Identification and Correction of Misspelled Drugs' Names in Electronic Medical Records (EMR)

Faiza Hussain and Usman Qamar
Department of Computer Engineering, College of E&ME, National University of Sciences and Technology (NUST), H-12 Islamabad, Pakistan

Keywords: Spelling Correction, Electronic Medical Record, EMR, Natural Language Processing, Text Mining, Information Retrieval, POST, Parts-of-Speech Tagging, Regular Expressions and Medical Text Processing.

Abstract: Medications are an important element of medical records but they usually contain significant data errors. This situation may result from haphazardness or possibly careless storage of valuable information. In either case, this misspelled data can cause serious health problems for the patients and can put their life at a major risk. Thus, the correctness of medication data is an important aspect so that potential harms can be identified and steps can be taken to prevent or mitigate them. In this paper, a novel and practical method is proposed for automated detection and correction of spelling errors in electronic medical record (EMR). To realize this technique, major relevant aspects is taken into consideration with the help of Parts-of-Speech tagging and Regular Expressions. The paper concludes with recommendations and future work for giving a new direction to the emendation of drug nomenclature.

1 INTRODUCTION

Processing medical texts is an emerging topic in natural language processing, especially in English. As it holds an immense amount of biomedical data which every person accumulates over a lifetime. A medical record can be divided into five major portions i.e. medical history, laboratory, diagnostic test results, patient's condition and therapeutic response, potential side effects, and drugs dosage and treatment notes (Wagner, et al., 1996). Prescribers may use this format as a procedure for their documentation but may use another format according to their own ease (Spooner, Linda, and Kimberly, 2013). Clinicians can also write additional information sometimes such as progress notes, or correspondence etc.

Nowadays, to ease the work of pharmacists, a variety of electronic systems are used to record health information of patients. A primary goal of these systems is to get simple yet important and complete data enough for a doctor's needs. There are numerous benefits of the systematic data collection such as it allows to regularly organize information of patients in a decent manner and also allows to transfer history and records of a patient to another pharmacist (Spooner, Linda, and Kimberly, 2013). Electronic medical record system (EMR) is one of the methods to automatically record health-related information of an individual that can be generated, collected, managed, and used by clinicians, researchers, managers, process- improvement teams, and decision-support systems (Wagner, et al., 1996).

Despite many benefits, general implementation of EMRs in the United States is very low; the latest study showed that only 4% of doctors use a fully functional EMR while 13%have a simple electronic system (DesRoches CM, et al., 2008).The typical cause is these clinical records are created in a rush without any proofing. Consequently, a large number of spelling errors are resulted especially a wide variety of data entry errors occur. These errors not only originate because of the complexity of the English language but also due to characteristics of the medical domain. (Siklósi et al., 2016) investigated the most frequent types of errors are the unintentional mistyping, grammatical errors, sentence fragments, and non-standardized abbreviations.

It is self-evident that the correctness of such information is important (Richard Pless, 2004). Also, many researchers concluded that accurate medication data is the need of the hour such as (Price, D., et al., 1986) found 70 percent omissions

Hussain, F. and Qamar, U.
Identification and Correction of Misspelled Drugs' Names in Electronic Medical Records (EMR).
In *Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016) - Volume 2*, pages 333-338
ISBN: 978-989-758-187-8

in dosages of patients and 46 percent were taking medicines that were not recommended by their doctor. (Monson and Bond, 1978) assessed 355 patients and also attained the same percentage of errors in dose i.e. 70%.

Thus, these errors support us for the improvement of EMR (medication) data as its accuracy is an important factor for patient's health. In this research, we proposed a methodology for the identification of errors in medication data and used them to increase the accuracy of medication data through correction of misspelled terms in Electronic Medical Records. The data accuracy describes data along two dimensions i.e. correctness and completeness (Wagner, et al., 1996). Correctness is the main concern according to the guidelines provided by Wiederholt and Perrault (Shortliffe, et al., 1990).

There are some issues that are considered while proposing a novel technique. Firstly, all the fields of an individual's record are not important, some fields are optional, and some are not useful in correcting of medication data. So, it is a prerequisite for choosing an appropriate subset of fields. Another issue is to understand the meaning of a medication record. In simple words, there are two kinds of useful data are recorded i.e. doctors instruct the patient what to do and what the patient is feeling and taking (Wagner, et al., 1996). Thirdly, the modern concepts of biomedical are different from the classical one. Thus, medical concepts are constantly changing (Nordenfelt and Lennart, 2013). Furthermore, it is difficult to identify and classify different medications as their names resemble each other (Wagner, et al., 1996) e.g. Pentobarbital and Phenobarbital.

