Abstract

Analyzing possible drug safety incidents and generating narratives in pharmacovigilance process have traditionally relied upon manual review of case reports from patients, consumers and healthcare professionals. However, due to the vast quantity and complexity of data to be analyzed and for ensuring timeliness, reduction of cost, consistency of reporting and quality of reporting; role of automated computational systems that can accurately detect adverse drug reactions attached to a suspected drug in a timely fashion have become critical. Pharmaceutical companies have started to realize the need for collaborative and integrative approaches and strategies to allow a faster identification of high-risk interactions between marketed drugs and adverse events, and to enable the automated uncovering of scientific evidence behind them. The fundamental requirement for the automatic processing of biomedical text is the identification of information carrying units such as the concepts or named entities. Additionally, there are regulatory guidance or rules with respect to identifiability of reporters, patients, drugs and interactions in the reports of suspected adverse reactions. Owing to these challenges, the problems of automated unambiguous identification of medical drugs and compounds, detection of adverse drug reactions, and generation of case narratives from the text of the reports are not considered to be adequately solved so far. In this paper, we present a novel text analysis platform that assists in bringing intelligent automation in the process by integrating a medical language processing pipeline and causal reasoning chain, with publicly available large-scale biomedical databases containing structure, bioassay, and genomic information, as well as comprehensive clinical data sets.

© 2014 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: Pharmacovigilance; Adverse Drug Reaction; Medical Text Processing

* Corresponding author. Tel: +91 80 43162608; E-mail address: anutosh.maitra@accenture.com
1. Introduction

Pharmacovigilance (PV), also known as Drug Safety Surveillance, is the pharmacologic science relating to the collection, detection, assessment, monitoring, and prevention of adverse effects with pharmaceutical products. This is an important process that allows health regulatory authorities to continue to assess benefits and risks throughout the life-cycle of a medicine and potentially detect serious adverse events and new drug safety signals that were not detected before marketing authorization. The process generally contains information received from patients, healthcare providers, medical literature, physicians, pharmaceutical company’s sales team or pharmacists. Information collected from different above mentioned sources should be processed in a defined consistent way for electronic submission to the regulatory authorities like FDA (Food and Drug Authority), WHO (World Health Organization), MHRA (Medicines and Health Regulatory Agency), EMA (European Medicines Agency) and other local authorities. Apart from regulatory requirements, pharmaceutical companies need to achieve this to serve public health, and to foster a sense of trust in patients in the medicines they use, and lastly but most importantly, to proactively monitor drug effects to prevent product withdrawal from market due to safety issues.

The main business challenges of PV, as seen today, are:

- Efficient and cost effective running of the process consistently without compromising on quality.
- Development of processes and tools to improve business decisions based on disparate and unstructured data.
- Continue to build deep skills in identifying the causation relationship, thus migrating PV from being a reactive to an anticipating process.

Maintaining a robust PV system relies on consistent and accurate acquisition, integration and analysis of adverse event data. Without a strong foundation, important safety signals may not be fully identified and evaluated. As much as 30% of all drug reactions result from concomitant use as an estimated 29.4% of elderly patients are on six or more drugs. Several published drug-safety papers have shown that adverse effects of drugs may be detected too late, when millions of patients have already been exposed to them. For a long time, researchers have been seeking a real time, continuous and prospective approach that could integrate vast, dispersed and unstructured information and knowledge bases to obtain unambiguous drug reaction relationships and automate the narrative generation process.

It is at this context, a novel text analysis platform has been built that uses advanced natural language processing (NLP) at its core to attain better accuracy in identifying nuanced attributions of disease conditions. The aim is to provide an automated system that would enable faster case processing and also nurture an integrative approach to quickly detect adverse drug reactions (ADR) with the associated scientific evidences properly identified.

2. State of the Art

At the heart, an effective ADR detection system finds out if an unexpected medical reaction has been caused due to an intake of a drug. In terms of text analysis, it means establishing a linkage between the event of a drug intake and an adverse reaction – where both are textual entities or concepts in a running narrative. An increasing number of text mining researchers focus on extracting and establishing associations between entities from textual data in structured and unstructured event reports, often through NLP. Co-occurrence statistics has been widely used to determine relations between the entities, and has proven to be effective in acquiring associations between biological and clinical entities. A number of systems with high throughput have been developed to extract adverse drug reactions from the biomedical literature and clinical documents, while a few other approaches have dealt with texts from social medial platforms or online health forums and websites. Disproportionality Analysis based data mining algorithms have been tried for years to detect ADRs, but these techniques have rarely been successfully commercialized. Many pharmaceutical companies use commercial platforms like Oracle’s ARGUS for adverse reaction event reporting activities. Some other organizations also possess home-grown platforms for the same process. However, these systems only use NLP in a limited capacity as the relationships between the drugs, the diseases (manifested through their system organ classes) and reactions are rarely investigated in an automated
fashion. This is primarily because the underlying clinical knowledge databases are very disparate in their structure and management. Despite the well-established understanding about the need of a collaborative knowledge framework to automate the PV process through semantic integration of these databases, not much concerted effort is seen in establishing a proper relationship amongst multiple databases to assist the practitioners in the PV business. Substantial manual effort is involved today in looking up different databases and reviewing and deciding on medical causation from these data sources. In other words, an effective PV evaluation today depends mostly on manual effort and on the ability and judgment of the practitioner.

