A Data-Driven Method of Discovering Misspellings of Medication Names on Twitter

Keyuan Jiang\textsuperscript{a,1}, Tingyu Chen\textsuperscript{a}, Liyuan Huang\textsuperscript{a}, Ricardo A. Calix\textsuperscript{a}, and Gordon R. Bernard\textsuperscript{b}

\textsuperscript{a}Department of Computer Information Technology and Graphics, Purdue University Northwest, U.S.A

\textsuperscript{b}Department of Medicine, Vanderbilt University, U.S.A

Abstract

Twitter, as a microblogging social media platform, has seen increasing applications of its data for pharmacovigilance which is to monitor and promote safe uses of pharmaceutical products. Medication names are typically used as keywords to query social media data. It is known that medication names are misspelled on social media, and finding the misspellings is challenging because there exists no \textit{a priori} knowledge as to how people would misspell a medication name. We developed a data-driven, relational similarity-based approach to discover misspellings of medication names. Our approach is based upon the assumption of the identical (or similar) association of a medicine with its effects whether the medication is correctly spelled or misspelled. With distributed representations of the words in tweets posted in recent 24 months, we were able to discover a total of 54 misspellings of 6 medicines whose indications containing headache. Our search results also show that Twitter posts with misspellings of codeine and ibuprofen can be more than 10\% of all the tweets associated with each of the medicines. Compared with the phonetics-based approach, our method discovered more actual misspellings used on Twitter.

Keywords

Information retrieval; Pharmacovigilance; Postmarking surveillance; Relational similarity; Twitter; Distributed word representation; Misspellings

1. Introduction

The primary goal of pharmacovigilance is to continuously monitor and promote the safe uses of pharmaceutical products, and an important task to achieve such a goal is to identify any suspected adverse effects related to uses of the products. Traditionally, drug adverse effects were collected through the spontaneous reporting systems (SPSs), but due to underreporting with the SPS systems, other data sources such as electronic health records (EHRs) have been sought. More recently, thanks to its prevalence and easy accessibility, social media has increasingly become an active data source for pharmacovigilance studies.
In a recent systematic investigation, Golder and colleagues [1] found in 16 databases a total of 3,064 publications related to “social media” and “adverse drug reaction” with an upward trend of publications for the last few years.

One of key challenges in using social media for pharmacovigilance is the variations of medication name: non-proprietary names (generic names) and proprietary names (brand names) as well as misspellings of medication names [2]. Misspelling can happen unintentionally if a person does not know how to exactly spell a medication name, or intentionally when there is a limited space to include all the characters of a medication name, which can happen on Twitter, a microblogging service, which limited each post to 140 characters.

Knowing the spelling of a medication name is important because it is typically used as a keyword to search social media posts or can be recognized by a named entity recognition (NER) tool. This is not an issue if the name is correctly spelled, but it becomes a challenge to deal with if it is misspelled name because there exists no a priori knowledge as to how people would misspell a medication name.

2. Related work

Pimpalkhute and colleagues are believed to be the first group to study misspellings of drug names on social media. They developed a phonetics-based approach to generate possible misspellings of any given medication name [3]. Their approach is a predictive method that first generates all possible misspellings of a medication name with the edit distance algorithm, and afterwards it filters these generated misspellings by the phonetic spelling algorithm. This phonetics-based method is able to find misspellings of the medicine names, but has limitations. First, it can generate an overwhelming number of misspellings if a medicine name is long. Taking acetaminophen as an example, there can be about 700 1-edit distance spellings and tens of thousands of 2-edit distance spellings before removing the duplicates. In addition, it fails to generate many misspellings people actually use on social media.

3. Method

3.1. Algorithm

Our approach to discover the actual misspellings of medicine names on Twitter is data-driven, relational similarity-based. Our method is based upon the assumption that a misspelled medicine is associated with its effects (or indications) in the identical or similar way that the correctly spelled medicine is associated with the same effects (or indications). For example, the association of ibruprofen (misspelling) with headache is similar to that of ibuprofen (correct spelling) with headache, indicative of a semantically similar relation. Mathematically, we have med\text{misspelled} : indication :: med\text{correct} : indication to represent similar relations. Because the indication at both sides is the same, they cancel out each other, and we have med\text{misspelled} :: med\text{correct} meaning that misspellings of a medication can be discovered from the terms semantically similar to the correct name. Recent advancement in distributed representations of word in vector space model (VSM) has made it possible to
find semantically similar terms through dense vectors: \( \text{vector(med}_{misspelled} \approx \text{vector(med}_{correct} \), which indicates that they are similar to each other.

