VOL. 10, NO. 3, FEBRUARY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

PATTERN-BASED SYSTEM TO EXTRACT AND DISTINGUISH DRUG-ADVERSE EFFECT RELATION FROM OTHER DRUG-MEDICAL CONDITION RELATIONS IN THE SAME SENTENCE

Safaa Eltyeb[1, 2], Naomie Salim[1] and Mubarak Himmat[1]
[1]Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia
[2]College of Computer Science and Information Technology, Sudan University of Science and Technology, Khartoum, Sudan
E-mail: safaa-82@hotmail.com

ABSTRACT

Extraction of drug-adverse effect causal relationship supports pharmacovigilance research and reduces the manual efforts for some tasks such as drug safety monitoring and building databases for adverse drugs effects from free text. In this study, we proposed a pattern-based method to extract drug-adverse effects causal relation from medical case reports and to distinguish this relation from other drug-medical condition relations exist in the same sentences. For training and evaluation purposes; we used 481 sentences from ADE corpus. Our method combined a concept recognition system with a module for drug-adverse effect relation extraction and discrimination task based on automatic generated numerous patterns and the position of matched pattern in a sentence. Our method achieved recall of 36.1, precision of 30.6 and an F-Score of 33.1 .The result of this study provides rapid extraction of machine-understandable drug-adverse effects pairs which can help in many computational drug researches.

Key words: drug, adverse effect, information extraction, relation extraction, pattern-based method.

INTRODUCTION

The rapid increase in the flow rate of newly published medical knowledge has resulted in a pressing need for techniques that can simplify the use of this knowledge. In medical text, after the named entity recognition (NER) information extraction (IE) task for entities (i.e. drugs, diseases, adverse effects, etc.); a later step is the extraction of relations between those recognized entities.

An adverse drug reaction effect (ADE) (sometimes called adverse drug event or adverse drug reaction) is defined as any undesired reaction which results from the use of a medicinal product for the purpose of prophylaxis, diagnosis or therapy (Edwards and Aronson, 2000). ADEs lead to further health complications or sometimes even death.

Automatic extraction of ADEs from free text assists drugs safety authorities in rapid information screening and extraction, instead of manual inspection or conventional searching. Consequently, this accelerates the medical decision support and risk factor estimation (Gurulingappa et al., 2011).

In this paper, we describe our method for relation extraction between drugs and adverse effects from medical case reports sentences and discriminate this relation from other drug-medical condition relations exist in the same sentence. The system has been trained and evaluated in a manually annotated corpus of sentences from Medline medical case reports.

The main issue addressed in this study is whether a sentence contains more than one drug-condition relation (i.e., drug-disease TREAT relation) in addition to the drug-adverse effect CAUSE relation as in many sentences in clinical trials and case reports; How can we distinguish the CAUSE relation from other types of relations between drugs and medical conditions entities to extract the drug-adverse effects pairs from these sentences? In this work, an extraction of drug-related adverse effect pairs is conducted in this type of sentence which contain other drug-medical condition relation in addition to drug-adverse effect CAUSE relation to evaluate the system ability in discriminating the drug-adverse effect relation from other dug-medical condition relations.

The rest of this paper is arranged as follows. Section 2 presents previous methods used in ADR extraction from text. Details of the relation extraction and evaluation methods are presented in Section 3. Section 4 presents the results obtained, Section 5 presents discussion as well as Section 6 presents the conclusion and future work.

RELATED WORK

In recent years, many systems have been developed for the automatic extraction of relations between drugs and adverse effects entities. Relation extraction approaches range from applying the simple co-occurrences search to classification-based approaches.

The co-occurrence search approach which concluded on collecting instances from entities occur in the same text or part of the text. The basic assumption here, if the entities mentioned together many times; this is an indicator that they may be related in some way. However, co-occurrence search approaches alone cannot determine the type and the direction of relations and commonly exhibit high recall and low precision (Simpson and Demner-Fushman, 2012). A co-occurrences-based approaches applied by (Wang et al., 2009) and (Leaman et al., 2010) for mining relationships between drugs and

adverse effects in discharge summaries by the former and in the user comments on health-related websites by the later.

