International Journal of Data Science and Analytics manuscript No.
(will be inserted by the editor)

Drug prescription support in dental clinics through drug corpus mining

Wee Pheng Goh · Xiaohui Tao · Ji Zhang · Jianming Yong · Wenping Zhang · Haoran Xie

Abstract The rapid increase in the volume and variety of data poses a challenge to safe drug prescription for the dentist. The increasing number of patients that take multiple drugs further exerts pressure on the dentist to make the right decision at point-of-care. Hence, a robust decision support system will enable dentists to make decisions on drug prescription quickly and accurately. Based on the assumption that similar drug-pairs have a higher similarity ratio, this paper suggests an innovative approach to obtain the similarity ratio between the drug that the dentist is going to prescribe and the drug that the patient is currently taking. We conducted experiments to obtain the similarity ratios of both positive and negative drug-pairs, by using feature vectors generated from term similarities and word embeddings of bio-medical text corpus. This model can be easily adapted and implemented for use in a dental clinic to assist the dentist in deciding if a drug is suitable for prescription, taking into consideration the medical profile of the patients. Experimental evaluation of our model's association of the similarity ratio between two drugs yielded a superior F score of 89%. Hence, such an approach, when integrated within the clinical work-flow, will reduce prescription errors and thereby increase the health outcomes of patients.

Keywords adverse relationship · word embeddings · term similarity · personalised prescription · drug properties

Wee Goh E-mail: wee.goh@usq.edu.au
Faculty of Health, Engineering and Sciences, University of Southern Queensland, Australia

Xiaohui Tao (corresponding author) E-mail: xtao@usq.edu.au
Faculty of Health, Engineering and Sciences, University of Southern Queensland, Australia

Ji Zhang E-mail: ji.zhang@usq.edu.au
Faculty of Health, Engineering and Sciences, University of Southern Queensland, Australia

Jianming Yong E-mail: jianming.yong@usq.edu.au
Faculty of Business, Education, Law and Arts, University of Southern Queensland, Australia

Wenping Zhang E-mail: wpzhang@ruc.edu.cn
School of Information, Renmin University, China

Haoran Xie E-mail: hxie@eduhk.hk
Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong SAR

1 Introduction

With the increasing number of patients who take multiple drugs, adverse drug interactions resulting from the prescription of additional drugs is a major concern for both patients and health professionals. Hence, the ability of a system to predict drug interactions at point-of-care to reduce prescription error is important as an adverse event can lead to serious health consequences. A common cause of hospital admissions worldwide is adverse drug reactions, with incidence being as high as 24% [1]. Naturally, many such admissions could have been avoided if more care was taken in drug prescription, such as by considering the patient's drug allergies.

Consider a toy illustration where a patient with heart problems is taking *warfarin* and is allergic to *penicillin*. The dentist wishes to prescribe a painkiller *paracetamol* before extracting a tooth. A drug prescription support system that ensures *paracetamol* does not adversely interact with *warfarin*, and also is not similar to *penicillin*, would be very useful.

A recent work by [2] derives similarity within a drug-pair by comparing textual descriptions between DrugBank and Medical Subject Headings. Although the experiment reported favorable results with *metformin*, a drug for treating diabetes, the focus was on drug repositioning to treat other conditions. So far, many methods have been developed to extract information on drug interactions [3,4], but these methods do not integrate with the patient's medical history within the clinical workflow. Having identified this gap in existing research, this paper describes innovative approaches in

determining if a drug-pair is similar as well as using such information to support the dentist's prescription decision.

Based on the assumption that similar drug-pairs have a higher similarity ratio than that of dissimilar pairs, the aim of this paper is to explain and evaluate a novel method in predicting if a drug-pair is similar. While the three-tier framework has been used in the extraction of feature vectors from drug attributes in our previous work [5], this study expands on this by describing the word embedding approach within the predictive layer in finding the similarity of a drug-pair. The text corpus is trained on Google's word2Vec platform where word embedding models are generated and used for the extraction of feature vectors. A higher similarity ratio suggests a higher probability that the drug-pair is similar. By assessing the number of correct predictions, the performance of our model can be evaluated. Our work performs well compared to other methods of prediction, having a F score of 89% with drug properties gathered from textual data obtained through bio-medical sources.

Traditionally, chemical structures and drug targets are used to decide if a drug-pair is interactive or non-interactive. By using data mining and feature extractions from text corpus which describes the drugs in terms of their attributes like adverse interactions and side effects, this paper contributes significantly in the way that useful information on drug interactions can be obtained. Moreover, the practical use of data mining techniques in supporting the dental prescription of drugs will benefit the ongoing research on machine learning methods and knowledge management within the medical domain.

