the model. The results without the *WSJ* are still quite good. An external corpus is not required to use internal features of a large corpus for detecting spelling errors.

Complex Misspellings include all the types of misspellings labeled Simple Misspellings, Incorrect Inflections, Incorrect Compounds, and Other Misspellings. As described in an appendix, we were able to detect Complex Misspellings with a feature set designed for Simple Misspellings. How is this possible? Some Complex Misspellings result from multiple simple mistakes. In these cases, there is a misspelling with one less single-error misspelling that plays the role of the "correct alternative." While incorrect, it appears more frequently and more often in its context than the badly misspelled word. It is surprising that even with Incorrect Inflections and Incorrect Compounds, the results are nearly as good as for Simple Misspellings. Apparently, it is nearly as effective to include them as misspellings as it is to exclude them.

Again, as noted in an appendix, a model with the same feature set used to predict misspellings was also able to predict abbreviations. In this model, the coefficient of the trigram feature was negative, indicating that the more frequently a word's trigrams are seen in MEDLINE®, the less likely the word is an abbreviation. Indeed abbreviations often include trigrams that are not frequently seen in ordinary text.

Once misspelled words have been found and identified, they can be used to spell check other works in the field. For example, they can be used to look for misspellings in the Unified Medical Language System (UMLS) (Humphreys, Lindberg, Schoolman, & Barnett, 1998) (such as "adenocarinoma" in concept C0238517, UMLS, version 2002AA). A potentially useful project would be the application of our methodology to full text articles. It may prove instructive to compare the misspelling rates of entire articles with those of the titles and abstracts considered in this current work.

A benefit to information retrieval would come if spelling mistakes could be not only detected, but corrected. This was not directly investigated, but a positive observation can be made. Many of the misspelled words only have one plausible alternative spelling. It is likely that, in most instances, that alternative would be the correct spelling. In cases with multiple alternatives, the available context should be very helpful in choosing the proper alternative. Of course, the misspelling rate is very small compared to the typical recall rate of information retrieval methods, so that the improvement would be modest.

A useful extension of this work would be to words with $\alpha > 10$. While misspelled words can be expected to have a lower α than their correct spelling, many still have $\alpha > 10$. In order to apply the learning methods used here, one would need a set of human judgments of words with a range of α values greater than 10.

Conclusion

These positive results support the value of α as a useful and reliable measure of the strength of a word's context. Otherwise, it would not have been possible to reliably detect misspelled words. Using α scores and the frequency of alternative spellings, it is possible to detect misspelled words without a dictionary. This is valuable in a corpus that contains many correct words that do not appear in any dictionary. Other than a modest benefit from comparison with the *Wall Street Journal* Corpus, these methods and models did not take into account any special features of the medical science literature. They should generally be applicable in any field where a large corpus of work is available.

References

Bentley, J. (1985). A spelling checker. Communications of the ACM, 28(5), 456–462.

Bowden, T. (1995). Cooperative error handling and shallow processing. Seventh Conference of the European chapter of the Association for Computational Linguistics, March, Dublin, Ireland.
Damerau, F.J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3), 171–176.
Duda, R.O., Hart, P.E., & Stork, D.G. (2000). Pattern classification. New York: Wiley.
Elmi, M.A., & Evens, M. (1998). Spelling correction using context. Proceedings of the 36th ACL, 17th COLING, Montreal, Quebec, Canada.
Golding, A.R., & Schabes, Y. (1996). Combining trigram-based and feature-based methods for context-sensitive spelling correction. Proceedings of the 34th annual meeting of the Association for Computational Linguistics, Santa Cruz, California.
Humphreys, B.L., Lindberg, D.A., Schoolman, N., & Barnett, G.O. (1998). The Unified Medical Language System: An informatics research collaboration. Journal of the American Medical Informatics Association, 5(1), 1–11.
Johnson, M., Geman, S., Canon, S., Chi, Z., & Riezler, S. (1999). Estimators for stochastic "unification-based" grammars. Proceedings ACL '99, University of Maryland, College Park.
Kim, W., & Wilbur, W.J. (2002). DNA splice site detection: A comparison of specific and general methods. AMIA annual symposium, November, San Antonio, Texas.
Kim, W.G., & Wilbur, W.J. (2001). Corpus-based statistical screening for content-bearing terms. Journal of the American Society for Information Science, 52(3), 247–259.
Kukich, K. (1992). Techniques for automatically correcting words in text. ACM Computing Surveys, 24(4), 377–439.
McIlroy, M.D. (1982). Development of a spelling list. IEEE Transactions on Communications, COM-30(1), 91–99.
Nylander, S. (1999). Statistics and phonotactical rules in finding OCR errors. 12th Nordiske datalingvistikkdager, Trondheim, Norway.
Schapire, R.E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. Machine Learning, 37(3), 297–336.
Witten, I.H., Moffat, A., & Bell, T.C. (1999). Managing gigabytes. San Francisco: Morgan-Kaufmann.
Zhang, T., & Oles, F.J. (2001). Text categorization based on regularized linear classification methods. Information Retrieval, 4(1), 5–31.

