Identification of Misspelled Words without a Comprehensive Dictionary Using Prevalence Analysis

Alexander Turchin, MD, MS[a,b,c], Julia T. Chu[c], Maria Shubina, ScD[b], Jonathan S. Einbinder, MD, MPH[a,b,c]

[a]*Clinical Informatics Research and Development, Partners HealthCare, Boston, MA*

[b]*Brigham and Women's Hospital, Boston, MA*

[c]*Harvard Medical School, Boston, MA*

Abstract

Misspellings are common in medical documents and can be an obstacle to information retrieval. We evaluated an algorithm to identify misspelled words through analysis of their prevalence in a representative body of text.

We evaluated the algorithm's accuracy of identifying misspellings of 200 anti-hypertensive medication names on 2,000 potentially misspelled words randomly selected from narrative medical documents. Prevalence ratios (the frequency of the potentially misspelled word divided by the frequency of the non-misspelled word) in physician notes were computed by the software for each of the words. The software results were compared to the manual assessment by an independent reviewer.

Area under the ROC curve for identification of misspelled words was 0.96. Sensitivity, specificity, and positive predictive value were 99.25%, 89.72% and 82.9% for the prevalence ratio threshold (0.32768) with the highest F-measure (0.903). Prevalence analysis can be used to identify and correct misspellings with high accuracy.

Introduction

Healthcare organizations are rapidly adopting electronic medical record systems[1]. Narrative medical documents contain a large fraction of the data stored in the electronic medical records[2]. However, retrieval of information from narrative documents presents a number of technical challenges.

A typical information retrieval task (e.g. searching World Wide Web using Google) includes identification of documents that contain some or all members of a set of semantic concepts entered by the user. Many narrative medical documents are created under significant time constraints and are not proofread afterwards, resulting in frequent misspellings[3, 4]. Misspellings may not be recognized as related to the semantic concepts that are being sought, decreasing the sensitivity of information retrieval. At the same time, lexical complexity of narrative medical texts prevents application of the lexicon-based methods for identification of misspellings that are commonly used elsewhere[5]. The vocabulary of medical texts is technical and constantly expanding. Additionally the texts contain many poorly (if at all) standardized abbreviations and acronyms and a wide variety of proper (e.g. patients' and health care providers') names[4]. Consequently no comprehensive vocabulary including all words that can be found in narrative medical texts exists, and building one is not feasible. Unsurprisingly, published reports that have evaluated identification of misspelled words in medical documents using existing vocabularies (e.g. UMLS) report relatively low sensitivity[6].

In most cases, in a sufficiently large body of text composed of documents created by different authors, any particular misspelling of a given word is expected to be encountered with lower frequency than the correct spelling of the word. We therefore evaluated the accuracy of an algorithm that identifies misspellings of a given set of words in a large body of narrative medical text based on the analysis of their relative prevalence in the text.

Materials and Methods

Algorithm

The prototype software was implemented in Perl. The software takes as input three sources of data:

1. One or more plain text files of unlimited size that contain representative narrative documents.
2. A text file that contains the words for which the software will identify misspellings in the medical narrative text provided in source # 1.
3. A text file that contains a list of words in the general English vocabulary. Linux.words[7] – a publicly available list of 45,402 words – was used in the prototype implementation.

For every word greater than four characters in length in each of the narrative text files, the algorithm performs the following steps:

1. Determine whether the word is an exact match for one of the words in the general English vocabulary.
2. If # 1 is false, then determine whether the word is an exact or plural match to one of the words whose misspellings are being identified. The number of times each of the words whose misspellings are being identified was found in the entire text body is counted and recorded.