The organization of the paper is as follows: In Section 2, Methodology and details of fundamental techniques are explained. Results are shown in Section 3 where dataset of the research is stated. Along with, discussion and analysis of the results are made in Section 4. In Section 5, Literature Review is presented. And finally, the paper is concluded in Section 6.

2 METHODOLOGY

The most effectual solution would be to expand the scope of medical records by not only considering drugs names but also the potentially related concepts i.e. disease, symptoms etc. in all dimensions to potentially identify correct medications.

2.1 Selection Criteria

Inclusion and exclusion criteria are illustrated below that made certain that only relevant work is considered.

2.1.1 Inclusion Criteria

Records those are considered to examine:

- Any patient consulted an authorized medical nurse, practitioner or a doctor.
- Any patient with a drug prescription with dosage and schedule.

2.1.2 Exclusion Criteria

Cleaning of entries is done through:

- Any patient having empty medication lists.
- Any patient who had been checked previously and no new amendment in prescription was made.
- Any medication with a brand name with the help of non-drug terms.
- Redundant prescriptions of the same patient found in the database.

2.2 Steps towards Solution

The complete flow of the proposed work is presented in Figure 1.

The main tasks are broken down into the following steps:

2.2.1 Data Preparation

For the preparation of data, first of all, records are extracted about different allergies and then it is analyzed. For this, NULL value and duplicate entries are removed through the use of array_unique and array_filter functions of PHP are used for the identification of particular required data.

Fields include date, time, reason, result, medication and instructions for multiple diverse patients. Each record is fetched without biasness.

Then all records fulfilling inclusion and exclusion criteria are saved in the database.

2.2.2 Pre-processing

Brown corpus's Parts Of Speech Tagging (POST) is a well-known method of relating a word with a particular part of speech in a given sentence. But tagging a word is not a straightforward process as multiple words can communicate in multiple ways.

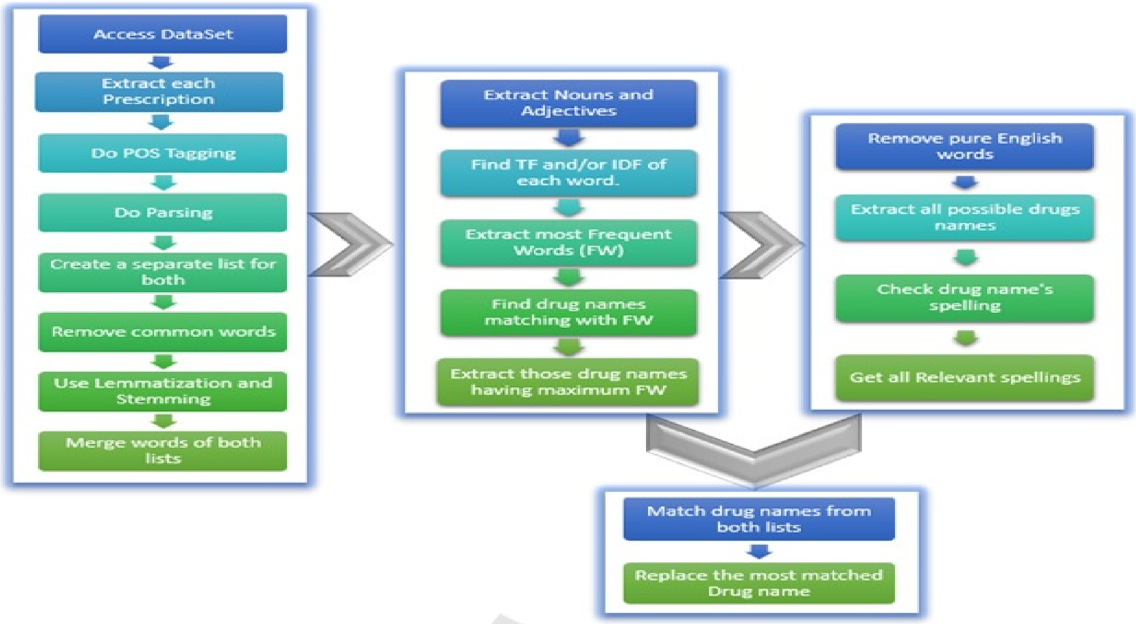
Access DataSet
Extract each Prescription
Do POS Tagging
Do Parsing
Create a separate list for both
Remove common words
Use Lemmatization and Stemming
Merge words of both lists
Extract Nouns and Adjectives
Find TF and/or IDF of each word.
Extract most Frequent Words (FW)
Find drug names matching with FW
Extract those drug names having maximum FW
Remove pure English words
Extract all possible drugs names
Check drug name's spelling
Get all Relevant spellings
Match drug names from both lists
Replace the most matched Drug name