The rest of the paper is organised as follows: Section 2 discusses the related work in data mining and how our model differs in the way the drug-drug relationship is detected and deployed for use. The research problem is described in Section 3, while Section 4 explains how data is collected for the study. Section 5 then describes the approach in the predictive layer within the three-tier framework in predicting the suitability of a drug-pair. Section 6 outlines the experimental design with comparison to other baseline models. Finally, Section 7 discusses the results, and Section 8 presents the conclusions obtained.

2 Related Work

2.1 Drug interactions

Many systems have been developed which use data mining techniques to explore DDI. In fact, such techniques are evolving quickly to improve the accuracy of the experiments, although in most situations the results may not be sufficient to derive DDI [6]. A recent work by [7] attempts to determine DDI by identifying neutral candidates, negation cues and scopes from bio-medical text. Features extracted from these articles include linguistic definitions of negation, the position of the drugs discussed in the sentence and the linguistic-based confidence level of an interaction. Text mining techniques have also recently been used to predict protein interactions from bio-medical text [8]. Another common way of examining DDI is to extract relevant information from text. For example, Tari *et al*. [9] developed a method that combines text mining and automated reasoning to predict enzyme-specific DDI. Yan *et al.* also used text mining techniques to create features based on relevant information such as genes and disease names extracted from drug databases to augment limited domain knowledge [10]. These features were then used to build a logistic regression model to predict DDI. Another method to extract information on DDI from bio-medical text was proposed by Bui *et al.* [3]. DDI pairs are mapped according to their syntactic structure, and subsequently generated feature vectors are used to produce a predictive model which classify the drug-pair as interactive or non-interactive [3]. Although these studies use data mining methods to extract relevant information to predict DDI, they do not take into account the drug profile of the patient.

The crucial need to integrate the patient's medical profile with the knowledge obtained from data mining motivated us to embark on this study. Although our system is similar to that proposed by [4] in terms of using information from the patient, the unique approach adopted in this paper goes one step further in using such information to support the decision-making process for the dentist at point-of-care within the clinical work-flow. Moreover, the word embedding method is also adopted which uses features that relate the similarity of a drug-pair in terms of how closely the words are related to each drug of the drug-pair. This approach distinguishes from our earlier work where feature vectors were constructed based on term similarities within the drug corpus [11].

2.2 Clinical prescription support

Though current clinical decision support systems (CDSS) utilise knowledge bases in their design, they are of a limited nature, restricted to a particular kind of treatment plan. Even if it is focused on diagnosis of a common disease such as tooth decay, the knowledge base is not self-learning. For example, Park *et al.*'s [12] shared CDSS for dental fillings needs to be expanded to include clinical guidelines from global dental ontology in a real-time manner and integrated with local knowledge for the system to be self-learning. This involves semantic annotation which requires complex machine learning techniques [13].

Since dental ontology can enable CDSS to automatically update their knowledge bases with expensive expert medical and dental knowledge, it will be easier and cheaper to maintain the CDSS with the current expertise of dentists and the

latest knowledge in scientific and clinical evidence [13]. Additionally, the efforts of researchers and dentists can be harnessed easily through the semantic web interfaces provided by dental ontologies which act as a consensual representation of knowledge in the dental domain[13]. Good design and fast response time will increase the appeal of such a system. Not only must the CDSS response time be fast enough to appear helpful to dentists, the system must also fit into the clinical workflow at point-of-care, which commonly requires it to handle multiple disease and drug allergy information. Bhatia and Singh [14] have designed a CDSS that uses the Bayesian Network to suggest treatment plans based on different degrees of oral symptoms. Another system also uses the Bayesian Network as an inference engine to produce treatment options based on oral health history and risk factors [15]. Though these systems help the dentist to treat patients more confidently, their applications are restricted to tooth decay.