Most of the text analytics based work on PV domain have been restricted to academic and research purposes and address only a few of the sub processes involved in the complete process chain. Complete end-to-end processing of adverse event (AE) reaction reports are usually not supported by these systems. Constant human intervention at data entry level as well as in the judgmental level creates lot of inconsistency and this has been a big hurdle to bringing high quality output and hence it could not gain the trust of the users or the authorities.

2.1. Motivation of the Present Work

Text Mining applied to the PV literature can be of great benefit, allowing identification and extraction of relevant information, and providing an interesting way to reduce the time spent by healthcare professionals and researchers who are trying to stay current by reviewing the literature. With the integration of an end to end text Analytics platform, that could be one of the first of its kind in the industry, one would aim to achieve:

- Reduction in the number of different human roles for case processing. Ideally not more than two roles, viz. a case processor or a narrative writer and a medical reviewer should be involved in the complete ADR detection and narrative generation process.
- A close to 50% productivity efficiency for the case processing, as the number of roles are limited and necessary knowledge support can be provided by an automated platform.
- More than 90% efficiency in medical review output in judging the seriousness of the event, if the event is already listed as related or expected, and in proposing possible causality relationship between event and the suspected drug. Efficiency in PV case processing today is heavily dependent upon the ability of the medical professionals to quickly find out the causation chain based on their training, experience and access to available data resources.

3. Collaborative Knowledge Management

NLP for information extraction from medical texts is a daunting task. The fundamental requirement is to identify the information carrying units, i.e. the named entities and the relationships existing between the entities. In a specialized domain like medical, disambiguation of named entities needs referral to authentic clinical and regulatory databases. Since no single database can provide the complete information, consistency in granularity of the codes corresponding to diseases and symptoms is essential. The entire analysis platform is thus supported by a series of databases belonging to largely three categories:

- Medical databases for anatomy, drug, device and medical condition related knowledge (e.g. SNOMED-CT, RxNORM and FMA)
- Medical databases that provide detailed identifier and medical coding for safety reporting purpose (e.g. WHO-DD and MedDRA)
- Databases that provide historical post marketing drug / active ingredient safety related information, also known as safety or package label databases. (e.g. DrugBank, EMC and DailyMed)

Clearly, collating and curating such vast amount of information from multiple database sources present us with its own set of challenges. However, these databases have to be collaborated for the purpose of surface form disambiguation to recognize medical entities and relationships, and also for drawing inferences by leveraging past clinical knowledge. All historical and new reaction data are coded with the new MedDRA (Medical Dictionary for Regulatory Activities) scheme. The strategy is to design an integrative data capture system which effectively
exploits current computing advances and technical performance to automate the aspects of initial adverse event review, supporting more efficient and effective clinical assessment of potential safety signals.

4. Proposed Bio-Medical Text Processing Approach

The scope of the proposed text analysis platform spans over a process zone starting from receipt of drug safety reports and ending with the submission of case narratives to the regulatory authorities. The information collected from the source documents usually include the drugs suspected to cause the ADR, concomitant drugs, indications, suspected events, and limited demographic information. While information embedded in the text is crucial for automating the process, the big challenge is the various shapes and structures these text documents can come with. As a preparatory step; the input documents, coming in various formats, are ingested into the system, extracted, and pre-processed into a canonical standard input XML formats. The step of extraction and canonicalization helps fusing structured and unstructured data from different sections of the safety forms such as CIOMS\textsuperscript{15}, into pre-defined elements, and handling of not-properly identified inputs as may be generated from optical character recognition or such techniques for extracting from document images. The canonical input XML also helps in processing the NLP engine without having to deal with multiple structural formats, or a free format running text inputs.

![High Level Block Diagram of Medical text Processing](image)

Fig.1. High Level Block Diagram of Medical text Processing

Fig.1 above presents the high level block diagram for the NLP process when applied to the medical texts. The first step in medical language processing is to identify and extract “features” that may help structure information. Generally, ontologies\textsuperscript{16} are excellent source of features and allow systematic normalization and aggregation when the feature set needs reduction. An ontology description was created depicting the main entities, relationships and other important concepts (we do not discuss it in details for sake of brevity). Primarily with the help of this ontology, two principal information extraction tasks are performed – that of identifying Semantic Classes and identifying Semantic Associations and Relationships among these classes. The semantic classes are key medical entities of interest. The semantic associations are key relationships. In particular, we are interested in capturing all relationships involving medical interventions, diagnostic results, and patient condition status and treatment outcomes; as these are the ingredients of reasoning and writing narratives. The information extraction task is done by cTakes\textsuperscript{17}, a parser trained in medical domain, with the help of the said PV oriented ontology and medical databases for surface form disambiguation. The existence of patient, drug, reporter and event is also established.