Figure 1: Flow Chart of the Methodology.

Hence, there are multiple parts of speech against a single term due to ambiguity and complexity of English language. Actually, "what" the word is less significant than "how" it is applied (Manning, Christopher D., 2011).

Therefore, our first step is tagging of words according to brown corpus methodology. Presently, we have eight parts of speech in English: adjective, adverb, conjunction, interjection, noun, preposition, pronoun and verb along with many categories and their sub-categories. But our focus is on the nouns (NN) and the adjectives (JJ).

In working with clinical data, information is written in a certain format such as "medication dose unit-of-measure time-period" for example. "Aspirin 60 mg 1 tablet q.d." This information share some common features, that are patterns in the text but they are fairly complex (Richard Pless, 2004). To address them, regular expressions are used to systematically collect data. As regular expression is a simple way of describing a search pattern and extract information. The advantage of regular expressions is the incredible flexibility that they offer. But the challenge of regular expressions is to understand the patterns that you want to find but it gives assurance to collect all valuable information with ease.

2.2.3 Term Frequency (TF) Calculation

Term Frequency used in text mining to evaluate the importance of a word in a document or corpus. The significance of terms enhances proportionally to the number of times a word occurs in a document. As each document has different length, so, it may be possible that a word would come more times in long documents than in shorter ones. Therefore, the term frequency is frequently divided by the document size for normalization.

So, Term Frequency is used to obtain all available matching drugs' names from the internet:

TF= (Number of times appears in a medical record) / (Total number of terms in the medical record) (1)

2.2.4 Information Retrieval (IR)

Since many online medication dictionaries are available, such corpus will be used to extract all possible spellings of the anticipated misspelled drug through information retrieval. In IR, what more we need is to:

- Process information quickly from the web documents as there is a bulk amount of data available on the web. The indexed web contains 4.39 billion web pages at the time of writing this. And to retrieve useful information from it is not an easy task.
- Flexibility in matching patterns is required.
- Ranking in retrieved results. As hundreds and thousands are results are retrieved what we need is to have best results for our query so for this ranked retrieval is very important

(Manning, Raghavan, and Schütze, 2008).

Further comprehension will be done with the help of Lemmatization and Stemming of words by considering the fact that ‘particular nouns’ can’t be altered in any case.

2.2.5 Cosine Similarity

This metric is frequently applied to determine the similarity between two documents due to the presence of multiple similar words. Here, words are treated as vectors to calculate the normalized dot product. The result 0 depicts that two documents do not share any common term while other scores show some similarity. In our case, comparison of most relevant drug name extracted through IR will be performed through Cosine Similarity.

3 RESULT

***Dataset*:** In order to evaluate, an incorrect test set of clinical documents is necessary. For this purpose, we randomly selected 250 electronic medical records. All the data is available in tabular form.

$$\text{Precision} = \text{True Positives}/\text{Total Prescriptions} \quad (2)$$

The formula (2) was applied to calculate the overall accuracy of the system that how well it performed on correcting erroneous terms in the test set. ‘Actual’ spellings were compared with the ‘Corrected’ ones in each prescription and ‘Total’ were all possible records having no NULL entry in them while duplicate entries were considered as their quantity and dosage method vary person to person. Analysis of every single prescription was done manually by comparing ‘Corrected’ results with multiple medical dictionaries.

Below graph shows the accuracy of the novel approach by applying the formula of precision.

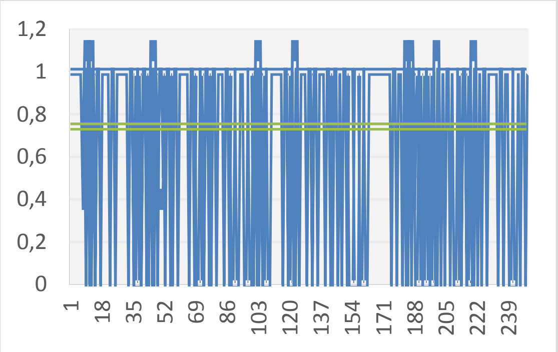

Figure 2: Comparison of Results.