Moreover, there are no features on drug detection to facilitate prescription of medicine. Such features are important as the wrong prescription can result in adverse outcomes for the patients. A decision support system with prescription support customised to patients' individual drug profiles could help to reduce the incidence of adverse drug reactions as well as associated hospital admissions [16]. Such a system will also relieve the dentist from having to rely on search engines such as Google which suffer from low recall and precision rates [17], with results that may not be relevant to the needs of the dentists. Although there has been much research on DDI using different techniques, there is no system that uses DDI information to facilitate drug prescription within a CDSS, notwithstanding the absence of a complete source of information on potential DDI [18]. A CDSS that conforms to our recommendations of a personalised system, which takes into account the drugs the patient is taking and is allergic to, will contribute to the productivity and efficiency of dental treatment. Therefore, a CDSS which integrates with drug knowledge bases to identify adverse drug events and advise on drug suitability before prescription will appear helpful to the dentist. With timely and accurate DDI information embedded within a CDSS, more comprehensive treatment options can be made available to patients and practitioners, thus contributing to a more positive treatment experience and better oral health outcomes.

3 Research Problem

Dentists are trained to rapidly and accurately diagnose oral disease using data gathered from patient history, observations and images. However, the increase in volume and variety of data and the way these data are presented may overload the cognitive skills of even the most experienced dentists, leading to inefficient treatment planning or even wrong diagnosis.

In terms of drug prescription, new drugs are introduced regularly, and to assist dentists to function with more accuracy and timeliness at point-of-care, a personalised CDSS will inform dentists as to whether these drugs are appropriate to prescribe to the patient. There have been many studies investigating the presence of DDI, and this study does not aim to repeat these findings, but instead, to use such information in an evidence-based approach to ensure drugs prescribed by the dentist are safe based on the individual patient's profile. Hence, assuming the dentist has knowledge of the patient's medical conditions, the drugs the patient is currently taking and any drug allergies the patient may have, the goal of this research is to ensure that the drug the dentist is prescribing does not interact adversely with the drugs that the patient is currently taking and does not belong to the group of drugs that the patient is allergic to. The research problem is formalised as follows.

Problem 1 Let $D = \{d_1, d_2, ...d_i\}$ be the set of drugs that patient is currently taking; $D^{'} = \{d'_1, d'_2, ...d'_j\}$ be the set of prescription drugs dentist is considering to prescribe to patient; $D^- = \{d^-_1, d^-_2, ...d^-_k\}$ be the known set of drugs that patient is allergic to, and $D^\perp = \{d^\perp_1, d^\perp_2, ...d^\perp_l\}$ be the unknown set of drugs that patient is potentially allergic to. For each pair of $(d, d') \in D \times D'$, a function $f(y^-|(d, d'))$ is required to uncover the likelihood of $d' \in D^\perp$ considering d and D^-. □

The research goal then is to evaluate the performance of the model in predicting the similarity ratio of a drug-pair in a test environment. The better the performance of the model, the better its ability in assisting the dentist to prescribe a drug that does not adversely interact with the current drugs that the patient is taking.

In this study, similarity ratios are computed from feature vectors obtained in terms of term similarities and word embeddings.

4 Data Collection

Data used in this research was extracted from DrugBank, a resource that contains a comprehensive corpus of information relating to various properties of drugs. It is maintained in collaboration with the US Food and Drug Administration (FDA), which acknowledges that easy access to useful information can help the public protect and improve their health[1].

This corpus contains 6811 drug entries including 1528 FDA-approved small molecule drugs, providing free, independent, peer-reviewed, and up-to-date information at both consumer and professional levels. Each drug is described

[1] http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm212844.htm

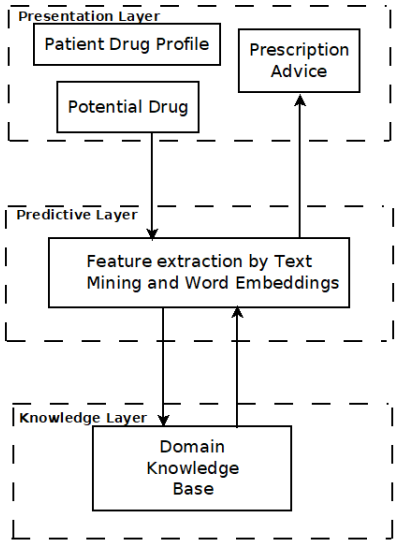

Fig. 1 Three-tier framework

from a different perspective to suit both patients (under the heading "Overview") and health professionals (under the heading "Professional") while information on side-effects are found under the heading "Side Effects". All this information is collectively stored in the drug taxonomy. References are also provided for drugs that are indexed to other databases such as the Kyoto Encyclopedia of Genes and Genomes, which is a collection of information on diseases and chemical substances useful for bio-informatics research and education. Many fields in this database are also hyperlinked to other resources including the RxList database which offers detailed and current pharmaceutical information on drugs useful for prescription and patient education.

