Journal of Biomedical Informatics 55 (2015) 188–195

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automated misspelling detection and correction in clinical free-text records

Kenneth H. Lai [a,*], Maxim Topaz [a,b], Foster R. Goss [c], Li Zhou [a,b,d]

[a] Division of General Internal Medicine and Primary Care, Brigham and Womens Hospital, Boston, MA, USA
[b] Harvard Medical School, Boston, MA, USA
[c] Department of Emergency Medicine, University of Colorado, Aurora, CO, USA
[d] Clinical Informatics, Partners eCare, Partners HealthCare System, Boston, MA, USA

ARTICLE INFO

Article history:
Received 30 September 2014
Revised 14 March 2015
Accepted 16 April 2015
Available online 24 April 2015

Keywords:
Electronic health record
Named entity recognition
Natural language processing
Spelling correction

ABSTRACT

Accurate electronic health records are important for clinical care and research as well as ensuring patient safety. It is crucial for misspelled words to be corrected in order to ensure that medical records are interpreted correctly. This paper describes the development of a spelling correction system for medical text. Our spell checker is based on Shannon's noisy channel model, and uses an extensive dictionary compiled from many sources. We also use named entity recognition, so that names are not wrongly corrected as misspellings. We apply our spell checker to three different types of free-text data: clinical notes, allergy entries, and medication orders; and evaluate its performance on both misspelling detection and correction. Our spell checker achieves detection performance of up to 94.4% and correction accuracy of up to 88.2%. We show that high-performance spelling correction is possible on a variety of clinical documents.
© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Accurate medical documents are critical for safe patient care and effective inter-provider communication. Spelling errors and incorrect information can lead to medical errors in patient care, which could put the patient at significant risk of harm. For example, errors in breast imaging reports can affect the understanding of the reports and patient care [1]. In particular, the confusion of two medication names that look or sound alike can have disastrous consequences [2,3].

While much of the patient's electronic health record (EHR) is now documented in a structured format (e.g. entered through checkboxes, dropdown lists, or other means), other important data is still recorded only using free-text. These unstructured data (e.g. clinical notes, reports, and free-text entries) are valuable both for patient care and research [4], but spelling errors continue to be a challenge in order to use and process these data.

Most previous studies of spelling errors in EHRs have focused on medication orders. Efforts have been made to curb look-alike, sound-alike mistakes, for example, by using "Tall Man Lettering", which involves writing part of a drug's name in uppercase letters to visually distinguish one drug from another (e.g. PENTobarbital and PHENobarbital) [5,6]. Few studies have focused on spelling errors in free-text EHRs. Ruch et al. reported that about one spelling error in every five sentences were found in a medical corpus of discharge summaries, surgical reports, and lab results [7], and they also found error rates up to 10% in follow-up notes [8]. Recently, Liu et al. reported that the occurrence of the potential errors in a corpus of 55 million notes at Mayo Clinic was about 0.4% [9]. Zhou et al. examined 2412 hypoglycemic drugs entered using free-text for 2091 patients in the Partners ambulatory EHR, and found that 17.4% of these free-text orders contained misspellings [10]. Physicians were responsible for 45.5% of misspellings (while writing 58.7% of orders), while registered nurses and nurse practitioners caused 19.0% of misspellings, writing 20.2% of orders.

It is crucial for misspelled words to be corrected in order to ensure that medical records are interpreted correctly. In addition, for natural language processing tasks, such as information extraction and encoding, it is essential that misspellings are handled appropriately [11]. For example, mapping of free-text to coded concepts is typically performed by exact string matching to controlled vocabularies. However, if words are misspelled, the information contained within them is lost.

In this study, we develop an automatic misspelling detection and correction system suitable for all kinds of medical text. In addition to clinical notes, we also train and test our spell checker on free-text medication orders and allergy entries, which can lead to medication errors and cause harm to patients when misspelled.

* Corresponding author.
E-mail address: klai5@partners.org (K.H. Lai).

http://dx.doi.org/10.1016/j.jbi.2015.04.008
1532-0464/© 2015 Elsevier Inc. All rights reserved.

2. Background

Automatic misspelling detection and correction is a subject that has attracted great interest. In Kukich's comprehensive survey of spell checking techniques [12], she identifies three increasingly difficult sub-problems: nonword error detection, isolated-word error correction, and context-dependent error correction.

Nonword error detection techniques have tended to fall into two categories. In n-gram analysis, mostly used in optical character recognition systems, unusual sequences of characters are indicators of recognition errors [13]. More common in spelling correction systems is dictionary lookup: any word not in the dictionary is probably misspelled.

Most isolated-word spelling correction systems use some form of minimum edit distance to generate or rank suggestions. Damerau found that over 80% of spelling errors consist of one of the following operations: an inserted letter, a deleted letter, a letter substituted for another, or two letters transposed or switched [14]. The Damerau–Levenshtein edit distance is the count of how many of these operations are needed to transform one word into another.