Figure 2 demonstrates how precision rate varies. X-axis represents a total number of records while Y-axis represents the precision of retrieved results while horizontal line represents the threshold.

4 DISCUSSION

The essence of this approach is detailed comparison instead of simply replacing drug name with the first ranked suggestion. The evaluation is done on the basis of:

- Whether the technique provided all possible spellings of the word.
- Whether a word can be presented in multiple senses.
- Whether results are close enough to the actual spellings.

The approximate average precision is 0.742. From results described in Figure 3, we observe that overall results are better than previous non-context based approaches. However, in some situations, accuracy of results is reduced as the data is chosen randomly, therefore, it is imbalanced in nature that becomes a major cause of result’s graph decline. A potential limitation of this approach is that complete disease and symptoms information is required. Although, a new technique is proposed in this paper with the combination of Information Retrieval (IR), Regular Expressions and POST, but sometimes, the nature of English language leads us towards the uncertainty of applying it in few circumstances. Ambiguous and unclear terms may cause drifting from the actual intent when replacement of the word is made. However, it can be said that, through detailed analysis of grammatical issues, we will able to overcome issues and improve this technique.

5 RELATED WORK

The problem of misspelling in the medical domain has been addressed in various publications.

(Ruch et al., 2003) classified spelling error detection and correction into two types. The first type is word-based or context-free spelling correction which consists of errors that cannot be found in the dictionary, for example, handel instead of handle. The other type is context-based or context sensitive which deals with the correct word but invalid within the situation, for instance, worm body instead of warm body. Also (Kukich, Karen., 1992) divided the problem into three subgroups as (a) nonword error detection; (b) isolated-word error correction; and (c) context-dependent word correction. However, many proposed techniques rely on a lexicon-based approach.

The typical spelling correction system by (Levenshtein, Vladimir I., 1966) is based on minimum edit distance which ranks suggestion by the least number of inclusion, removal, replacement and reversal required to convert one string into the other.

(Turchin, Alexander, et al, 2007) identified incorrect words by comparing them to some predefined list of words, but this baseline method is extended by doing prevalence analysis, i.e. determining the frequency ratio of a word and its one edit distance alternatives in the corpus.

(Patrick, Jon, and Dung. 2011) used numerous knowledge bases of English clinical terms in addition to utilizing statistical methods.

Mass noun errors in English are solved by (Brockett et al., 2006), who focused on grammatical errors rather than on orthographical. Their work is related to the (Ehsan, Nava, and Faili, 2013) where the traditional SMT algorithm is used for spelling error preciseness. Though, in the approach, errors were initiated artificially.

(Siklósi et al., 2016) presented a new method for automatic correction in Hungarian clinical records by means of a SMT decoder. Due to the lack of a corpus normalized, a realistic aim was not fully achieved.

Some spelling suggestion tools such as Aspell and Gspell also exist in the English language for use and exploration. Aspell is a mixture of the Metaphone algorithm and near-miss strategy. While NGrams, metaphone, common misspellings, and homophone retrieval tools are present in Gspell. Candidates are evaluated by the Levenshtein edit distance, and similar ranked candidates are reordered by (Divita, G., 2003).

A frequency-based approach joining a medical dictionary configuration was built to improve recommendations of Aspell and Gspell by (Crowell et al., 2004). Turchin et al. used prevalence analysis for correction. (Senger, Christian, et al., 2010) made use of Aspell and user activities to analyze medication misspellings in a drug query system.

6 CONCLUSIONS AND FUTURE WORK

As we know that medical records play a significant role in everyone's life. So, the focus of this study is to illustrate the problems of Electronic Medical Records. Based on the causes of data errors, an effective improvement to the EMR would be to expand its scope to classify possible medications. In our paper, we presented a method to correct single spelling errors by concentrating more on 'how' and 'why' part of the searching instead of 'what', making a firm base for extending it to the correction of multiple errors as well. POST, Regular Expressions, and Information Retrieval played an important role in substitution. The overall accuracy of the system is a technically better than traditional techniques.