In this study, the properties that are of interest are those related to interactions and side effects. Data related to "Overview", "Professional" and "Side Effects" can be extracted from this database. Text mining is conducted on each property in order to construct feature vectors for computing similarity between drugs. Take for example the drug *warfarin*. The "Overview" section explains in lay language the effects of *warfarin*, the dosage and other relevant advice such as the kinds of food to avoid while taking it. For the professional, under the property "Professional", more in-depth description of *warfarin* is provided such as the chemical structure, warnings and precautions as well as the recommended dosage for various symptoms to achieve maximum effect. The "Side Effects" page lists the major and minor side effects that are associated with *warfarin*.

5 Proposed Framework

As shown in Figure 1, the predictive layer obtains data from the domain knowledge base in the knowledge layer, and the drug profile of patients in the presentation layer. Feature vectors are then extracted and similarity ratios calculated to determine if the drug prescribed by the dentist is suitable for the patient.

The next two sections described the two approaches used within the predictive layer.

5.1 Term similarity

The quote by Tobler [19] that "Everything is related to everything else, but near things are more related than distant things" can be applied not just to spatial similarity but also to textual similarity. It is expected that a higher set of common terms are used to describe a pair of drugs that are similar in functions. Text mining is used to extract the term frequency for each drug. Given that each drug has k terms each with their *tf*idf* computed, the task of this model within the predictive layer of our framework is to construct feature vectors for each attribute of the drug. This feature vector comprises of a set of pairs of terms and their respective *tf*idf*. Thus, the feature vectors extracted during the data mining process for each attribute ("Overview", "Professional", "Side Effects") of the drug is respectively given as:

$$\overrightarrow{f_i^v} = \{(t_{1i}^v, v_{1i}^v), (t_{2i}^v, v_{2i}^v), ...(t_{xi}^v, f_{xi}^v)\} \quad (1)$$

$$\overrightarrow{f_i^p} = \{(t_{1i}^p, v_{1i}^p), (t_{2i}^p, v_{2i}^p), ...(t_{yi}^p, f_{yi}^p)\} \quad (2)$$

$$\overrightarrow{f_i^s} = \{(t_{1i}^s, v_{1i}^s), (t_{2i}^s, v_{2i}^s), ...(t_{zi}^s, f_{zi}^s)\} \quad (3)$$

Similarity between a drug-pair for an attribute is computed by comparing common terms within that attribute. For example, the similarity ratio $\mathcal{S}(d_j, d_k)$ within the attribute "Professional" for drug d_j and druq d_k was obtained by comparing these two feature vectors, with each feature vector sorted in descending order of the size of the term frequency:

$$\overrightarrow{f_j^p} = \{(t_{1j}^p, v_{1j}^p), (t_{2j}^p, v_{2j}^p), ...(t_{nj}^p, v_{nj}^p)\} \quad (4)$$

such that $v_{nj}^p >= v_{(n+1)j}^p$

$$\overrightarrow{f_k^p} = \{(t_{1k}^p, v_{1k}^p), (t_{2k}^p, v_{2k}^p), ...(t_{nk}^p, v_{nk}^p)\} \quad (5)$$

such that $v_{nk}^p >= v_{(n+1)k}^p$ where n is the size of each feature vector and for any n^{th} term, $v_{nj}^p == v_{nk}^p$.

The similarity ratios obtained from individual drug properties are used to decide if the drug-pair is similar. For example, if $s^v(i,j)$ is the similarity ratio between feature vectors $\overrightarrow{f_i^v}$ and $\overrightarrow{f_j^v}$ taken from drug property "Overview", then the number of similar drug pairs that were correctly predicted as similar can be found by counting the number of similar pairs. The number of true positives and true negatives can then be used to compute the F score. A drug pair is considered to be similar if $s^v(i,j)$ is above a threshold value α which occurs

at F^v_{max}. Thus

$$S^v(i,j) = \begin{cases} 0, \text{ if } s^v(i,j) < \alpha \\ 1, \text{ otherwise} \end{cases} \quad (6)$$

To obtain the overall gross similarity ratio, the performance of the individual drug property is taken into consideration. Depending on how accurately the similarity of each drug-pair is predicted, the similarity associated with each property is normalised by a factor as indicated by the F score. Thus if F^v_{max}, F^p_{max} and F^s_{max} is the maximum F score for drug attribute "Overview", "Professional" and "Side Effect" respectively, then the weight w_1 against the similarity ratio for "Overview" is given by:

$$w_1 = \frac{F^v_{max}}{(F^v_{max} + F^p_{max} + F^s_{max})} \quad (7)$$

$$(8)$$

The weights w_2 and w_3 against the similarity ratio "Professional" and "Side Effects" respectively can also be calculated in a similar manner.