Context-dependent error correction is used in instances where a correctly spelled word is replaced with another. These techniques make use of statistical language models [15] to detect ill-formed sequences of words [16].

2.1. Noisy channel spelling correction

More recent spelling correction systems have been based on the noisy channel model. The concept of a noisy channel in communication was introduced by Shannon in his seminal paper [17]. In the model, a signal (e.g. a sequence of letters) is generated by an information source according to a statistical process. However, before it reaches its destination, the signal may be distorted by noise, also modeled by a statistical distribution. It is generally not possible to recover the original message with certainty; however, the most probable message can be calculated from the source and noise models.

Kernighan, Church, and Gale introduced a spelling correction program based on the noisy channel model [18]. According to the model, the most probable correction $\hat{c}$ for some misspelled word m is $\hat{c} = \text{argmax}_c P(m|c)P(c)$. $P(c)$ is the probability of c being generated by the source, while $P(m|c)$ is the probability that some correct word c will be misspelled (distorted via noise) as m. In practice, $P(c)$ is estimated using the frequency of c in a training corpus, and $P(m|c)$ is estimated using the inverse of the Damerau–Levenshtein edit distance between m and c. Kernighan's program successfully corrected 87% of typos found in newswire text.

Toutanova and Moore extended the noisy channel model to consider not only the raw spelling of a word, but also its pronunciation, using a letter-to-phone model [19]. They were able to decrease their error rate by 23.8–46.8% compared to a purely letter-based model.

2.2. Spelling correction in the medical domain

Spelling correction has been applied to several different problems in the medical domain. Tolentino et al. looked at spelling correction on a corpus of vaccine safety reports [20]. They built a comprehensive dictionary containing both medical and general English words, and preprocessed the data using regular expressions to eliminate certain abbreviations. They then used modified versions of the edit distance to both generate and score corrections. They achieved a precision (or positive predictive value: the percentage of misspellings the spell checker identified that were also actual misspellings found by humans) of 47%, and a recall (or sensitivity: the percentage of misspellings humans identified that were also found by the spell checker) of 74%.

Crowell et al. looked at the problem of spell checking medical queries to improve information retrieval [21]. Instead of generating their own suggestions, they used the open-source GNU Aspell program [22]. Aspell generates possible corrections using the Metaphone phonetic algorithm [23] and sorts them according to their orthographic and phonetic edit distances. The Metaphone algorithm maps the misspelling to a code; words with the same or similar code are returned as suggestions. Crowell et al. found that performance was greatly improved by re-sorting the suggestions list based on the frequencies of the possible corrections, up to a top-suggestion accuracy rate of 76.2%.

Mykowiecka and Marciniak investigated automated spelling correction of Polish-language mammography reports, for the purpose of improving information extraction [24]. Whereas Kernighan [18] and Crowell [21] used only the word frequency (a unigram language model) to estimate $P(c)$, Mykowiecka and Marciniak used a bigram language model, which also considers the context of the word. Additionally, instead of compiling a dictionary from outside resources, they built their own from a manually corrected gold standard. 93.4% of the misspelled word types were successfully corrected; however, 57.6% of the correct words not in their dictionary were wrongly changed.

Patrick et al. corrected lists of misspellings found in a large corpus of clinical notes [25]. Their dictionary was compiled from several sources, including both previously corrected notes and outside resources, including the Systematised Nomenclature of Medicine–Clinical Terms (SNOMED CT) [26] and the Moby lexicon, a collection of public-domain lexical resources [27]. To generate suggestions, they used edit distance-based rules, ranking the suggestions using a trigram language model. Accuracy on each set ranged from 84% to 94%. After adding the language model, system accuracy dropped from 84.36% to 84.04% on one of their test sets. Patrick suggests that a possible reason for this is that using single word frequencies works very well already, and so adding other ranking methods is not likely to improve performance.

Finally, Ruch, Baud, and Geissbühler explored the use of named entity recognition (NER) to improve spelling correction [8]. Many mistakes made by automated spell checkers result from the erroneous correction of names, for which comprehensive dictionaries are generally not available. Using a rule-based NER system to detect names to be ignored, the authors reduced their false correction rate on a corpus of French-language surgery discharge summaries from 21.8–23.5% to 2.6–3.6%.

2.3. Rationale

Previous studies of medical spelling correction have been limited in their scope, focusing on short query-type data [21,25], or long note-type data [8,20,24], but not both. Furthermore, much work has been done using automatically generated misspellings [8,21], rather than naturally occurring ones. We believe this is problematic because the misspellings were generated according to Damerau's operations (the same ones used to score corrections). In addition, constructing a dictionary that includes all possible abbreviations and names is difficult, especially in the medical domain. Those studies that do consider this problem show that there is much work still to be done in this area.

Finally, most work on spelling correction in general has measured only the true correction of misspellings, not considering the false correction of correctly spelled words, either treating it as a separate problem or ignoring it altogether. Only Tolentino reports the performance of their system in detecting true misspellings, in terms of precision and recall [20].

