Semantic Similarity of Side Effect and Indication Relations of Drugs Inferred from Neural Embedding[1]

Keyuan Jiang[§], Tingyu Chen[§], Liyuan Huang[§], Gelareh Karbaschi[§] and Gordon R. Bernard[¶]

[§] Purdue University Northwest, Hammond, IN 46323, USA
[¶] Vanderbilt University, Nashville, IN 37232, USA
kjiang@pnw.edu, chen2694@pnw.edu, huanglydd@gmail.com,
gkarbasc@pnw.edu, gordon.bernard@vanderbilt.edu

Abstract. Patient-reported information on medication-effect experience can contribute to pharmacovigilance, and nowadays patients share their experience on social media which have been investigated for an alternative data source. Extracting the relations between pairs of medication-effect terms from social media data is a challenging task, but inferring the medication-effect relations from known (base) relations using the neural embedding technique seems to be a promising solution. This study aimed at understanding how the similar semantics is carried over from the base relations to inferred relations in the neural embedding of Twitter data. From a set of 99 randomly chosen inferred medication-effect relations whose associated tweets were manually annotated, we observed that the accuracies of having the inferred relations with the similar semantics to the base relations are 0.586 for medication-side effect relations and 0.688 for medication-indication relations. This demonstrated the utility of inference through relational similarity based upon neural embedding technique.

Keywords: Medication-effect Relations, Semantic Similarity between Relations, Neural Embedding

1 Introduction

Pharmaceutical products are widely used in modem medical practices, and it is known that they may have unwanted side effects on human subjects. Typically, some side effects are identified in pre-market clinical trials, while others are observed after the medications are put on the market. Some side effects can cause harmful effects to patients, while others may generate effects with benefit of therapeutic treatments of unintended symptoms, syndromes or diseases. Reporting of discovery of medication effects may come from physician's notes which can be kept in electronic medical records or from published literature of clinical research. Only those effects of adverse nature are reported to regulatory agencies mandatorily by manufacturers and voluntarily by healthcare professionals and consumers.

Patients are the consumer of the pharmaceutical products and they have the first-hand experience of medication effects. However, their venues of reporting side effects are very limited albeit the voluntary nature of reporting adverse events to regulatory agencies. Knowing medication effects directly from consumers of pharmaceutical products can help advance medical sciences and improve healthcare. Studies show that information reported by patients is different than that by healthcare professionals, in terms of better understanding of adverse experience, better explanation, and more detailed information [1].

The emergence of online social media provides a platform where patients can easily share experiences including the ones related to medication effects. Various studies have been conducted in leveraging social media data for possible use in pharmacovigilance. In 2015, Golder and colleagues collected over 3,000 published articles on investigating social media data for pharmacovigilance [2]. Among the published efforts, much was focused on identifying expressions of adverse events in social media text data or pairs of medication and effect, but little has been done in understanding how the identified effects are related to the medications within the same context. Understanding such relations can help generate hypotheses that may discern the association between the medications and effects, thus enhancing our understanding of medication effects.

However, identifying the relations between a pair of words (medicine and effect in our case) is a challenging task in natural language processing (NLP). Posts on general purpose social media, Twitter in particular, do not necessarily follow the spelling and grammatical rules, making methods and tools, such as SemRep [3] and dependency parsing [4], designed for formal writing, behave unsatisfactorily.

Inferring potential medication effect relations through the use of relational similarity, which reasons for less known or unknown relations from known relations, seems to be a promising approach. This approach does not require formal writing, and it bases upon the similarity of the relations expressed in the text. For example, to understand any potential medication-effect relations of Humira (adalimumab), one may uncover the potential relations by inferring (reasoning) from similar known relations of medicines other than Humira.

Development in neural embedding of word representations demonstrated state-of-art results in discovering similar relations between word pairs [5, 6], based upon the similarities known as linguistic regularities or relational similarities in that the similarities are between relations [5-8]. Neural embedding is a technique of generating vector representations of text by learning from a large corpus of unlabeled data, and vectors are the input weights of a neural network and embed the semantic and syntactic information from the context.