Thus the overall similarity ratio $s(p,q)$ for drug-pair with feature vector p and q is given by:

$$s(p,q) = w_1 * s^v(p,q) + w_2 * s^p(p,q) + w_3 * s^s(p,q) \quad (9)$$

where
s^v is the similarity ratio for the drug property from "Overview",
s^p is the similarity ratio for the drug property from "Professional",
s^s is the similarity ratio for the drug property from "Side Effects".

5.2 Word Embeddings

One of the reasons for its popularity lies in the fact that analogical linguistic relationships among words can be easily discovered through word embeddings. Interest in word embedding has intensified with Mikolov *et al.*'s introduction of a simplified architecture, which eliminates the non-linear hidden layer, allowing training on much larger datasets than was previously possible [20]. Thus, for the word pairs a:b :: c:d (*a is to b as c is to d*), the vector for the word d, $\vec{d}$, can be obtained by finding the vector closest to the vector $\vec{c} - \vec{a} + \vec{b}$. Hence, in the ubiquitous proportional analogy *man is to woman as king is to queen*, the vector for the word *queen*, $\vec{queen}$, can be found from the linear vector operation $\vec{king} - \vec{man} + \vec{woman}$ [21].

As the aim is to find the extent of similarity between two drugs, it will be appropriate to obtain words most connected to the name of the drug. To achieve this, Google's word2Vec platform is used to generate pre-trained word embedding models from the same text corpus within the domain knowledge base as used for extracting feature vectors based on term similarity (see Section 5.1). One of the reasons for the popularity of word2Vec is that the output vector which is produced in numerical format can be easily understood by other deep learning networks making it very suitable for use in such works. For a detailed mathematical explanation behind word2Vec, refer to [22].

Once the text corpus has been trained by word2Vec, the output vector for any name of a drug can be conveniently obtained through built-in Java methods included in word2Vec. For example, given a keyword, the output vector comes in an array of numbers, and the number of such arrays depends on the number of nearest neighbors specified in the experiment. Refer to the definitions above for the feature vectors obtained for computing similarity.

6 Evaluation Design

We conducted an experimental evaluation to assess the accuracy and efficiency of the proposed method, testing the hypothesis that similar drug-pairs have a higher similarity ratio compared to that of dissimilar pairs.

6.1 Experimental Design

The training set consisted of sample drug-pairs that were either similar (true positive) or dissimilar (true negative) according to the drug taxonomy.

In the term similarity approach, feature vectors were based on terms in the text corpus and their respective term frequencies. By performing data mining on the corpus, models of feature vectors were constructed from the three properties "Overview", "Professional" and "Side Effects". By computing the similarity ratio of drug-pairs in the training set, the experiment aims to find how accurately the model can predict if the drug-pair is similar. The performance can be evaluated by varying the threshold value which determines if the drug-pair is similar or not similar.

The same dataset is used in the experiment using the word embedding approach. By using the skip-gram model from word2Vec[20], a predictive model was constructed for learning word embeddings from the raw corpus that described the properties of the drugs. Since the problem domain aims to extract related words to determine the extent of similarity from bio-medical text, word2Vec fitted well in the context of our experiment. Given a keyword, for example, the name of a drug, this method formulated a feature vector that best predicts a window of surrounding words that occur in some meaningful context. Such semantic similarity also conforms to the important criteria for selecting good word pairs [23].

With this model, word vectors were constructed by sending a keyword. In this experiment, a number of keywords associated with the nearest neighbor of the drug name was retrieved from the model. Similarity ratio between each set of vectors produced from the keywords could then be computed. In order to observe the behavior of this approach, the model was constructed with individual properties of the drug ("Overview", "Professional" and "Side Effects") while varying the number of nearest neighbors. In each model, word vectors were constructed from different combinations of keywords associated with the drug name. For example, if d_{11}, d_{12}, d_{13} were the three nearest keywords for a given drug d_1, a word vector would be obtained from the specified model by combining the three word vectors from the respective three keywords.

To obtain the weighted feature vector, the number of nearest neighbors that yielded the maximum F score for each attribute of the drug ("Overview", "Professional" and "Side Effects") was used to normalise the similarity ratio.