In our study, we aimed to synthesize the above approaches and create a general medical-domain spell checker with good performance on a wide variety of types of text.

3. Methods

Our spell checker is based on the noisy channel model of Shannon [17], applied to spelling correction by Kernighan, Church, and Gale [18], considering both the orthography and phonetics of a word. We built a comprehensive dictionary from a wide variety of lexical resources, and used named entity recognition to prevent the mistaken correction of names. We trained and tested our spell checker on three different data sets found in electronic health records: free-text medication orders, allergy entries, and full clinical notes. Unlike most other studies on this topic, in addition to assessing the accuracy of misspelling correction, we also evaluated the quality of misspelling detection (using precision and recall).

3.1. Preprocessing and named entity recognition

Following the approach offered by Ruch [8], we used named entity recognition to avoid misclassification of person names as misspellings. However, while Ruch used a rule-based system, we took a machine learning approach, based on the Stanford NER [28]. The Stanford NER uses a conditional random field (CRF) to classify words as part of named entities or not. Wellner et al. used CRFs to achieve the best overall performance in the 2006 i2b2 challenge, using NER to de-identify medical records [29]. We trained a linear chain conditional random field (CRF) on the clinical note training set, annotated with a single entity type – "Person".

Because the clinical note data set contains very few misspellings (see Section 3.4), a missed name is much more likely to be incorrectly treated as a misspelling, than a misspelling is to be wrongly left alone. Therefore, we found it more important to capture all possible names than to make sure only names are identified. For this reason, we also used a pre-built 3-class model (considering only the Person class; Location and Organization were ignored), trained on newswire text, and included in the Stanford NER distribution [28]. We found that some named entities missed by our clinically trained system were identified by the pre-built system, and vice versa. Because we wanted to capture all possible names, we decided to combine the two systems in the following way. If either model tagged a word as a named entity, the word was ignored by the spell checker.

During training, we found a few patterns of errors that the spell checker made, such as attempting to correct email and website addresses. Based on these patterns, we therefore introduced regular expressions to clean the data prior to the NER or spelling correction steps. We also replaced certain commonly misspelled terms, such as replacing "alot" with "a lot".

3.2. Misspelling detection

According to Kukich [12], the problem of spelling correction can be split into the separate problems of error detection and error correction. We used dictionaries to detect misspellings: any word not in the dictionaries was a possible misspelling. Like others [20,21,25], we used the Unified Medical Language System (UMLS) SPECIALIST lexicon (2014 edition) as our medical terminology source [30]. We enhanced this with a number of terminologies included in the RxNorm drug lexicon (April 2014 monthly release): the World Health Organization's Anatomical Therapeutic Chemical classification system, Elsevier's Gold Standard drug database, the Micromedex RED BOOK, the Food and Drug Administration Structured Product Labeling, the Veterans Health Administration National Drug File (and its reference terminology), and RxNorm's own normalized drug name vocabulary [31]. We also included a list of previously seen abbreviations, and Aspell's official English dictionary [22].

The above dictionaries contain a total of 418,632 unique words. In addition, like Patrick [25], we created custom dictionaries from each set of gold standard training data (we added any correctly spelled words, including names and abbreviations, not in any of the other dictionaries). Lastly, if a misspelling appeared in one of the other dictionaries, we removed it manually.

3.3. Misspelling correction

Once misspellings were identified, we then obtained a list of suggested corrections for each misspelling using Aspell [22]. However, instead of using Aspells suggestion ranking algorithm, we developed our own scoring algorithm. Aspell ranks the suggestions according to their orthographic and phonetic edit distances. Our scores roughly correlate with $1/(P(m|c)P(c))$, so this is equivalent to maximizing $P(m|c)P(c)$ in the noisy channel model. Our system selects the correction with the lowest score, as long as it was below a threshold that varies by the length of the misspelled word. If the lowest score was above the threshold, the word was left unchanged: these cases were more likely to be correctly spelled words that are not in the dictionaries.

We used the Damerau–Levenshtein distance between the misspelling and the suggestion, both in terms of their orthography and their phonetics, to estimate $1/P(m|c)$. For our phonetic algorithm, we used a simplified version of the Double Metaphone algorithm (the successor to Metaphone) [32]. Specifically, most of the changes between the original Metaphone and Double Metaphone were designed for rare names (e.g. "cz" maps to the metaphones "SX", except in "wicz", where it maps to "K"); we did not include these in our algorithm. Empirically, we weighed the actual spelling twice as much as the phonetics, so that our "spell score" equals 2 times the orthographic edit distance, plus the phonetic edit distance. Crowell, who also used a "spell score" based on these edit distances (i.e. using Aspell's internal scoring system), experimented with varying the relative impact of this score by scaling it by a power of C [21]. While Crowell used a value of $C = 3$, we found that a value of $C = 2$ maximized performance on our training sets.