However, as described in [5, 7], relational similarities can be computed by simple vector operations: offset of two vectors and cosine similarity between the two offset vectors. The mathematical operations on the vectors do not intuitively demonstrate how the similar semantics is carried over or inferred from the known (base) relations to the less known or unknown relations. In this study, we seek to understand how relational similarities of medication-effect relations handles the semantic similarities from the neural embedding of Twitter data. The outcome will help determine the utility of inferring potential medication-effect relations from neural embedding of text data.

2 Related Work

Research of semantic similarity has mainly been focused on medical concepts/terms rather than relations. Pakhomov and colleagues [9] developed a reference standard of medical terms annotated by 8 medical residents, and their results indicated the existence of a measurable mental representation of semantic relatedness between medical terms which is distinct from similarity and independent of the context. Leveraging the neural embedding generated by Google's word2vec[2], Zhu and colleagues [10] investigated semantic relatedness and similarities of biomedical terms by examining the effects of recency, size and section. Fathiamini and colleagues investigated discovery of therapeutically relevant drug-gene relationships from unstructured text of Medline abstracts [11], and their results demonstrated better performance of the method of relational similarity than that of attributional similarity.

3 Method

In this research, we first infer medication-effect relations through relational similarity from neural embedding of Twitter data, and later examine the semantic similarity between the base relations and inferred relations.

3.1 Relational similarity

If we have the knowledge of known medicine-effect relations, the task of inferring potential medication-effect relations becomes finding similar relations of the medicines of interest. For example, if we want to answer a question like: what is the word or phrase that is to *Adderall* in the same sense as *seizure* is to *Gabapentin*? Here, the relation between *Gabapentin* and *seizure* is known, and we wish to solve (or find) an effect of *Adderall* that has a relation similar to *Gabapentin-seizure* relation. If we denote *medicine*:*effect* as a medicine-effect relation, and use :: for similarity, the example can be expressed as

Gabapentin:*seizure* :: *Adderall*:?

Mikolov and colleagues [6] demonstrated that such task can be accomplished by simple algebraic operations of vectors embedding the words: offset and cosine similarity. The relation of a pair of words can be represented as the offset of the two word

[2] https://code.google.com/archive/p/word2vec/

vectors, and the most similar relation to the known one can be determined by choosing the relation with the highest cosine similarity to the known relation.

Therefore, we have

$$medicine_{base}{:}effect_{base} :: medicine_{potential}{:}effect_{potential} \tag{1}$$

to represent that base relation $medicine_{base}{:}effect_{base}$ and target (or inferred) relation $medicine_{potential}{:}effect_{potential}$ are similar. Our goal is to find $effect_{potential}$ of $medicine_{potential}$ such that their relation is most similar to the known relation $medicine_{base}{:}effect_{base}$. In the vector space model of neural embedding, where each term is a vector, (1) above becomes

$$v(medicine_{base}) - v(effect_{base}) \approx v(medicine_{potential}) - v(effect_{potential}) \tag{2}$$

which can be rearranged as

$$v(effect_{base}) - v(medicine_{base}) \approx v(effect_{potential}) - v(medicine_{potential}) \tag{3}$$

or

$$v(effect_{potential}) \approx v(effect_{base}) - v(medicine_{base}) + v(medicine_{potential}) \tag{4}$$

Therefore, the task of inferring medication-effect relations becomes finding effect vectors which are most similar to the any of $v(effect_{base}) - v(medicine_{base}) + (medicine_{potential})$. In implementation, we use all possible known relations $medicine_{base}{:}effect_{base}$ except those for $medicine_{potential}$ to infer potential relations for $medicine_{potential}$. Utilization of multiple base relations can help cover more linguistic variations of expressing the same relation, and increase the confidence of inference.

3.2 Semantic similarity

In NLP research, there exists a broad spectrum of relations such as class-inclusion, part-whole, contrast and cause-purpose, and they have been used in shared tasks such as SemEval [12]. Many of the relations are irrelevant to our interest of studying medication-effect relations. In medical and healthcare domain, there also exist a large number of relations. U.S. National Library of Medicine published a list of hierarchical semantic relations in its Unified Medical Language System® (UMLS®)[3], and many of the UMLS Semantic Relations pertain to medication-effect relations, such as *treats*, *causes*, *occurs_in*, *disrupts*, *exhibit*, and *produces*. An ambitious repository of semantic predicates[4] extracted from the sentences of all Medline citations based upon the UMLS Semantic Relations has been developed (and regularly updated) [13], reflecting the fact that there are various linguistic ways of expressing a single semantic relation.