3. If # 2 is false, then determine the Levenshtein distance[8] (also known as edit distance) between the word and each of the words whose misspellings are being identified. In this study, the standard definition of Levenshtein distance as the total number of letter insertions, deletions and transpositions necessary to convert one word to another was used. Subsequently Levenshtein distance ratio is calculated as the ratio of the Levenshtein distance to the length of the word whose misspellings are sought.
4. If the smallest Levenshtein distance ratio between the word in the text and any of the words whose misspellings are being identified is below the threshold value of 0.25, the text word is recorded. The number of times this word is found in the entire text body is counted and recorded.
5. After the entire text body has been analyzed, prevalence ratio is computed for each of the words within the threshold Levenshtein distance ratio from one of the words whose misspellings are being identified. Prevalence ratio is calculated as the number of times the word was found in the entire text body divided by the number of times the word to which it was closest by the Levenshtein distance ratio was found. For example, if the word "accupral" is found twice, and the word "accupril" is found 100 times, prevalence ratio is calculated as 0.02.
6. The words in the text body whose prevalence ratio to one of the words whose misspellings are being identified is below the *prevalence ratio threshold* are identified as misspellings. The purpose of this evaluation was to identify the optimal prevalence ratio threshold that confers the highest accuracy in this algorithm.

Dataset

We used the prototype software to identify misspellings of 200 names of anti-hypertensive medications in the text of narrative physician notes. Accuracy of identification of misspelled words was assessed on the dataset comprised of 2,000 words that were within the Levenshtein distance ratio threshold of 0.25 from one of the anti-hypertensive medication names. These words were randomly selected from narrative physician notes from the electronic medical record system at Partners HealthCare System. All of the notes were created by physicians in outpatient practices affiliated with either Massachusetts General Hospital or Brigham and Women's Hospital (both in Boston, MA).

Each of the words was analyzed by an independent reviewer who did not participate in the design of the software. The line of text on which the word was found was supplied to the reviewer to provide context. For each of the 2,000 words the reviewer made the determination whether the word represented a misspelling of one of the anti-hypertensive medication names or was a different word. The reviewer's analysis was subsequently used as the gold standard to which the software results were compared.

Evaluation

The goal of the evaluation was to assess the accuracy of the software and to determine the optimal prevalence ratio threshold that permitted the best discrimination between misspellings and real words.

Based on the distribution of the prevalence ratios in the narrative medical text under analysis, we selected prevalence ratio thresholds for evaluation that started at 10^{-5} and then increased in geometric progression by a factor of two until 671.0886. For each of these 27 thresholds we compared the software results with the manual review and calculated the following parameters:

1. Sensitivity (recall)
2. Specificity
3. False positive fraction (1-Specificity)
4. Positive predictive value (precision)
5. F-measure

F-measure was calculated according to the following formula:

$$F = \frac{(1+\beta^2)\times R\times P}{(\beta^2\times P)+R}$$

where R represents recall and P represents precision. The β coefficient indicates the value of precision relative to recall; we used β = 1 which gives equal weights to recall and precision. Receiver operator characteristic (ROC) curve was generated by plotting sensitivity against false positive fraction for each of the thresholds.

Statistical Analysis

Normal approximation[9] was used to calculate confidence intervals for sensitivity, specificity, and positive predictive value. Trapezoidal rule was used to calculate the area under the ROC curve. The conservative variance estimator based on the maximum variance over distributions with the same expected area under the ROC curve (AUC) was used to calculate AUC confidence interval[10].

IRB

The study protocol was reviewed and approved by Partners Human Research Committee.

Results

The software processed 4.31 GB of narrative medical text over 34.2 hours. A total of 748,147,019

words were identified. Of these, 484,145,968 (64.7%) matched to one of the words in the general English vocabulary. Among remaining words, 2,244,266 matched one of the 200 test set words representing the names of anti-hypertensive medications. The software identified 172,415 occurrences of 4,658 unique words that were within Levenshtein distance ratio of 0.25 or less from the test set words.