To estimate $1/P(c)$, we used the frequency of the correction in the training data. We used the same formula as Crowell: $1/(1 + \ln(\mathit{frequency}))$ [21]. The spell and frequency scores were then multiplied to give the final score.

A pseudocode description of our correction algorithm can be found in Appendix A.

3.4. Data

We used three different sources of data for training and testing. The first data set consisted of randomly selected clinical notes of patients who visited two primary care clinics at Brigham and Women's Hospital, dated between January 2010 and May 2014. These patients' notes were retrieved from a centralized clinical data repository across Partners HealthCare; therefore, the notes had a wide variety of formats, including inpatient and outpatient, structured and free-form patient records (e.g. clinic visit notes, discharge summaries, letters to patients, phone call summaries, lab results, and more). We believe that this data set represents the diversity of text generated in a medical setting. This set was divided into a training set of 275 notes, containing 106,668 words (where a "word" is defined as a sequence of non-blank, non-

punctuation characters separated by blanks or punctuation), 475 (0.45%) of which were misspelled, and a test set of 40 notes, containing 15,247 words and 78 misspellings (0.51%). In the test set, physicians and clinical psychologists authored 29 notes, which contained 63 misspellings in 13,617 words (0.46%). Registered nurses wrote 6 notes, which together contained only one misspelling in 861 words (0.12%). Physician assistants wrote 5 notes containing 769 words, four of which had no misspellings, but one of which had 15, for an overall rate of 1.95%.

The second data set was constructed from randomly selected free-text allergy entries taken from the Partners Enterprise Allergy Repository (PEAR). PEAR contains all text entered by clinicians in the allergy sections of Partners' EHR systems since May 1993 [33]. 2184 entries containing 6460 words and 307 misspellings (4.75%) made up the training set; the testing set consisted of 442 entries with 1380 words and 55 misspellings (3.99%).

The third data set was comprised of randomly selected free-text medication orders entered by clinicians through Partners' ambulatory EHR system between March and September 2010. 2,351 entries (402 misspellings out of 5,069 words–7.93%) formed the training set, while 392 entries (872 words and 59 misspellings–6.77%) formed the test set.

The error rate in our clinical note corpus is similar to the 0.4% found in Mayo Clinic's notes [9], while the error rates in the allergy entries and medication orders are much higher. We should note that a spell checker is available (although it does not highlight misspellings by default) when entering clinical notes in Partners' EHR, but not when entering allergies or medications.

Each data source was analyzed independently from the others (e.g. the allergy entry training set was not used to train the medication order spell checker, etc.) and performance measured.

3.5. Misspelling analysis

We conducted an analysis of the types of misspellings found in our test sets. First, we determined the percentage of misspellings that were of clinical terms. We defined a "clinical term" as a word that was neither a name, nor present in Aspells default dictionary. These are generally less common medications or other specialized medical vocabulary items. Clinical terms made up 28.2% of our clinical note test sets misspellings, 65.5% of the allergy sets misspellings, and 78.0% of the medication sets misspellings.

We also analyzed the types and numbers of edits needed to transform the correct words into the misspellings in each test set. Table 1 shows the percentages of edits in each category (insertions, deletions, substitutions, and transpositions), and Table 2 shows the minimum edit distances required (all four categories of edits were considered to have an edit distance of one). In all three test sets, at least 80% of the misspellings were edit distance 1 from the correct word; this is consistent with Damerau's observation [14]. These misspellings were most likely typos, rather than cases where the clinician had no idea how to spell the word. The allergy entry data set had the highest proportion of misspellings with edit distance 3 or more (9.1%), while the clinical notes had none. In all three sets, deletions were more common than insertions. Substitutions were more common in the allergy and medication sets, while transpositions were more common in the clinical note set.

Table 1
Types of edits needed to transform correct words into misspellings, test sets.

	Notes (%)	Allergies (%)	Medications (%)
Insertions	32.2	14.7	18.8
Deletions	46.0	44.0	30.4
Substitutions	5.7	36.0	44.9
Transpositions	16.1	5.3	5.8

Table 2
Minimum edit distances needed to transform correct words into misspellings, test sets.

	Notes (%)	Allergies (%)	Medications (%)
1	88.6	80.0	86.4
2	11.4	10.9	10.2
3+	0.0	9.1	3.4

4. Results

We evaluated our system's performance on misspelling detection and correction separately. For the task of misspelling detection, on the test sets, we report precision (the percentage of changed words that were actual misspellings) and recall (the percentage of actual misspellings that were changed), as well as their harmonic mean (F measure). On the training sets, we only report recall: because we added all the correctly spelled words in the training sets to our dictionary, all changed words are guaranteed to be misspellings (i.e. precision is 100%). For misspelling correction, we only considered the true misspellings that were changed, and report the accuracy – the percentage that were changed to the correct word. All judgments were made in comparison with the test set gold standards, manually corrected by a team consisting of a physician, two pharmacists, a postdoctoral researcher in medical informatics, and two pharmacy students. Uncorrected text was provided in the form of plain text files, which the team edited using the text editors of their choice.