In our method, we first collect Twitter posts with known effects or indications of a particular medication. After removing the stop words, punctuations, and non-English tweets, a VSM is built. A list of terms similar to the given medication is generated. All the known, correctly spelled medicines are removed from the list – note that many medicines share common effects or indications. The similar term list is sorted and inspected manually and checked with regular expressions for leading and trailing characters for possible misspellings. Later, the identified misspellings are checked against the drug names at drugbank.ca for final confirmation.

### 3.2. Experiment

Using word “headache” which is an indication of many medicines (e.g., Aspirin and Ibuprofen), we collected a total of 9,335,201 tweets posted from September 2015 to September 2017. After preprocessing, there were 6,555,535 remaining tweets. A word vector space model was built with Google’s word2vec tool [4]. A collection of medications whose indications contain “headache” was compiled, based upon the information available in SIDER 4.1, a resource of side effects of pharmaceutical products hosted at EMBL [5]. For each medication in the collection, a list of terms similar to the medication was generated with similarities of 0.20 or higher – mathematically this is done by checking the vector similarity which is measured by the cosine value between two vectors. Each similar term list was filtered with the corrected medicine names downloaded from the FDA’s National Drug Code (NDC) Directory 2. Afterwards, each filtered list was sorted and manually inspected and checked with regular expressions to select candidates of misspellings. Finally, misspelling candidates were compared with the drug names at drugbank.ca.

To compare our approach with the phonetics-based method, we first reproduced a program that implements the algorithm described by Pimpalkhute et al. [3] – we studied the authors’ Java code but a fragment seemed to be missing. The results of the comparison are presented in the next section.

### 4. Result

Out of 34 medicines whose indications contain “headache,” six were found to have misspellings. Table 1 lists the misspellings of the 6 medicine names.

Table 2 shows the statistics of misspellings corresponding to the 6 medicine names. The discovered misspellings are the results of our approach. The predicted misspellings are the results of the phonetics-based method. The common misspellings are the ones that were found by both phonetics-based and our approaches. In addition, we collected tweets both with the correct spelled medicine names and with their misspellings from the inception of Twitter to September 2017.

---

2. [https://www.fda.gov/drugs/informationondrugs/ucm142438.htm](https://www.fda.gov/drugs/informationondrugs/ucm142438.htm)
Examples of tweets containing misspellings are listed in Table 3.

5. Discussions

With our approach, we were able to discover misspellings without \textit{a priori} knowledge. This is to say, misspelled words can be discovered even if we do not know how and why they are misspelled. The ability to do so may be attributed to the power of distributed word representation that embeds rich semantic information in the word vectors and supports relational similarity. This is advantageous over the phonetics-based predictive method in that the phonetics-based method focuses on discovering the similar pronunciations of a medication but ours relies on the semantical relation.

As can be seen in Table 2, a total of 54 misspellings were discovered for 6 medications, and we were able to gather tweets with all these misspellings. We did not find many tweets containing the misspelling of Acetaminophen, but there were a noticeable number of tweets containing misspellings of ibuprofen and codeine, representing more than 10\% of each of the medications and making up a noticeable amount of tweets with the medicines. This indicates the importance of considering their misspellings when collecting Twitter data.

Although we do not know why our method only discovered the misspellings of 6 out of 34 medications whose indications include headache, we offer a few possible explanations. First, some misspellings may not appear a sufficient number of times in Twitter data in order to be included in the VSM. Second, headache may not be the major indication of some medications in this study. Finally, in this research, only single word medication names were considered, and this may have left out the tweets with two or more word medication names.

Misspellings discovered by our approach were actually posted by Twitter users. There is a substantial disagreement between the misspellings discovered by our approach and the ones predicted by the phonetics-based approach. Out of 54 discovered misspellings, only 13 were predicted by the phonetics-based approach (Table 2). This may indicate that the predictions generated by the phonetics-based approach are far from the reality and being useful. It is observed that many tweets contain the plural form of medication names, and we considered the plural form of a medication name correct spelling. This is true if stemming is applied in processing the textual data.

We acknowledge that this study only used a very common indication (headache) and only tweets posted in recent 24 months and containing the word headache were collected and analyzed. Ideally, tweets with all the indications and effects and posted from the inception of Twitter should be queried in order to discover all possible misspellings of a particular medication.