6.2 Performance Assessment Methods

Precision, recall and F score were used to evaluate the performance of our model. Precision indicated how accurately the model predicted drug-pairs as similar, while recall indicated how accurately similar drug-pairs were predicted. Accuracy was also used to measure the percentage of correct predictions combining both the similar and dissimilar predictions. Thus,

$$precision = \frac{TP}{TP+FP} \quad recall = \frac{TP}{TP+FN} \tag{10}$$

where TP is True Positive, FN is False Negative, and FP is False Positive. F score (see equation 11) was based on the precision and recall.

$$Fscore = \frac{2 * precision * recall}{precision + recall} \tag{11}$$

6.3 Baseline Models

Our work was evaluated against other works to highlight how adoption of this approach results in superior performance. The work of [9] predicted DDI by parsing bio-medical text for syntactic and semantic information on biological entities like induction and inhibition of enzymes by drugs. These relations were then mapped with the general knowledge about drug metabolism and interactions to derive the DDI.

Just like our work, DrugBank was also used by [10]. However, one of the methods in their preparation of data was to represent each drug by a vector of drug targets. The values in each vector are either 1 or 0, depending on whether the drug target is associated with the given drug. In our work, we chose to construct feature vectors from textual information related to the properties of each drug.

	Tari [9]	**Yan** [10]	**Proposed Model**
Aim	Discover drug interaction	Predict drug interaction (non-personalised)	Predict drug interaction (personalised)
Source	Drug Bank and MeSH	Drug Bank and MeSH	Drug Bank
Method	Combine text mining and reasoning approach based on biological entities	Compose feature vectors based on names of disease and genes	Create feature vectors from textual drug description
Accuracy	78%	69%	89%

Table 1 Comparison with baseline models

In terms of accuracy, which indicates the percentage of correct predictions taking into consideration both the similar and dissimilar predictions, our model yielded a superior performance of 89%. Table 1 summarises the accuracy of our model as well as the baseline models.

7 Results and Discussions

With the unique three-tier conceptual framework where knowledge is extracted from the knowledge base and delivered to the predictive layer, the ensuing results demonstrate the efficiency and robustness of our model. Not only is the algorithm able to compute the similarity of the drug-pair based on the hypothesis that a drug-pair is similar if the cosine similarity ratio between their frequency terms is high but such information can also be adopted as a decision support tool for the health professional in drug prescription.

7.1 Experiment using term similarity

By building feature vectors of terms and their corresponding *tf*idf*, the model predicts if the drug-pairs are similar. Table 2 shows the number of correct predictions by applying to our model the datasets or attributes "Overview", "Professional" and "Side effects".

	Overview	Professional	Side Effects
Similar (n=24)	19	13	23
Dissimilar (n=24)	19	23	4
Recall	0.79	0.54	0.96
Precision	0.79	0.93	0.53
F score	0.79	0.68	0.69

Table 2 Correct predictions based on different attributes of drug-pairs

By computing the similarity ratio between drug-pairs, their average values were obtained as a guide to set the threshold ϕ in order to maximise the F score. For a given value of ϕ, the number of correct predictions for the dataset that was supposed to be similar (true positives) and dissimilar (true negatives) was counted. If the similarity ratio was above ϕ, it was considered "similar", otherwise it was considered to be "dissimilar". For example, from the "Professional" attribute, there were 13 and 23 correct predictions from the similar and dissimilar datasets respectively.

ϕ	**Overview**	**Professional**	**Side Effects**
0.45	0.60	0.60	0.67
0.48	0.61	0.56	0.66
0.50	0.60	0.57	0.67
0.53	0.56	0.51	0.68
0.55	0.53	0.52	0.69
0.58	0.51	0.54	0.69
0.60	0.54	0.52	0.68
0.63	0.56	0.46	0.73
0.65	0.53	0.44	0.70

Table 3 F score at different threshold values of ϕ

Table 3 shows the F scores obtained for a range of values for ϕ, applied for each of the drug properties "Overview", "Professional" and "Side Effects". For example, a ϕ of 0.45 was used as a threshold to compute the recall, precision and F score for features gathered from the drug property in the "Professional" attribute as the maximum value of F score occurs at this value.

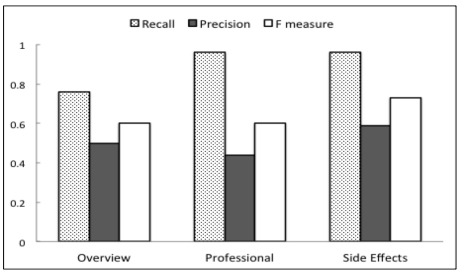

Fig. 2 Performance comparison against different drug properties

By using this threshold value that yields the maximum F score, Figure 2 compares the results achieved with drug properties gathered from "Overview", "Professional" and "Side Effects". As indicated in Figure 2, a recall rate of 96% was achieved from drug properties obtained from "Side Effects", showing that our model performed much better than other methods of prediction. In contrast, the work by [9] achieved 48.5% with predictions based on the inhibition properties of drugs in the knowledge base.