After the identification of the classes (entities) and relationships, the next step is to create the temporal and causal textual chains. These textual chains identify if there are hints about associations between an indication for which a drug is taken, the actual event of drug consumption with related dosage and the subsequent reaction noticed. Thereafter, we “codify” the drug and the disease tokens. The coding for diseases that could include past medical conditions, the medical indications and adverse reactions, are done against the appropriate terms from medical dictionary for regulatory authorities (MedDRA) database promoted by International Council for Harmonization (ICH). The coding of the drugs involved is done against WHO governed Drug Dictionary database. Such coding is essential for proper PV reporting.
The next step is to drawing of clinical and process related inferences from the text models. The inferences are drawn with the help of backend databases, ontologies and knowledge sources. To illustrate, to find out if an adverse event is significantly serious, we compare the extracted token against a database containing Important Medical Event (IME) list. A set of medium to high complexity rules to carry out different types of inferences has been implemented, a snapshot of which is presented in the succeeding section.

Finally, creation of case narratives is also automatic. A narrative describes a chronological and etiological summary of the case. It is important to capture the physical intervention and clinical status events through a chronology chain. Similarly all the clinical diagnosis, treatment and clinical causality logic are captured through an etiology. We employ a template based natural language generation for narration creation. Such a narration, after review and approval from a medical practitioner, gets routed out in another canonical XML format as a reviewed case for internal purposes or for electronic submission to the authorities.

5. Inference Process

A Drools based Rules Engine has been integrated to draw machine assisted inference for drug event seriousness and causality. Sometimes keywords like “Death” could define serious events, but in general, event seriousness is primarily determined by the existence of certain key concepts and relationships in the source document. One important step in writing the narratives is to determine the causation chain for the ADR reported so that the suspect drug can be identified. Association between the drug and the reaction event does not necessarily imply causation as causation not only requires correlation but also a counterfactual dependence. This is achieved by a sequence of rules and database checks, to individually identify drug-drug interactions, drug-reaction associations and remote indication associations. For the sake of brevity, every causality rules are not described here, but suffice to say that effective automatic inferences were drawn by creation of a large number of simple rules as shown in Fig.2 below. Special causality rules are used to address special cases like expectant mothers and lactating mothers.

Many Suspect Drugs, One Adverse Event
FOR Each Drug
IF
Adverse Event == known Adverse Reactions ["section48"] EMC/ Daily Med
Then, Tag: Primary Suspect
ELSE
Unknown

Fig.2. Example Rule Set

6. Proposed Platform Schematic

The system follows a pipelined approach to process the electronic drug safety reports being generated in the field through a series of document processing and natural medical language processing steps. We have adopted a stage-wise processing through a client-server architecture as depicted in the figure 3 (a) below.

The case processors, who could be a safety advisor or a medical reviewer, are expected to access the system through a web UI. The server-side hosts the core engine consisting of a NLP run-time, a workflow engine, an ontology server and a rules engine. The NLP run-time performs a series of information extraction tasks like semantic class and relationship extraction, and processing tasks like temporal and causal chain creation. The ontology server hosts and serves the PV specific ontology that captures the essential knowledge required for processing. The ontology is edited and managed through Protégé. The Drools Engine hosts the business rules concerning drug event seriousness and causality tests like what conditions must be satisfied so the suspect drug involved can be taken to cause the adverse reaction and so forth.

The entire architecture, although running in a pipeline, is also event driven. Any event – such as user edits or updates are propagated through the stack to invalidate certain computed data and force re-computation. We also attach one screenshot of the application in the figure 3 (b).
7. Conclusion

The proposed NLP based platform for PV is aimed at achieving automation and machine assisted decision support in narrative creation and has the ability to pick and understand the data sources, detect associations and causation and is regulatory compliant. It also integrates strong life sciences practices and decades of rich experience of universal pharmaceutical research in form of rules. The framework is already deployed in the PV process of a couple of large pharma companies and is expected to reduce the cost as well as increase the quality and efficiency to a large extent. Actual productivity figures are now being monitored over prolonged usage of the platform and the efficacy of the platform can be reported thereafter.

References

7. SNOMED Clinical Terms description In http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
12. Drug related bioinformatics and cheminformatics database In http://www.drugbank.ca/
15. In http://www.cioms.ch/
17. In cttakes.apache.org/