Our system was able to process approximately 550 words per second, with a constant 2.4 s of time needed to load the named-entity recognition models, on a computer with a 3.3 GHz dual-core Intel Core i5-2500 processor. NER requires a maximum of 284 MB of RAM, while the remaining parts of the system use up to 47 MB.

The results on the training sets are shown in Table 3, and the results on the test sets are shown in Table 4. On the task of misspelling detection, our system achieved F measures ranging from 75.7% (on clinical notes) to 94.4% (on allergy entries) on the test sets. Correction accuracy on the test sets ranged from 78.1% to 88.2% and ranged from 90.6% to 92.1% on the training sets.

Table 3
System performance on three data sources, training sets.

	Notes (%)	Allergies (%)	Medications (%)
Recall	90.3	97.7	95.6
Accuracy	92.1	91.7	90.6

Table 4
System performance on three data sources, test sets.

	Notes (%)	Allergies (%)	Medications (%)
Precision	71.1	96.2	90.0
Recall	81.0	92.7	91.5
F measure	75.7	94.4	90.8
Accuracy	78.1	88.2	81.5

Table 5
Effects of rescoring the suggestion list using word frequencies and named entity recognition on system performance, clinical note test set.

	Precision (%)	Recall (%)	*F* measure (%)	Accuracy (%)
Aspell default	48.9	82.3	61.3	58.5
With frequency rescoring	59.3	81.0	68.4	78.1
With NER	71.1	81.0	75.7	78.1

4.1. Effects of frequency and NER

To measure the effects of each part of our methods on performance, we used Aspell's default settings as a baseline: the first suggestion returned in Aspell's suggestion list was used as the correction, and named entity recognition was not used. Results on our clinical note test set are shown in Table 5. The baseline achieved a detection precision of 48.9% and a correction accuracy of 58.5%. The baseline precision is similar to that found by Tolentino (47%) on vaccine reports [20], and the accuracy is similar to those found by Crowell (59.5% using Aspell and a comprehensive dictionary) [21] and Patrick (60–64%) [25] on lists of misspelled words.

After rescoring the suggestion list according to word frequency, we observed an 11% increase in precision, and a nearly 20% increase in accuracy. Crowell [21] and Patrick [25] also saw improvements of between 15% and 25% in accuracy after frequency-based resorting of their suggestion lists. Meanwhile, detection recall decreased slightly. For one misspelling, the baseline system previously selected an incorrect suggestion. After rescoring, none of the suggestions fell under the score threshold, so the misspelling was left alone.

We used two named entity recognition systems to detect names so that they would be ignored. We trained an NER system on the clinical note training set, which achieved a precision of 95.2% but a recall of 69.2% on the clinical note test set. The pre-built Stanford NER model reached a precision of 74.4% and a recall of 60.7%. Combining the two systems, we were able to capture 81.1% of the names in our clinical note test set.

Adding named entity recognition to our system resulted in an additional 12% improvement in precision. Ruch also found that using NER led to a large decline in their correction error rate (which included false corrections of names).

5. Discussion

Our system performed well on all three corpora on which it was tested. Performance in all categories was best on the PEAR allergy entries, followed by the medication orders, and finally the clinical notes. We found that rescoring the suggestion list using word frequencies led to both a notable increase in the precision of misspelling detection and an increase in the correction accuracy. Exclusion of named entities also resulted in an increase in the precision.

5.1. Clinical applications

Correction of spelling errors in real-time has the potential to impact not only the accuracy of medical documentation but also the clinical care of a patient. Clinicians' overwhelming preference is still to use free-text and only recently have EHRs integrated misspelling correction. Incorporating automatic spell checking, particularly in areas that are critical for patient safety or research (e.g., entry of a drug allergy or medication, entry of diagnosis or problem) has the potential to markedly improve the quality and accuracy of electronic medical records. This is particular true with patient allergies, which when correctly spelled, can be encoded to a standard terminology and used for clinical decision support or alert the clinician of any drug-allergy interactions. With an increased attention to adherence to evidence based pathways and capture of specific quality metrics, correct spelling and encoding of these entries will play an important role in creation of a safer healthcare system guided by automated data correction and clinical decision support.

5.2. Error analysis

Errors made by the spell checker resulted from a variety of causes. First, some misspellings were very complex; in some cases, the spell checker either selected an incorrect replacement, or if no suggestions fell under the score threshold, it did not correct the misspelling. An example is "Penethol" (misspelling of "Pentothal"). This problem occurred more often in the allergy and medication data sets, which contain a higher proportion of harder-to-spell medication names.

In our implementation of the noisy channel model, more frequent words have a higher probability than less frequent words. This is usually a good outcome, but occasionally results in errors where a frequently occurring word is incorrectly selected over the correct, but infrequent, word. For example, the misspelling "alxity" is corrected to the very common "anxiety" instead of the relatively rare "laxity". In contrast to the complex-misspelling errors, this problem occurred more often in the clinical note set, possibly due to the larger training set and wider range of frequencies.