The second type of approaches for relation extraction is the rule-based approaches where the linguistic patterns exhibited by the relations between entities are utilized to generate rules to extract them. The rules can be manually specified by domain experts or derived from annotated corpora using machine learning algorithms. However rule-based approaches exhibit high precision and low recall unlike the systems based on the term co-occurrences.(Simpson and Demner-Fushman, 2012). A heuristic rule-based patterns implemented by (Aramaki et al., 2010), and used to identify the relations between drugs and adverse effects from clinical records. (Liu et al., 2011) applied statistics and heuristic methods to build up a hierarchical ontology of side effects from patient-submitted drug reviews on health-related websites. (Kang et al., 2014) developed a knowledge-based relation extraction system from Medline case reports. The knowledge base is a graph representation of concepts and relations between them, populated from the Unified Medical Language System (UMLS) which combined with a concept identification module to identify drugs and adverse effects.

(Sampathkumar et al., 2012) applied a Hidden Markov Model-based text mining system that can be used to extract the adverse drugs effects of from online medical forums.(Gysbers et al., 2007) adapted the Cancer Text Information Extraction System (CaTIES) for identifying terms suggestive of adverse drug events in documents. Another machine learning-based system by (Gurulingappa et al., 2012) based on a Support Vector Machine tool to identify and extract the drug-related adverse effects in Medline case reports.

Our work is characterized by it does not require human effort for building patterns manually to identify the relation, like all previous pattern-based systems and doesn't need large training corpus as machine learning-based systems. Furthermore, our method can discriminate between the type of relations (i.e. CAUSE, TREAT relations) between the drug and medical condition entities that mentioned in the same sentence; whereas this discrimination matter was not explored by most of the previous studies.

MATERIALS AND METHODS

Corpus

The data set used for training and testing is the ADE corpus (Gurulingappa et al., 2012). The corpus contains annotations of 5,063 drugs, 5,776 conditions (e.g. diseases, signs, symptoms) and 6,821 relations between drugs and conditions comprising drug-adverse effect relations in 4,272 sentences. Drugs and conditions that do not comprise a potential adverse event relation are not annotated.

Due to the sensitivity of our method in pattern generation, all names of drugs and conditions should be removed before generating patterns. Subsequently, all sentences containing drugs and conditions which got un suitable mapping to UMLS semantic concepts (i.e myotonia, hair loss, neutropenia, etc..), or are not covered by UMLS (i.e pruritic bullous eruption, decrease in the D-dimers, etc.) are discarded. After the corpus is cleaned, there are 3,180 drug-adverse effect relations in 2,362 sentences. For training and testing purposes we used all sentences containing more than one type of relation between drugs and medical conditions entities which represented in drug-adverse effect CAUSE relation in addition to another type of drug-medical condition relation(s) (i.e., drug-disease TREAT relation).

Method

Case-Based Reasoning (CBR) is a methodology that based on finding a previous case similar to the new one for solving problems (Aamodt and Plaza, 1994). Here, in the relation extraction and discrimination situation, the cases are patterns or expressions for drug-adverse effects relations are learnt in the training phase, and then saved in a case base. During the testing phase, the system searches the case base for cases most similar to the problem case.

A set of automatic generated patterns are exploited to identify the existence or non-existence of a drug- adverse effect relation between drug(s) and medical condition(s) pair(s) in a sentence. The automatic patterns generation distinguished over the manual creation of patterns by it doesn't require manual efforts to build the patterns and can guarantee a complete differentiation among cases. So, similar to our work in (Eltyeb and Salim, 2015); the Minimal Differentiator Expressions (MDE) algorithm proposed in (Moreo et al., 2012) was adapted in this work. This algorithm was used to generate a set of linguistic patterns (expressions) used to retrieve the case (i.e., case of sentences which most appropriate to the input sentence). Examples of the generated patterns are: {*occurring *after*},{*induced*after*},{*presented*with*after*},{*developed*during*},{*developed *caused*}, etc.. where the character '*' denotes to zero or more words. These generated patterns are used to distinguish the drug-adverse effect CAUSE relation from other drug-medical condition relations (i.e TREAT) mentioned in the same sentence, and extract the arguments of CAUSE relation (i.e., drugs and adverse effects pairs) in a further step.

The Differentiator Expression (DE) is an expression consists from token(s) of a sentence after excluding the names of drugs and medical conditions which differentiates a sentence from other sentences in the case base and not necessarily on the sentence in its own case. Here we have two cases; the first one for sentences most appropriate for drug-adverse effect patterns and the other case for sentences most appropriate for drug-disease/symptom patterns.