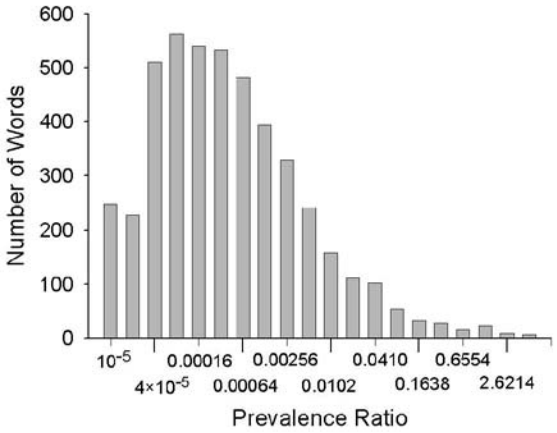

Figure 1
Distribution of Potentially Misspelled Words in Narrative Documents by Prevalence Ratio

The distribution of the prevalence ratios of the 4,658 words identified as potential misspellings of the test set words was skewed to the left (Figure 1). Over 90% of the words had prevalence ratios below 0.01. The lowest prevalence ratio identified in our dataset was 4×10^{-6} (for "atenololg" compared to "atenolol"), and the highest 459.4 (for "retic" compared to "oretic").

Out of the 2,000 words in the evaluation dataset, manual review identified 667 (33.3%) as misspellings and 1,333 (66.7%) as real words. Sensitivity, specificity, precision and F-measure of the software for selected prevalence ratio thresholds are shown in Table 1. The software achieved the highest F-measure (0.90) at the prevalence ratio threshold of 0.32768. This threshold resulted in sensitivity of 99.3% (95% CI ± 0.75%), specificity of 89.7% (95% CI ± 1.6%) and precision of 82.9% (95% CI ± 1.6%).

Table 1
Accuracy of the Algorithm at Different Prevalence Ratio Thresholds

Threshold	0.00001	0.00008	0.00064	0.00512	0.04096	0.16384	0.32768	1.31072	10.48576	83.886	671
Sensitivity	0.75%	4.05%	18.44%	45.58%	67.92%	86.36%	99.25%	99.70%	100%	100%	100%
Specificity	100%	100%	99.70%	98.27%	96.02%	92.65%	89.72%	67.52%	30.01%	28.43%	9.15%
Precision	100.00%	100.00%	96.85%	92.97%	89.53%	85.46%	82.85%	60.56%	41.69%	41.15%	35.52%
F-measure	0.015	0.078	0.310	0.612	0.772	0.859	0.903	0.754	0.588	0.583	0.524

Table 2
Examples of the Words Correctly and Incorrectly Identified by the Software

True Positives	True Negatives	False Positives	False Negatives
labetolol (labetalol)	titrate (nitrate)	nocardia (Procardia)	Serapes (Ser-Ap-Es)
cardiazem (Cardizem)	Alan (Calan)	Caplan (Calan)	Cardia (Cartia)
linisopril (lisinopril)	eosinophil (fosinopril)	MonoJet (Monoket)	Betalol (Betaloc)

The words from the test set closest to text words identified as potential misspellings are given in parentheses

Examples of the words correctly and incorrectly identified by the software at this threshold are provided in Table 2.

The area under the ROC curve for prevalence ratio was 0.967 (95% CI ± 0.014), indicating excellent accuracy (Figure 2).

Discussion

Misspellings are common in narrative medical documents[4, 11] and several reports have highlighted them as one of significant obstacles to effective information retrieval from medical texts[4, 12]. Most modern spellcheckers[11] identify misspelled words based on a) lack of a match to a word in a lexicon and b) a measure of proximity to one of the words in the lexicon, most commonly Levenshtein distance[8], tri-grams[5] or similarity keys[13]. Identifying misspellings in medical texts based on these algorithms can be particularly challenging because of the constantly evolving technical vocabulary and frequent use of acronyms and abbreviations as well as proper names. Consequently creation of a comprehensive lexicon is not feasible, limiting the efficacy of both steps of the algorithm.