Our dictionaries also contained a few misspellings of their own. These misspellings were not corrected in the test data, and sometimes appeared as suggestions for other misspellings. We removed all misspellings found in the training data from the dictionaries, but some erroneous words like "releif" and "vacine" remained.

5.2.1. False positives

We used named entity recognition to detect names of persons in the clinical note data set, which were then ignored by our spell checker. Some names were not recognized, however, and a few were subsequently corrected to other words.

A larger source of errors in the clinical note data set occurred from the mistaken correction of abbreviations. Our abbreviation dictionaries contain 11,829 distinct abbreviations, not including those in our other dictionaries or those previously seen in the training data. Even so, abbreviations such as "abnml" ("abnormal") and "PEERL" ("pupils equal and equally reactive to light") were wrongly changed.

The above problems were generally confined to clinical notes, which have a greater proportion of named entities and abbreviations than the other two data sets. In the medication set, there were a few instances in which correctly spelled words did not appear in any of our dictionaries or the training data and were wrongly changed. These included non-medication health supplements (e.g. "Fiberwise") and medications not found in the United States (e.g. "Foraseq").

The impact of these false positive errors can be seen in the system's precision. The lowest precision of 71.1% occurred in the clinical note data, while the highest precision of 96.2% was in the allergy entry data, with the medication orders intermediate between the two at 90.0%.

Of the previous studies done of misspellings in the medical domain, only three – Tolentino, Mykowiecka, and Ruch [8,20,24] – applied their spell checkers to complete texts, as opposed to

individual misspelled words. Of these, only Tolentino reports a precision (positive predicative value), of 47% [20]. Mykowiecka notes that of the correct word types not in the dictionary, 57.6% were wrongly corrected [24], while Ruch presents correction error rates of 2.6% and 3.6% after using NER, depending on the score threshold used [8]. However, we should note that our precision and Ruch's correction error rates depend on the number of misspellings in the data, not only on the number of false positives. Because Ruch artificially corrupted their data up to a misspelling rate of 13.6% (compared to roughly 0.5% in our clinical note test set), their true positives should be expected to greatly outnumber their false positives.

5.2.2. Real-word errors

Real-word errors, in which the misspelling is another correctly spelled word in an inappropriate context, made up a small portion of the set of mistakes. There were three real-word errors (3.8% of the misspellings) in the clinical note test set, one (1.8%) in the allergy entry test set, and none in the medication order test set. Examples include "pen" for "pcn" in a list of allergens, or "our" for "out" in the sentence "Discussed her values in the context of choosing to go *our* to celebrate 4th of July".

These real-word errors can affect the recall of our system. In the test sets especially, there are other factors that also affect the recall, but in our case, recall can be seen as a measure of how many misspellings can be identified using only nonword error detection. The lowest recall and highest prevalence of real-word errors were found in the clinical note data, yet in the training set, over 90% of misspellings were detected, and therefore less than 10% were real-word errors. This indicates that context-dependent misspelling detection and correction, using language models to treat such errors, would be of limited use. Patrick and colleagues found that adding a language model to a clinical note spell checker did not lead to an improvement in performance [25].

Overall, our system's results are comparable to Tolentino's system, which achieved a recall of 93% on the training data and 74% on the test data.

5.3. Limitations and future work

We are unable to properly compare our results to previous results. Neither the system implementations nor the data sets used in previous studies are publicly available. We tested our system on three different corpora in part so that for each previous study, one of our data sets will be at least somewhat similar. However, our differing evaluation metrics make even indirect comparisons problematic. In addition, the fact that many of the previous studies used data sets containing only misspellings means that they can be more aggressive in trying to correct misspellings, since they do not have to worry about changing correctly spelled words.

Our training sets contain between 300 and 500 misspellings each. Performance can be improved with the addition of more training data, which will add correct words and remove misspellings from our dictionaries, as well as refine our word frequency counts.

Our spell checker uses Aspell to generate suggestions. Therefore, our correction accuracy is limited by the coverage of Aspell's suggestion lists. Crowell found that the correct word was generated as a suggestion for 92.3% of their misspellings [21]. Future work may include developing our own suggestion search algorithm to create lists of possible corrections.

Our system also does not implement context-dependent correction methods. For clinicians, real-word errors are most important: errors such as the confusion of two medication names can result in serious consequences (non-real-word errors are less critical, as long as the words are not totally unrecognizable). Fortunately, we see that real-word errors were very rare in our data sets, and none involved the confusion of two clinical terms. From a system performance perspective, real-word errors are not common enough to make a significant difference. However, if performance reaches a very high level, the effect of real-word errors will be more significant compared to other errors, and then it may be worthwhile to use a language model to detect and correct these misspellings.