6. Conclusion

We developed a data-driven, relational similarity-based approach to discover misspellings of medication names on Twitter. Our method is based on the assumption of the identical (or similar) relation of a medication with its effects whether the name is correctly spelled or not, and was able to effectively discover medication misspellings actually used by Twitter users.
We believe that our approach can be extended to discovering misspellings of all the medications, and can ultimately assist in collecting more relevant tweets for pharmacovigilance. It is conceivable that our method can also be applied to many other types of health surveillance based on social media where misspelling can happen.

Acknowledgments

Authors wish to thank anonymous reviewers for constructive comments. This work was supported by the National Institutes of Health Grant 1R15LM011999-01.

References

1. Golder S, Norman G, Loke YK. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. British journal of clinical pharmacology. 2015; 80(4): 878–888. [PubMed: 26271492]
2. Sloane R, Osanlou O, Lewis D, Bollegala D, Maskell S, Pirmohamed M. Social media and pharmacovigilance: a review of the opportunities and challenges. British journal of clinical pharmacology. 2015; 80(4):910–920. [PubMed: 26147850]
3. Pimpalkhute P, Patki A, Nikfarjam A, Gonzalez G. Phonetic spelling filter for keyword selection in drug mention mining from social media. AMIA Summits on Translational Science Proceedings. 2014:90–95.
4. Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient estimation of word representations in vector space. Proceedings of Workshop at International Conference on Learning Representations; 2013.
5. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic acids research. 2015; 44(D1):D1075–D1079. [PubMed: 26481350]
Table 1

Discovered misspellings of 6 medicine names

| Medicine   | Misspellings                                                                 |
|------------|-----------------------------------------------------------------------------|
| Acetaminophen | acetomenophen                                                               |
| Aspirin    | aspirine, asprine, asperine, asperines, aspin, asperin, asprin, assprin, aspririn |
| Codeine    | codine, codiene                                                              |
| Ibuprofen  | ibprofen, ibuprofen, ibuprofin, ibprophen, ibprofen, ibproefen, ibprofeen, ibprofens, ibprofen, ibprofen, ibprofin, ibprofien, ibproufen, ibprofin, ibprofens, ibprofene, ibprophens, ibprofuens, ibprouphens, ibprofue, ibprophen, ibproifen, ibprofin, ibprofens, ibprofene, ibprophen, ibpro phen, ibproufen, ibprofhene, ibprofenen |
| Naproxen   | niproxin, neproxen, naproxin, neproxin                                      |
| Sertraline | sertaline                                                                    |
Table 2

Statistics of misspellings of 6 medication names.

| Medicine    | No. of Misspellings Discovered | No. of Misspellings Predicted | No. of Common Misspellings | No. of Tweets w Correct Name (2006 – 2017) | No. of Tweets w Misspellings (2006 – 2017) |
|-------------|--------------------------------|--------------------------------|-----------------------------|-------------------------------------------|--------------------------------------------|
| Acetaminophen | 1                              | 158                            | 0                           | 160,908                                   | 194                                        |
| Aspirin     | 9                              | 99                             | 2                           | 1,354,690                                 | 96,938                                     |
| Codeine     | 2                              | 103                            | 1                           | 2,407,864                                 | 287,560                                    |
| Ibuprofen   | 37                             | 126                            | 8                           | 1,628,993                                 | 183,931                                    |
| Naproxen    | 4                              | 108                            | 2                           | 83,317                                    | 4,296                                      |
| Sertraline  | 1                              | 143                            | 0                           | 31,680                                    | 819                                        |
Table 3  
Examples of tweets with misspelled medicine names

| Medicine    | Misspelling | Tweets                                                                 |
|-------------|-------------|------------------------------------------------------------------------|
| Acetaminophen| Acetomenophen| *that’s right it would have! I git a great deal at Target on some acetomenophen… so much better than aspirin* |
| Aspirin     | Asprine     | *I took asprine for my cramps earlier and now I need them for my headache. :(* |
| Ibuprofen   | Ibprefin    | *A nice blend of ibprofin for my fever. And melatonin and Benadryl for sleep. Let’s see how this goes.* |
| Codeine     | Codiene     | *codiene will hopefully knock me out and take away my cold #pleaseallah* |
| Naproxen    | Neproxin    | *are you sure the side effects of neproxin is worse that ibuprophen* |
| Sertraline  | Sertaline   | *Anyone else feel sleepy in the afternoons whilst taking #sertaline? Bloody nightmare* |