Appendix A: Method Details

Data

We looked at the words in MEDLINE® with an α score less than 10. Casual observations of words with low α scores showed that many of them were misspellings. The rate of misspellings seemed to drop near $\alpha = 10$. A word needs to occur at least 3 times to allow the calculation of an α score. In MEDLINE® references through the end of 2000, 535,533 words appear three or more times. Kim and Wilbur (2001) calculated the α score and a *Wall Street Journal* (*WSJ*) score for these words. The *WSJ* is the log of the ratio of the probability of seeing the word in MEDLINE® to the probability of seeing the word in the *Wall Street Journal*. Since we were looking for misspellings, we considered only purely alphabetic words, i.e., contain only letters. MEDLINE® contained 42,340 purely alphabetic words occurring three or more times and with $\alpha < 10$.

Test Set

Human judges examined a portion of the $\alpha < 10$ words and reported on their use. From the set of alphabetic $\alpha < 10$ words, 2,000 were randomly selected. For each of these 2,000 words, three occurrences in MEDLINE® were selected. (Due to updates in MEDLINE®, for a few words, only two occurrences were available.) For each of the nearly 6,000 occurrences, a human judge was provided with the word in question highlighted in the context of the title or abstract in which it appeared. The judge was also shown a list of strings in MEDLINE® that differed from the word in question by a single-error transformation. Based on this information about the word and the context of its use, the judge selected one of the options listed in Table A1. The occurrences of a word were treated independently, so a word could be assigned to one, two, or even three categories.

The Correct, Abbreviation, and Proper Name categories are clear. (The Proper Name category was treated as correct in this project.) A Simple Misspelling is one of the single-error misspellings identified by Damerau (1964). An Incorrect Inflection is a word that was inflected by usual rules, but is a mistake because the word does not follow those rules or another phrase should be used instead. For example, "growther" versus "more growth." An Incorrect Compound is formed when two correctly spelled words are combined into a single word when they should remain as individual words.

The judges assigned each of the three occurrences of a word to a category. Table A1 reports the number of words assigned to a category at least once. The total is more than 2,000 because some words were assigned to different categories in their different occurrences. Of these 2,000 words, 1830 were rated the same in each occurrence. Examples include "Borchert" a proper name, "gmba" an abbreviation, and "pletelets" a simple misspelling. The remaining 170 words received different ratings in their different occurrences. One example is "whch": an abbreviation for wheelchair or a misspelling of which. Another example "bais" played three different roles: a person's name ("C. Bais et al."), an abbreviation (plural of BAI, for "breath-actuated inhaler"), and a misspelling of basis ("on the bais of two criteria").

Of the 2,000 words judged, 1,000 were randomly selected as the training set, while the remaining 1,000 words became the test set.

TABLE A1. Word usage options and occurrence counts.