From the F score of each attribute, a weightage was computed in proportion to the respective F_{max}. In our experiment, F_{max} for "Overview" was 0.6 and the total F_{max} for the three attributes was 1.94, so the feature vector for "Overview" was weighted by a factor of 0.6/1.94 which was 0.32. The weights of the other attributes were computed in a similar manner. By combining the normalised feature vectors for all the three attributes, an aggregated similar ratio was obtained for each drug-pair.

	Predicted: Similar	Predicted: Dissimilar
Actual: Similar	17	13
Actual: Disimilar	7	49
F Score	0.63	

Table 4 Results with feature vectors normalised

	Predicted: Similar	Predicted: Dissimilar
Actual: Similar	24	3
Actual: Disimilar	15	33
F Score	0.73	

Table 5 Results from word embeddings method

In the same manner, different F score values were obtained at different threshold levels by counting the number of true positives and true negatives produced from the model. Table 4 shows the results based on the aggregated similarity ratio obtained from the normalised feature vectors. In terms of accuracy, the percentage of correct predictions combining both the similar and dissimilar predictions, our system achieved 76% compared to 69% where drug predictions were based on the relationship between drug targets [10].

7.2 Experiment using word embeddings

Table 5 shows the results of the experiment trained using the word embeddings approach. At different threshold values of ϕ, results are obtained and compared for the two different approaches. With the common data-set used for both methods, results show that the word embeddings approach performs better than the term similarity approach (Figure 3). This is expected as the former approach in computing similarity was to gather the term frequency by means of a bag of words. In the latter approach, the feature vectors used to find the similarity were obtained from closely related words.

To illustrate the conceptual framework of this study, the same model can be used to decide if the drug is suitable for prescription. Based on the overall similarity from the three properties of the drug-pair, the system can help dentists to

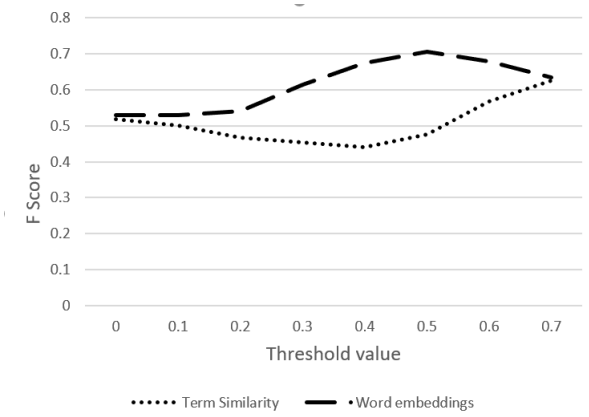

Fig. 3 Comparison with word similarity approach

ensure that the drug prescribed does not adversely interact with the drugs that patient is currently taking. This approach highlights the usefulness of our framework where knowledge generated can become useful to the user - in this case, as a decision support tool for the health professional. In future work, we will investigate the drug taxonomy for more complex semantic relations existing between drugs, for example, *neutral* and *advantageous*, and use a more comprehensive database for the CDSS.

8 Conclusions

This paper presents a novel idea in prescription support for the dentist by predicting the similarity ratio of a drug-pair using the feature vectors of a bio-medical text that describes the drugs. Term similarity and word embeddings approach are used to predict the similarity of a drug-pair. Empirical results show that our approach performs better than other approaches using a similar data set. Additionally, our design which incorporates the patient's medical condition can be readily implemented within the clinical work-flow of a dental clinic as it takes into account the drugs that the patient is currently taking and the drugs that the patient is allergic to.