A number of investigations have been carried out to address this challenge using both specialized medical lexicons, such as UMLS[6], and context-based approach[14]. In this report we describe the first, to our knowledge, analysis of using word prevalence in a representative collection of medical text to identify

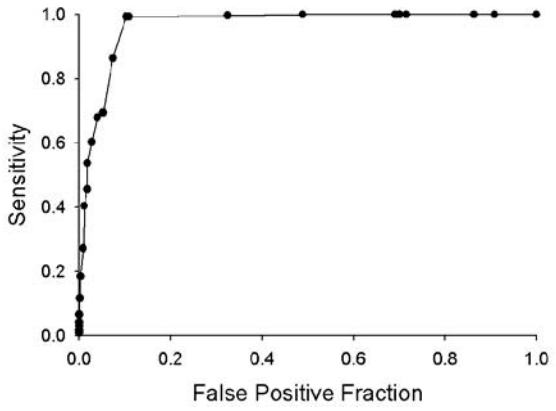

Figure 2
ROC Curve for Prevalence Ratio Threshold

misspelled words and correct errors in medical narrative.

The goal of the algorithm we developed is to identify all misspellings of a given set of words in a body of text. While not designed for real-time error correction, it has important applications in information retrieval from narrative medical documents. It is common for users of large medical databases to search for a set of documents that contain a given set of keywords, where comprehensiveness of the results may be more important than speed[12, 15, 16]. An algorithm similar to the one we present could be used to identify all common misspellings of the keywords using a representative body of narrative documents prior to running the query. The common misspellings could then be included in the keyword list to improve sensitivity of the query. Another potential application of the algorithm is in semantically-driven document classifiers focused on a specific domain. A number of these applications have been described in the literature[17-19], and some of them explicitly address the detection of misspellings in their design[18]. Our group's experience with similar tasks[20] indicates that incorporation of misspellings in the definition of semantic fields can boost sensitivity of information retrieval.

As demonstrated by the area under the ROC curve, prevalence analysis has a high discriminative ability for detection and correction of spelling errors. At the optimal prevalence ratio threshold it detected nearly all misspellings except names of several medications that were both very uncommon and counter-intuitively spelled (e.g. Ser-Ap-Es), leading to misspellings being more common that the correct medication name. At the same time, while achieving nearly total recall, the algorithm maintained high specificity of nearly 90%, resulting in an F-measure of over 0.9. This level of performance is sufficient for most applications and could likely be improved further if combined with other previously reported approaches, such as context analysis. While the algorithm requires availability of a large body of representative text to achieve this level of accuracy, narrative medical documents are increasingly becoming available in digital format[21] and this constraint is unlikely to be a significant obstacle in the future.

This study has several limitations. Only words longer than four characters were included in the evaluation and the list of potential misspellings was limited by the ratio of Levenshtein distance to the word length of 0.25. Previous investigations showed, however, that 80% of spelling errors are within Levenshtein distance of one from the original word[22]; therefore most spelling errors were likely to be included in our evaluation scheme. On the other hand, it is recognized that misspellings of words less than five characters long present a particular challenge for identification and correction[23]. A single letter change in one of these words is more likely to result in another word in the lexicon and consequently algorithms based on grammatical analysis of the sentence will likely be necessary to identify these errors. The technique we describe would not be able to identify a misspelling of a word from the given word list that resulted in another word on that list – a problem common for spellcheckers[24]. The evaluation was limited to spelling errors of names of anti-hypertensive medications. It is possible that the findings would not be applicable to other semantic domains. The body of medical text we used was geographically limited to practices affiliated with two academic hospitals in a single city and it is possible that the findings could not be extended to other geographic locales. However, previous reports showed that differences in accuracy of analysis of narrative medical documents between different geographic domains are not high[25]. The large number of documents in our dataset created by physicians who trained in many different institutions across the country could partially compensate for the geographic localization.

Conclusion

In summary, in this paper we report the results of the evaluation of an algorithm that identifies and corrects spelling errors for a set of words using analysis of word prevalence in a body of text. The algorithm demonstrated very high accuracy and could be used in a number of applications of information retrieval from narrative medical documents. The source code for the prototype software is available upon request.