Even with EMR extensions, 100% accuracy can't be guaranteed, some minor error chances will remain (Wagner, et al., 1996). As the domain of searching is very gigantic; still more work is required to gather more accurate and close results. So, we have some future plans for including text parsing in it and will do the implementation of this technique for free-text clinical records to provide more ease to practitioners.

REFERENCES

Brockett, C., Dolan, W.B. and Gamon, M., 2006, July. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 249-256). Association for Computational Linguistics.

Crowell, J., Zeng, Q., Ngo, L., and Lacroix, E.M., 2004. A frequency-based technique to improve the spelling suggestion rank in medical queries. *Journal of the American Medical Informatics Association, 11*(3), pp.179-185.

DesRoches, C.M., Campbell, E.G., Rao, S.R., Donelan, K., Ferris, T.G., Jha, A., Kaushal, R., Levy, D.E., Rosenbaum, S., Shields, A.E. and Blumenthal, D., 2008. Electronic health records in ambulatory care—a national survey of physicians. *New England Journal of Medicine, 359(*1), pp.50-60.

Divita, G. 2003. Spelling Suggestion Tools (Gspell), *National Library of Medicine*. Available at http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/gSpell/current/GSpell.html.

Ehsan, N., and Faili, H., 2013. Grammatical and context□sensitive error correction using a statistical machine translation framework. Software: *Practice and Experience, 43*(2), pp.187-206.

Kukich, K., 1992. Techniques for automatically correcting words in text.*ACM Computing Surveys (CSUR), 24*(4), pp.377-439.

Levenshtein, V.I., 1966, February. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707-710).

Manning, C.D., Raghavan, P. and Schütze, H., 2008. *Introduction to information retrieval* (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press.

Manning, C.D., 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? *In Computational Linguistics and Intelligent Text Processing* (pp. 171-189). Springer Berlin Heidelberg.

Monson, R.A., and Bond, C.A., 1978. The accuracy of the medical record as an index of outpatient drug therapy. *JAMA, 240*(20), pp.2182-2184.

Nordenfelt, L., 2013. Identification and classification of diseases: Fundamental problems in medical ontology and epistemology. *Studia Philosophica Estonica*, 6(2), pp.6-21.

Patrick, J.D. and Nguyen, D., 2011. Automated Proof Reading of Clinical Notes. In *PACLIC* (pp. 303-312).

Perreault, L.E. and Shortliffe, E.H., 1990. *Medical Informatics: Computer Applications in Health Care*. Addison-Wesley.

Price, D., Cooke, J., Singleton, S. and Feely, M., 1986. Doctors' unawareness of the drugs their patients are taking: a major cause of overprescribing? *BMJ*, 292(6513), pp.99-100.

Richard Pless. 2004. "An Introduction to Regular Expressions with Examples from Clinical Data", *IL, SUGI Proceedings.* Available at http://www2.sas.com/proceedings/sugi29/043-29.pdf.

Ruch, P., Baud, R. and Geissbühler, A., 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine, 29*(1), pp.169-184.

Senger, C., Kaltschmidt, J., Schmitt, S.P., Pruszydlo, M.G., and Haefeli, W.E., 2010. Misspellings in drug information system queries: characteristics of drug name spelling errors and strategies for their prevention. *International journal of medical informatics, 79*(12), pp.832-839.

Siklósi, B., Novák, A. and Prószéky, G., 2016. Context-aware correction of spelling errors in Hungarian medical documents. *Computer Speech & Language,* 35, pp.219-233.

Spooner, L.M. and Pesaturo, K.A., 2013. The Medical Record. *Fundamental Skills for Patient Care in Pharmacy Practice*, p.37.

Turchin, A., Chu, J.T., Shubina, M. and Einbinder, J.S., 2007. Identification of misspelled words without a comprehensive dictionary using prevalence analysis. In *AMIA Annual Symposium Proceedings* (Vol. 2007, p. 751). American Medical Informatics Association.

Wagner, M.M. and Hogan, W.R., 1996. The accuracy of medication data in an outpatient electronic medical record. *Journal of the American Medical Informatics Association,* 3(3), p.234.

https://github.com/darylc123/diabetes-prediction/blob/master/trainingSet/training_SyncAllergy.csv.

http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/sphilip/cos.html.

http://www.tfidf.com/

http://www.regexbuddy.com/regex.html.

http://www.worldwidewebsize.com/