Future work may include studying the impact of misspellings and spelling correction on information extraction from medical text. Errors involving clinical terms, as well as those involving non-clinical terms that provide context (e.g. negation words), can negatively affect the performance of natural language processing systems. Correction of real-word errors may be useful for this task, since certain real-word errors (that replace clinical terms with other clinical terms) can result in both a false negative (since the correct term is not identified), and a false positive (due to extraction of the misspelling).

Future work may also include testing our system on other kinds of medical text, such as problem lists, or laboratory test orders. In addition, all of this study's data is from Partners' data repositories; we may evaluate our performance on data from other institutions and EHR systems in the future.

Finally, the system developed in this study is designed for automatic spelling correction. Fully automatic correction is useful for processing of very large corpora for research purposes, where manual verification of suggested corrections is infeasible. Another next step may be to modify our algorithms and dictionaries (e.g. to improve the detection recall rate) if some human input is needed, such as selection from a computer-generated suggestion list.

6. Conclusions

In this study, we developed a spelling correction system for medical text, based on the noisy channel model. We trained and tested our spell checker on three diverse data sets, and evaluated both the precision and recall of misspelling detection and the accuracy of misspelling correction, while maintaining performance comparable to previous studies of misspellings in the medical domain. Our methods may be used to improve coding and extraction of information from unstructured medical text.

Conflict of interest

None.

Acknowledgments

We would like to thank Neil Dhopeshwarkar, Devki Patel, Diane Seger, and Sarah Slight for their help creating our correctly spelled gold standard corpora. We would also like to thank Frank Chang, Jason Lau, and Joseph Plasek for their assistance in this research.

This study was funded by the Agency for Healthcare Research and Quality (AHRQ) Grant 1R01HS022728-01.

Appendix A

The following is a pseudocode description of our spelling correction algorithm.

```
function correct(misspelling) returns correction:
  suggestion_list = Aspell's suggestion list
  if suggestion_list is empty:
    return misspelling
  else:
    threshold = maximum allowable score (varies by length of misspelling)
    best_score = threshold
    misspelling_metaphone = get_metaphone(misspelling)
    for suggestion in suggestion_list:
      suggestion_metaphone = get_metaphone(suggestion)
      orthographic_edit_distance = Damerau–Levenshtein (D–L) edit distance
        between misspelling and suggestion
      phonetic_edit_distance = D–L edit distance between
        misspelling_metaphone and
        suggestion_metaphone
      spell_score = (2 * orthographic_edit_distance +
        phonetic_edit_distance)^ 2
      frequency = frequency of suggestion in training data
      frequency_score = 1/ (1 + ln(frequency))
      score = spell_score * frequency_score
      if score < best_score:
        best_score = score
        best_suggestion = suggestion
      if best_score < threshold:
        return best_suggestion
      else:
        return misspelling
```

The following is a partial description of the Metaphone algorithm. Implementations of both the Metaphone and Double Metaphone algorithms are available on the Aspell website at aspell.net/metaphone [22].

```
function get_metaphone(word) returns metaphone
  for character in word:
    if character is the first character and a vowel:
      metaphone += character
    else if character = 'B':
      if previous character != 'M' or character is not the last character:
        metaphone += character
    else if character = 'C':
      if previous character != 'S' or next character is not a front vowel
      or current character is the first character:
        if character is not the first character and next character = 'I'
        and character after that = 'A':
          metaphone += 'X'
        else if next character is a front vowel:
          metaphone += 'S'
        else if character is not the first character and previous
        character = 'I' and next character = 'A':
          metaphone += 'K'
        else if next character = 'H':
          if current character is the first character and character
          after next is not a vowel:
            metaphone += 'K'
          else:
            metaphone += 'X'
        else if previous character = 'C':
          metaphone += character
        else:
          metaphone += 'K'
    else if character = 'D':
      ...
```

References

[1] S. Basma, B. Lord, L.M. Jacks, M. Rizk, A.M. Scaranelo, Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription, Am. J. Roentgenol. 197 (4) (2011) 923–927.

[2] B.L. Lambert, Predicting look-alike and sound-alike medication errors, Am. J. Health-Syst. Pharm. 54 (10) (1997) 1161–1171.

[3] J. Aleccia, Look-Alike, Sound-Alike Drugs Trigger Dangers, 2010 <http://www.nbcnews.com/id/37386398/ns/health-health_care/t/look-alike-sound-alike-drugs-trigger-dangers/> (accessed September 2014).

[4] Ö. Uzuner, I. Solti, E. Cadag, Extracting medication information from clinical text, J. Am. Med. Inf. Assoc. 17 (5) (2010) 514–518.

[5] R. Filik, K. Purdy, A. Gale, D. Gerrett, Labeling of medicines and patient safety: evaluating methods of reducing drug name confusion, Hum. Fact. 48 (1) (2006) 39–47.

[6] D. Gerrett, A.G. Gale, I.T. Darker, R. Filik, K. Purdy, Tall Man Lettering, Final Report of the Use of Tall Man Lettering to Minimise Selection Errors of Medicine Names in Computer Prescribing and Dispensing Systems, Loughborough University Enterprises Ltd.