MDE is a DE which does not contain any other DEs (Moreo et al., 2012). For example, if we have the two DEs {*presented *with* after*} and {*presented *with*}, the DE {*presented *with* after*} is not minimal because

it contains the MDE {*presented *with*} which has fewer terms and it also allows differentiation. Any sentence in the case base is represented by its MDEs (patterns or expressions).

MetaMap[1] (Aronson, 2001), a Java API from the National Library of Medicine, is used to map the biomedical text to concepts in the UMLS metathesaurus. To recognize drugs and medical conditions in sentences, the 'Chemicals' , 'Drugs' and 'Disorders' semantic type groups are used.

For the training phase, the modules on (Eltyeb and Salim, 2015) are used to fill the case base with specific cases which they are represented on:

- Module for annotating sentences with MetaMap API;
- Module for preprocessing the annotated sentences by removing the annotated names of entities (i.e., drugs and medical conditions) and remove token(s) before first mentioned entity and token(s) after last mentioned entity to short sentences.
- Module to implement the MDEs algorithm(Moreo et al., 2012) and get MDEs for each sentence.

Then a module to calculate the MDEs' weights for each sentence according to Equation (1) and its subsequent Equation (2), Equation (3) and Equation (4) (Moreo et al., 2012) is implemented. The MDE with highest weights from training set is saved and the other MDEs discarded. Consequently, less important MDEs will not lead to a classification process.

For the testing phase; the following modules are implemented:

- Module to assign each sentence to the most relevant case relying on MDEs with the highest weights.
- Module to extract drug-related adverse effects pairs to evaluate the relation extraction and discrimination task. The extraction is based on the assignment of sentences resulted from the previous module and the position of the matched MDE in the sentence. Here our assumption is; the type of medical condition (i.e., side effect or disease/symptom) is identified by the type of the most nearest matched MDE (i.e., MDE related to the first case or MDE related to the second case).

As mentioned above the weight ω of the expression e is calculated using $CRelevance(e, C)$ (Eq.(2)) and $DRelevance(e, B)$ (Equation (4)) (Moreo et al., 2012) as displayed below.

$$\omega_{e,C} \text{ wrt } = \beta . CRelevance(e, C) + (1 - \beta) . DRelevance(e, B) \quad (1)$$

$CRelevance(e, C)$(Moreo et al., 2012)is calculated as the proportion of sentences S that satisfy expression e, according to the following equation:

$$CRelevance(e, C) = \frac{|\{S \in C \mid e \text{ occurs in } S \text{ wrt } B\}|}{|\{S \in C\}|} \quad (2)$$

Where B is the case base of cases C.

The $DRelevance(e, B)$ depends on the number of words in the expression and the importance λw of each word in the expression. The $idf(w)$ of the $tf.idf$ algorithm has been used, divided by $log|B|$ to normalize it (Moreo et al., 2012):

$$\lambda w = \frac{idf(w)}{log|B|} = \frac{log\left(\frac{|B|}{|\{C \in B | f_C(w) > 0\}|}\right)}{log|B|} \quad (3)$$

$$DRelevance(e, B) = \frac{\sum w \in e \lambda^2 w}{\alpha} \quad (4)$$

RESULTS

Our system evaluated in all sentences containing a drug-adverse effect CAUSE relation plus other drug-medical condition relation (i.e TREAT relation) to assess the ability of system in discrimination between types of relations mentioned in one sentence. Since the gold standard dataset annotates the arguments of drug-adverse effect CAUSE relation only; we also extracted and evaluated those arguments only. Performance was evaluated in terms of precision (P), recall(R) and F-measure (F) as in the following table.

Table-1. Performance evaluation (in %) of the drug-adverse effect relation extraction task evaluated by 10-fold cross-validation.

Adjustments of sentence match	P	R	F
Sentence matches at least one MDE after selecting MDEs with the highest weight.	30.6	36.1	33.1

DISCUSSIONS

We have investigated the use of automatic generated patterns to extract and distinguish drug-adverse effect relation from other drug-medical condition relations exists in the same sentence.

This considered the first study implements the extraction on specific type of sentences (i.e., sentences contain more than one different relation between drugs and medical conditions entities) to investigate the discrimination of relation types between drugs and medical conditions entities. The main advantage of the proposed method is the automation of the patterns generation process which requires less effort to build patterns. At the same time, the main limitation of method is that the results are highly dependent on the richness of the training data and to what extent the test data include sentences that match the generated patterns. We attribute the weakness of results to the few numbers of used

[1] http://metamap.nlm.nih.gov/

sentences for training and testing because we just used specific type of sentences as mentioned above.