Acknowledgements This research is partially supported by Glory Dental Surgery (Roxy Square) Pte Ltd, Singapore (*http://glory.sg*) and undertaken collaboratively with their panel of dentists. The authors would like to thank the reviewers, the handling editor and the Editor-in-Chief for their constructive comments on the manuscript. We would also like to thank Dr Elizabeth Goh (e-mail: eg@glory.sg) for enriching the authors' understanding of dentists' requirements and for proof-reading the manuscript.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. N. S. R. Lieber, E. Ribeiro, Adverse drug reactions leading children to the Emergency Department, Revista Brasileira de Epidemiologia 15 (2012) 265–274.
2. A. S. Brown, C. J. Patel, Meshdd: Literature-based drug-drug similarity for drug repositioning, Journal of the American Medical Informatics Association 27 (2016) 1–5.
3. Q. Bui, P. Sloot, E. vanMulligen, J. Kors, A novel feature-based approach to extract drug-drug interactions from biomedical text, BioInformatics 30 (23) (2014) 3365–3371.
4. A. Casillas, A. PÃlrez, M. Oronoz, K. Gojenola, S. Santiso, Learning to extract adverse drug reaction events from electronic health records in spanish, Expert Systems with Applications 61 (2016) 235–245.
5. W. P. Goh, X. Tao, J. Zhang, J. Yong, Mining drug properties for decision support in dental clinics, in: J. Kim, K. Shim, L. Cao, J.-G. Lee, X. Lin, Y.-S. Moon (Eds.), PAKDD 2017, Part 2, Springer International Publishing, 2017, pp. 375– 387.
6. H.-Y. Wu, C.-W. Chiang, L. Li, Text mining for drugâĂŞdrug interaction, Methods in molecular biology 1159 (2014) 47–75.
7. B. Bokharaeian, A. Diaz, H. Chitsaz, Enhancing extraction of drug-drug interaction from literature using neutral candidates, negation, and clause dependency, PLoS ONE 11 (10) (2016) 1–20.
8. A. Li, Q. Zang, D. Sun, M. Wang, A text feature-based approach for literature mining of lncrna-protein interactions, Neurocomput. 206 (C) (2016) 73–80. doi:10.1016/j.neucom.2015.11.110.
9. L. Tari, S. Anwar, S. Liang, J, C. Baral, Discovering drugâĂŞdrug interactions: a text-mining and reasoning approach based on properties of drug metabolism, Bioinformatics 26 (18) (2010) 547–553.
10. S. Yan, X. Jiang, Y. Chen, Text mining driven drug-drug interaction detection, 2013 IEEE International Conference on Bioinformatics and Biomedicine (2013) 349–354.
11. W. P. Goh, X. Tao, J. Zhang, J. Yong, Decision support systems for adoption in dental clinics: A survey, Knowledge-Based Systems 104 (2016) 195–206.
12. S. G. Park, S. Lee, M.-K. Kim, H.-G. Kim, Shared decision support system on dental restoration, Expert Systems with Applications 39 (14) (2012) 11775–11781.
13. P. R. Smart, M. Sadraie, Applications and uses of dental ontologies, in: Proceedings of the 2012 IADIS International Conference, 2012, pp. 499–504.
14. A. Bhatia, R. Singh, Using bayesian network as decision making system tool for deciding treatment plan for dental caries, Journal of Academia and Industrial Research 2 (2) (2013) 93–96.
15. M. Bessani, E. Lins, A. Delbem, C. Maciel, Construction of a clinical decision support system for dental caries management using BN, in: Brazilian Congress on Biomedical Engineering, 2014, pp. 517–520.
16. S. Dechanont, S. Maphanta, B. Butthum, C. Kongkaew, Hospital admissions/visits associated with drug-drug interactions: a systematic review and meta-analysis, Pharmacoepidemiology and Drug Safety 23 (5) (2014) 489–497.
17. Y. Cai, C.-m. Au Yeung, H.-f. Leung, Knowledge representation on the web, in: Fuzzy Computational Ontologies in Contexts: Formal Models of Knowledge Representation with Membership Degree and Typicality of Objects, and Their Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 15–21.
18. S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N. P. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, M. Dumontier, R. D. Boyce, Toward a complete dataset of drug-drug interaction information from publicly available sources, Biomedical Informatics 55 (2015) 206–217.
19. S. Sen, A. B. Swoap, Q. Li, B. Boatman, I. Dippenaar, R. Gold, M. Ngo, S. Pujol, B. Jackson, B. Hecht, Cartograph: Unlocking spatial visualization through semantic enhancement, in: Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI '17, ACM, New York, NY, USA, 2017, pp. 179–190.
20. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, CoRR abs/1301.3781.
21. A. Drozd, A. Gladkova, S. Matsuoka, Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, 2016, pp. 3519–3530.

22. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013, pp. 3111–3119.
23. Y. Zhang, A. Jatowt, K. Tanaka, Towards understanding word embeddings: Automatically explaining similarity of terms, 2016 IEEE International Conference on Big Data (Big Data) (2016) 823–832.