Word usage options	Count
Correct	595
Abbreviation	172
Proper name	348
Simple misspelling	798
Incorrect inflection	101
Incorrect compound	102
Other misspelling	72

Categorization Methods

We tested four categorization methods which were appropriate for the task. All of these methods learn weights for the different features. An instance is scored by combining its features with the weights. The objective is to learn weights that score members of one category high and members of the other low so they can be separated based on their features. These scores do not always provide a clear distinction between the sets, but they do provide a ranking from most likely to be in the first set to least likely. The methods tested were Mahalanobis distance (Duda, Hart, & Stork, 2000), a log-linear model (Johnson, Geman, Canon, Chi, & Riezler, 1999), the CMLS wide margin classifier of Zhang and Oles (2001), and linear boosting (Schapire & Singer, 1999). CMLS provided our best results.

The Mahalanobis distance approach assumes the distributions of correct and misspelled words can be described by multidimensional normal distributions over the features. Using the correct and misspelled words from the training set, the means and covariance matrices for the two distributions are determined. A measure of how well each word in the test set fits into each distribution can then be calculated (Duda, et al., 2000).

Our LogLinear model is based on the ideas expressed in Johnson et al. (1999).

CMLS is a wide margin classifier (Zhang & Oles, 2001). It is related to Support Vector Machines (SVM), but has a smoother penalty function. This allows the calculation of gradients which can provide faster convergence (Kim & Wilbur, 2002).

Boosting is a technique for improving learning by repeatedly refocusing a learner on those parts of the training data where it has not yet proved effective. We used the linear AdaBoost algorithm presented by (Schapire & Singer, 1999).

Evaluation

The natural measures of information retrieval are recall and precision. Recall is the proportion of the desired items actually recovered. Precision is the proportion of the returned items that were desired. Unfortunately, each measure is trivial to maximize if the other is ignored.

Two measures that combine precision and recall in a useful way are F scores and 11-point average precisions. The F score, as the harmonic mean of precision and recall, represents one recall-precision pair. Generally, the F score favors a balance between precision and recall. Eleven-point average precision (11-pap) is a value that reflects the quality of categorization over a range of recall levels (Witten, Moffat, & Bell, 1999). To determine 11-pap, the highest precision seen at or after 11 different recall levels (0.0, 0.1, 0.2, . . ., 0.9, 1.0) is recorded. The average of these 11 precision values is the 11-pap.

Appendix B: Additional Results

The feature set we used was designed to detect simple spelling errors. In fact, the alternative words used to derive several of the features were obtained by considering possible simple spelling errors. However, detecting all mistakes would be preferable. As noted above, our judges also indicated Incorrect Inflections, Incorrect Compounds, and Other Misspellings. We used CMLS to generate a model where any word with an occurrence of any of these types was considered a misspelling. This obtained an 11-point average precision of 0.881. This is referred to as the "complex misspelling" model. The precisions, F scores, and raw scores obtained appear in Table B1.

TABLE B1. Eleven precision points for complex misspellings.

Recall	Precision	F score	Raw score
0.0	1.000	0.000	2.041
0.1	1.000	0.182	1.282
0.2	1.000	0.333	1.115
0.3	0.994	0.461	0.924
0.4	0.986	0.569	0.818
0.5	0.964	0.658	0.635
0.6	0.923	0.727	0.385
0.7	0.856	0.770	0.157
0.8	0.767	0.783	−0.106
0.9	0.671	0.769	−0.329
1.0	0.529	0.692	−1.302

Abbreviation detection is not the point of this paper. However, the judges identified "words" which are abbreviations. Using these same features, CMLS generated a model that did a good job of identifying abbreviations. It achieved an 11-point average precision of 0.853. The details are in Table B2.

TABLE B2. Eleven precision points for abbreviations.

Recall	Precision	F score	Raw score
0.0	1.000	0.000	2.931
0.1	1.000	0.182	1.898
0.2	1.000	0.333	1.517
0.3	1.000	0.462	1.318
0.4	0.980	0.568	0.779
0.5	0.980	0.662	0.779
0.6	0.951	0.736	0.382
0.7	0.929	0.798	0.091
0.8	0.899	0.846	−0.186
0.9	0.519	0.659	−0.889
1.0	0.131	0.231	−1.350