Acknowledgments

This research was supported in part by a grant from Partners Information Systems Research Council.

References

1. de Lusignan S, Teasdale S, Little D, et al. Comprehensive computerised primary care records are an essential component of any national health information strategy: report from an international consensus conference. *Inform Prim Care.* 2004;12(4):255-264.

2. Hicks J. *The potential of claims data to support the measurement of health care quality.* [PhD]. San Diego, CA, RAND; 2003.

3. Shapiro AR. Taming variability in free text: application to health surveillance. *MMWR Morb Mortal Wkly Rep.* Sep 24 2004;53 Suppl:95-100.

4. Hersh WR, Campbell EM, Malveau SE. Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical records: a lexical analysis. *Proc AMIA Annu Fall Symp.* 1997:580-584.

5. Kukich K. Techniques for automatically correcting words in text. *ACM Computing Surveys.* 1992;24(4):377-439.

6. Tolentino HD, Matters MM, Walop W, et al. A UMLS-based spell checker for natural language processing in vaccine safety. *BMC Med Inform Decis Mak.* Feb 12 2007;7(1):3.

7. Faith R. Linux.words. http://www.ibiblio.org/pub/linux/libs/linux.words.2.lsm. Accessed 02/17/2007.

8. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.* 1966;10(8):707-710.

9. Rosner B. Section 6.8.2: Estimation. *Fundamentals of Biostatistics*. Boston, MA: Duxbury; 2005.

10. Birnbaum ZW, Klose OM. Bounds for the variance of the Mann-Whitney Statistics. *Annals of Mathematical Statistics.* 1957;28(4):933-945.

11. Ruch P, Baud RH, Geiddbuhler A, Lovis C, Rassinoux AM, Riviere A. Looking back or looking all around: comparing two spell checking strategies for documents edition in an electronic patient record. *Proc AMIA Symp.* 2001:568-572.

12. Fisk JM, Mutalik P, Levin FW, Erdos J, Taylor C, Nadkarni P. Integrating query of relational and textual data in clinical databases: a case study. *J Am Med Inform Assoc.* Jan-Feb 2003;10(1):21-38.

13. Pollock JJ, Zamora A. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM.* 1984;27(4):358-368.

14. Comeau DC, Wilbur WJ. Non-word identification or spell checking without a dictionary. *Journal of the American Society for Information Science and Technology.* 2004;55(2):169-177.

15. Einbinder JS, Scully KW, Pates RD, Schubart JR, Reynolds RE. Case study: a data warehouse for an academic medical center. *J Healthc Inf Manag.* Summer 2001;15(2):165-175.

16. Murphy SN, Gainer V, Chueh HC. A visual interface designed for novice users to find research patient cohorts in a large biomedical database. *AMIA Annu Symp Proc.* 2003:489-493.

17. Mendonca EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. *J Biomed Inform.* Aug 2005;38(4):314-321.

18. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *J Am Med Inform Assoc.* Sep-Oct 2005;12(5):517-529.

19. Turchin A, Pendergrass ML, Kohane IS. DITTO – a Tool for Identification of Patient Cohorts from the Text of Physician Notes in the Electronic Medical Record. *AMIA Annu Symp Proc.* 2005:744-748.

20. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inform Assoc.* Nov-Dec 2006;13(6):691-695.

21. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Fam Pract.* Apr 2006;23(2):253-263.

22. Damerau FJ. A technique for computer detection and correction of spelling errors. *Communications of the ACM.* 1964;7(3):791-801.

23. Pollock JJ, Zamora A. Collection and characterization of spelling errors in scientific and scholarly text. . *Journal of the American Society for Information Science and Technology.* 1983;34(1):51-58.

24. Pedler J. Computer spellcheckers and dyslexics - a performance survey. *British Journal of Educational Technology.* 2001;32(1):23-37.

25. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods Inf Med.* Jan 1998;37(1):1-7.