An error analysis was carried out on a sample of 50 randomly selected errors that were made by the relation extraction and discrimination module. The largest source of errors (29 of FPs and 21 of FNs) was the FPs (58%) that occurred in our system when there were medical conditions annotated as side effects instead of diseases, and consequently introduces the FP relations. Most of these false annotations were generated by the extraction module which discriminates between medical conditions according to the position of the matched MDE. In about 3% of FP relations, there were entities annotated by the system as drugs and not annotated in the gold standard data set; consequently FP relations were established. FN relations were generated because some medical conditions were labeled wrongly, and consequently the right relations were not rendered by the system.

CONCLUSION AND FUTURE WORK

The current study has the potential to reduce the time required for manual drug safety monitoring and to assist drug safety professionals for collecting information from free text. For future improvement of the study outcome, the training data should be increased to incorporate more patterns. Also we expect using a parsing technique to select the word(s) syntactically connect(s) medical entities instead of module for shorting the sentences will increase the quality of generated patterns and consequently the quality of relation extraction task.

ACKNOWLEDGEMENTS

This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Research University Grant Category [Vot: .J130000.2528.07H89]. We also would like to thank Sudan University of Science and Technology (SUST) for sponsoring the first author.

REFERENCES

Aamodt A. And Plaza E. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. Ai Communications, 7, 39-59.

Aramaki E., Miura Y., Tonoike M., Ohkuma T., Masuichi H., Waki K. and Ohe K. 2010. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform, 160, 739-43.

Aronson A. R. Year. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium, 2001. American Medical Informatics Association, 17.

Edwards I. R. and Aronson J. K. 2000. Adverse drug reactions: definitions, diagnosis, and management. The Lancet, 356, 1255-1259.

Eltyeb S. And Salim N. 2015. Pattern-based System to Detect the Adverse Drug Effect Sentences in Medical Case Reports. Journal of Theoretical And Applied Information Technology, p. 71.

Gurulingappa Rajput A. and L T. 2012. Extraction of Adverse Drug Effects from Medical Case Reports. Journal of Biomedical Semantics, Volume 3.

Gurulingappa H., Fluck J., Hofmann-Apitius M. and Toldo L. Year. Identification of adverse drug event assertive sentences in medical case reports. In: First international workshop on knowledge discovery and health care management (KD-HCM), European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD), 2011. 16-27.

Gurulingappa H., Rajput A. M., Roberts A., Fluck J., Hofmann-Apitius M. and Toldo L. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of biomedical informatics, 45, 885-892.

Gysbers M., Reichley R., Kilbridge P. M., Noirot L., Nagarajan R., Dunagan W. C. and Bailey T. C. Year. Natural language processing to identify adverse drug events. In: AMIA... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium, 2007. 961-961.

Kang N., Singh B., Bui C., Afzal Z., Van Mulligen E. M. and Kors J. A. 2014. Knowledge-based extraction of adverse drug events from biomedical text. BMC bioinformatics, 15, 64.

Leaman R., Wojtulewicz L., Sullivan R., Skariah A., Yang J. and Gonzalez G. Year. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: Proceedings of the 2010 workshop on biomedical natural language processing, 2010. Association for Computational Linguistics, 117-125.

Liu J., Li A. and Seneff S. Year. Automatic drug side effect discovery from online patient-submitted reviews: Focus on statin drugs. In: IMMM 2011, The First International Conference on Advances in Information Mining and Management, 2011. 91-96.

Moreo A., Navarro M., Castro J. L. and Zurita J. M. 2012. A high-performance FAQ retrieval method using minimal differentiator expressions. Knowledge-Based Systems, 36, 9-20.

Sampathkumar H., Luo B. and Chen X.-W. Year. Mining Adverse Drug Side-Effects from Online Medical Forums. In: Healthcare Informatics, Imaging and Systems Biology

(HISB), 2012 IEEE Second International Conference on, 2012. IEEE, 150-150.

Simpson M. S. and Demner-Fushman D. 2012. Biomedical text mining: A survey of recent progress. Mining Text Data, 465-517.

Wang X., Hripcsak G., Markatou M. and Friedman C. 2009. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. Journal of the American Medical Informatics Association, 16, 328-337.