In this study, we focus on two specific types of semantic relations: medication-side effect (SE) and medication-indication relations (IND). This treatment is based upon

[3] https://www.nlm.nih.gov/research/umls/META3_current_relations.html

[4] https://skr3.nlm.nih.gov/SemMedDB/index.html

the available data and facilitates our data analysis tasks. The known relations for base relations come from the SIDER database (more discussions below) which only contains the SE and IND relations. The SE relations may be considered for adverse effect relations whereas IND relations for beneficial effect relations.

3.3 Measuring semantic similarity

Given the nature of our data, accuracy of the semantic similarities between base relations and inferred relations was measured using a method modified from the one described in [12]. If an inferred relation whose base relations are of the same type as itself – for example, an SE inferred relation and its base relations being SE relations, then the base and inferred relations are said to have a similar semantic relation. If an inferred relation is different from its corresponding base relations, then they are semantically dissimilar. Some inferred relations may be yielded from both SE and IND base relations, they are also considered to have a similar semantic relation. In the case where a single type of base relations yields inferences of both types of relations, they are still considered to have similar relations. This treatment helps measure the degree of semantic similarity between relations. If they are of the same type, they are semantically similar. Otherwise, they are dissimilar. The definition of our accuracy is as follows

$$Accuracy_t = I_t / B_t \qquad (5)$$

where t is the relation type, side effect (SE) or indication (IND); I_t is the count of inferred relations of type t, whereas B_t is the count of inferred relations whose base relations containing type t.

3.4 Annotation

To study the semantic similarity between the base relations and the inferred relations, a subset of inferred relations was randomly chosen and their tweets were annotated to determine their relation type – it was cost prohibitive and almost impractical to annotate tweets associated with all the inferred relations. If a tweet pertains to a medication-side effect relation, it is labeled as SE, and if it describes a medication-indication (or beneficial effect), it is marked as IND. If a tweet is neither about SE nor IND relation, it is labeled as "_" (underscore). A draft of annotation guideline was developed, and 100 tweets were first annotated based upon the draft guideline. The guideline and annotation were refined to establish a good standard of annotation, with which the rest of tweets were annotated and reviewed.

3.5 Data

Several sets of data were utilized in this study: a list of medication names, a corpus of unannotated tweets related to medications, a collection of known medicine-effect pairs, and the Consumer Health Vocabulary (CHV).

Twitter data were chosen for the rationale that in many instances, medication and effect(s) can be found in a single post, and hence a relation can be contained in an individual post. The Twitter data were gathered by searching for tweets with medication names as keywords. Two lists of top 100 drugs, by sales and by units, were obtained from drugs.com, and they were combined by removing the duplicates. The combined drug list was further expanded by including generic and brand names of these medications to facilitate querying related tweets.

A collection of unlabeled tweets, related to medication names discussed above, was retrieved through the use of a home-made crawler of twitter.com. Twitter has its own spam filter for its web interface, and Twitter posts gathered at twitter.com seem to be cleaner than those collected via Twitter APIs. In the summer of 2017, a total of 53 million tweets were collected with the time span between the inception of twitter.com (March of 2006) and the time of collection. After preprocessing which removes non-English, duplicate tweets, and tweets with a URL (which are considered mostly commercial), there were 12 million "clean" tweets. Phrases in the tweets were learned with GenSim[5] to treat multiple word terms as single unites. This set of tweets was further filtered by a list effect terms to ensure that each tweet contains at least a medication name and an effect expression. The effect term list was created by compiling Consumer Health Vocabulary (CHV) terms related to effects listed in the SIDER database. The resultant corpus of 3.6 million "clean" and filtered tweets served as the data for learning the neural embedding representation with word2vec.