[7] P. Ruch, A. Gaudinat, Comparing corpora and lexical ambiguity, in: Proceedings of the Workshop on Comparing Corpora, 2000, pp. 14–19.

[8] P. Ruch, R. Baud, A. Geissbühler, Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record, Artif. Intell. Med. 29 (1) (2003) 169–184.

[9] H. Liu, S.T. Wu, D. Li, S. Jonnalagadda, S. Sohn, K. Wagholikar, P.J. Haug, S.M. Huff, C.G. Chute, Towards a semantic lexicon for clinical natural language processing, in: AMIA Annual Symposium Proceedings, 2012, p. 568.

[10] L. Zhou, L.M. Mahoney, A. Shakurova, F. Goss, F.Y. Chang, D.W. Bates, R.A. Rocha, How many medication orders are entered through free-text in EHRS?–A study on hypoglycemic agents, in: AMIA Annual Symposium Proceedings, 2012, p. 1079.

[11] W.R. Hersh, E.M. Campbell, S.E. Malveau, Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical records: a lexical analysis, in: AMIA Annual Symposium Proceedings, 1997, pp. 580–584.

[12] K. Kukich, Techniques for automatically correcting words in text, ACM Comput. Surv. 24 (4) (1992) 377–439.

[13] L.D. Harmon, Automatic recognition of print and script, Proc. IEEE 60 (10) (1972) 1165–1176.

[14] F.J. Damerau, A technique for computer detection and correction of spelling errors, Commun. ACM 7 (3) (1964) 171–176.

[15] P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai, Class-based n-gram models of natural language, Comput. Ling. 18 (4) (1992) 467–479.

[16] E. Atwell, S. Elliott, Dealing with ill-formed english text, Comput. Anal. Engl.: Corpus-Based App. (1987) 120–138.

[17] C.E. Shannon, A mathematical theory of communication, Bell. Syst. Tech. J. 27 (1948) 379–423. 623–656.

[18] M.D. Kernighan, K.W. Church, W.A. Gale, A spelling correction program based on a noisy channel model, in: Proceedings of the 13th Conference on Computational Linguistics, vol. 2, 1990, pp. 205–210.

[19] K. Toutanova, R.C. Moore, Pronunciation modeling for improved spelling correction, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 144–151.

[20] H.D. Tolentino, M.D. Matters, W. Walop, B. Law, W. Tong, F. Liu, P. Fontelo, K. Kohl, D.C. Payne, A UMLs-based spell checker for natural language processing in vaccine safety, BMC Med. Inf. Dec. Mak. 7 (1) (2007) 3.

[21] J. Crowell, Q. Zeng, L. Ngo, E.-M. Lacroix, A frequency-based technique to improve the spelling suggestion rank in medical queries, J. Am. Med. Inf. Assoc. 11 (3) (2004) 179–185.

[22] K. Atkinson, GNU Aspell, version 0.60.6.1 <http://aspell.net/> (accessed September 2014).

[23] L. Philips, Hanging on the metaphone, Comput. Lang. 7 (12) (1990).

[24] A. Mykowiecka, M. Marciniak, Domain-driven automatic spelling correction for mammography reports, in: Intelligent Information Processing and Web Mining, Springer, 2006, pp. 521–530.

[25] J. Patrick, M. Sabbagh, S. Jain, H. Zheng, Spelling correction in clinical notes with emphasis on first suggestion accuracy, in: 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining, 2010, pp. 1–8.

[26] International Health Terminology Standards Development Organisation, SNOMED CT <http://www.ihtsdo.org/snomed-ct/> (accessed September 2014).

[27] G. Ward, Moby Project <http://icon.shef.ac.uk/Moby/> (accessed September 2014).

[28] J.R. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by gibbs sampling, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005, pp. 363–370.

[29] B. Wellner, M. Huyck, S. Mardis, J. Aberdeen, A. Morgan, L. Peshkin, A. Yeh, J. Hitzeman, L. Hirschman, Rapidly retargetable approaches to de-identification in medical records, J. Am. Med. Inf. Assoc. 14 (5) (2007) 564–573.

[30] National Library of Medicine, The SPECIALIST Lexicon <http://lexsrv3.nlm.nih.gov/Specialist/Summary/lexicon.html> (accessed September 2014).

[31] National Library of Medicine, RxNorm <http://www.nlm.nih.gov/research/umls/rxnorm/> (accessed September 2014).

[32] L. Philips, The double metaphone search algorithm, C/C++ Users J. 18 (6) (2000) 38–43.

[33] G. Kuperman, E. Marston, M. Paterno, J. Rogala, N. Plaks, C. Hanson, B. Blumenfeld, B. Middleton, C.D. Spurr, R. Kaushal, T.K. Gandhi, D.W. Bates, Creating an enterprise-wide allergy repository at partners healthcare system, in: AMIA Annual Symposium Proceedings, 2003, p. 376.