SIDER is an online resource of side effects, hosted at the European Molecular Biology Laboratory (EMBL), and contains side effect information of marketed pharmaceutical products [14]. Two sets of data from SIDER were compiled. The first one contains all the terms for medication effects and their corresponding CHV expressions. The alignment of the SIDER and CHV terms was mapped through the UMLS CUIs (Concept Unique Identifiers). This set was used to filter out tweets without any effect expressions. The second set is a collection of lists of medication-effect pairs for each study medicine that exists in SIDER. In SIDER, a medicine has a list of side effects and a list of indications. This collection of medication-effect pairs served as the guidance for base relations, which are known, to infer potential medicine-effect relations.

The Consumer Health Vocabulary (CHV), a collection of words and phrases which consumers use to express health concepts and represent the mapping between the consumer expressions and technical terms used by healthcare professionals [15], was utilized to cover various ways of expressing concepts related to medication effects. Each individual effect concept in SIDER was expanded by including the corresponding CHV terms. Mapping between the SIDER terms and CHV terms was done by linking the identical CUIs. The expanded version of effect terms was then used to identify base relations and infer potential relations.

[5] https://radimrehurek.com/gensim/

4 Results

Our inference using the data described above generated a total of 5,182 potential medicine-effect relations from a collection of 3,184 unique base relations. Among inferred relations, 1,448 relations are known, meaning that they are found in the SIDER database, and 3,734 relations do *not* exist in the SIDER database. In inferring potential relations, the 3,184 unique base relations were utilized in a total of 78,369 times, indicating that many relations were used multiple times for inference for different medications.

To verify the semantic similarity, a collection of 100 inferred relations was randomly chosen using the random number generator at random.org – one of the 100 relations selected was dropped due to the ambiguity of the medication, leaving 99 inferred relations for annotation. A total of 3,492 unique tweets related to this set of inferred relational were annotated to determine the relation type of each inference.

Table 1. Statistics of relations.

	Side Effect (SE)	Indication (IND)
# Base relations	1,000	89
# Corresponding inferred relations	99	16

Table 1 summarizes the counts of relations for both base and inferred relations for SE and IND respectively. Among the base relations which are known, there are 1,000 SE relations and 89 IND relations, and among the inferred relations, there are 99 SE relations – every inferred relation contains the SE relation, and 16 IND relations.

Table 2. Counts of inferred relations from both base SE relations and base IND relations.

Inferred relations	Counts from base SE relations	Counts from base IND relations
SE only	**41**	3
IND only	26	**4**
Both SE and IND	**17**	**7**
Neither SE nor IND	15	2
Total	99	16

Shown in Table 2 are the counts of inferred relations from different base relations by inference type: SE only, IND only, both SE and IND, and neither SE and IND. Forty-one (41) inferred relations are SE only and their corresponding base relations contain SE relations. And three (3) inferred relations are SE and their corresponding base relations contains IND relations, indicating that the inferred relations are semantically dissimilar to the base relations.

Table 3. Accuracy of semantic similarities.

	SE	IND
# Actual inferred relations	58	11
# Corresponding inferred relations	99	16
Accuracy	0.586	0.688

Numbers in the first data row of Table 3 come from combining the boldfaced numbers of the corresponding column of Table 2. In other words, 58 = 41 + 17, representing the counts of inferred relations containing SE relations. Figures in the second row are the counts of inferred relations whose base relations are of the same type. For example, there are sixteen (16) inferred relations whose base relations contain IND relations. That is to say that there are supposed to be 16 inferred IND relations, but the results show that there are only 11 inferred IND relations. The accuracy of semantic similarity for IND relations is 0.688 (=11/16). Similarly, the accuracy of semantic similarity for SE relations is 0.586 (=58/99).

5 Discussions

For 99 inferred relations, all of them are associated with base SE relations, and only 16 of them are associated with base IND relations (Table 1). This implies that in an ideal situation, there would be 99 inferred SE relations and 16 inferred IND relations. Please note that for the 16 inferred relations, their corresponding base relations contain both SE and IND relations. Or in other words, both SE and IND base relations were used to draw the same inferred relations.

Either type of base relations does not always generate the same (correct) type of inferred relation. For SE relation inferences (Table 2), SE relations are the base relations for all 99 inferred relations, but only 41 inferences are solely SE relations, and 21 have a mixture of both SE and IND relations. Interestingly, there are 26 inferred IND relations which are based upon known base SE relations, and 15 are neither SE nor IND relations. The inferences from base IND relations are similar. Sixteen inferences are based upon known IND relations: 4 are solely IND relations, 3 are SE relations, 7 are a mixture of SE and IND relations, and 2 are neither.

If we combine inferred SE only and both SE and IND relations for SE relations (boldfaced numbers Table 2), then 58 out of 99 relations were inferred correctly, and for the IND relations, 11 out of 16 inferred IND relations were correct. This yields the accuracy of semantic similarity for SE relations (0.586) and that for IND relations (0.688), demonstrating the utility of the approach of relational similarity.

There may be two possible reasons why opposite (dissimilar) relation types are observed. First, the information of negation may not be embedded properly in the neural embedding, yielding inferences of opposite (dissimilar) relation type. Another possible reason may come from the fact that there does not exist a practical way to extract tweets by any particular relations because a relation is a vector of real values which do not corresponding to particular tweets. Instead, we extracted tweets associated with a particular relation by string match of the medication-effect pair. This may cause extraction of some unrelated tweets.

For the observation of inferred relations which are neither SE nor IND, this may be attributed to the nature of inference based upon the vector manipulations: offset and cosine similarity. They are pure mathematic operations whose results may not correspond to any relations in the data.

6 Conclusion

In this study, we investigated the accuracy of semantic similarity between the known base relations and inferred relations. Accuracies for both SE and IND relations demonstrated the utility of the approach using relational similarity to infer potential medication-effect relations, although further improvement will be needed to improve the accuracy and human annotation will be needed to remove false inferences.

Acknowledgement

Authors wish to thank anonymous reviewers for their critiques and constructive comments which helped improve the manuscript. This work was supported in part by the U.S. National Institutes of Health Grant 1R15LM011999-01.

Ethics Compliance

The protocol of this project was reviewed and approved for compliance with the human subject research regulation by the Institutional Review Board of Purdue University.

References

1. Härmark, L., Raine, J., Leufkens, H., Edwards, I. R., Moretti, U., Sarinic, V. M., & Kant, A. (2016). Patient-reported safety information: a renaissance of pharmacovigilance?. Drug safety, 39(10), 883-890.
2. Golder, S., Norman, G., & Loke, Y. K. (2015). Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. British journal of clinical pharmacology, 80(4), 878-888.
3. Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of biomedical informatics, 36(6), 462-477.
4. De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In LREC (Vol. 6, pp. 449-454).
5. Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chap-

ter of the Association for Computational Linguistics: Human Language Technologies (pp. 746-751).

6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. International Conference on Learning Representations (2013).
7. Levy, O., & Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In Proceedings of the eighteenth conference on computational natural language learning (pp. 171-180).
8. Turney, P. D. (2006). Similarity of semantic relations. Computational Linguistics, 32(3), 379-416.
9. Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T. and Melton, G.B., 2010. Semantic similarity and relatedness between clinical terms: an experimental study. In AMIA annual symposium proceedings (Vol. 2010, p. 572). American Medical Informatics Association.
10. Zhu, Y., Yan, E., & Wang, F. (2017). Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. BMC medical informatics and decision making, 17(1), 95.
11. Fathiamini, S., Johnson, A.M., Zeng, J., Holla, V., Sanchez, N.S., Meric-Bernstam, F., Bernstam, E.V. and Cohen, T., 2019. Rapamycin-mTOR+ BRAF=? Using relational similarity to find therapeutically relevant drug-gene relationships in unstructured text. Journal of biomedical informatics, 90, p.103094.
12. Jurgens, D. A., Turney, P. D., Mohammad, S. M., & Holyoak, K. J. (2012, June). Semeval-2012 task 2: Measuring degrees of relational similarity. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 356-364). Association for Computational Linguistics.
13. Kilicoglu, H., Rosemblat, G., Fiszman, M., & Rindflesch, T. C. (2011). Constructing a semantic predication gold standard from the biomedical literature. BMC bioinformatics, 12(1), 486.
14. Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2015). The SIDER database of drugs and side effects. Nucleic acids research, 44(D1), D1075-D1079.
15. Zeng, Q. T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. Journal of the American Medical Informatics Association, 13(1), 24-29.