Drug name recognition and classification in biomedical texts

A case study outlining approaches underpinning automated systems

Isabel Segura-Bedmar[1], Paloma Martínez[1] and María Segura-Bedmar[2]

[1] Computer Science Department, Carlos III University of Madrid, Madrid, Spain
[2] Drug Information Center, Pharmacy Department, Mostoles University Hospital, Madrid, Spain

This article presents a system for drug name recognition and classification in biomedical texts. The system combines information obtained by the Unified Medical Language System (UMLS) MetaMap Transfer (MMTx) program and nomenclature rules recommended by the World Health Organization (WHO) International Nonproprietary Names (INNs) Program to identify and classify pharmaceutical substances. Moreover, the system is able to detect possible candidates for drug names that have not been detected by MMTx program by applying these rules, achieving, in this way, a broader coverage. This work is the first step in a method for automatic detection of drug interactions from biomedical texts, a specific type of adverse drug event of special interest in patient safety.

Introduction

The pharmaceutical industry is increasingly becoming a knowledge-based discipline. Scientists need to access relevant information and knowledge in the process of developing drugs. The deluge of published research has overwhelmed most health care professionals because it is not possible to be kept up-to-date of everything published about, for instance, drug interactions.

Although some of the information is available in a structured form, a great deal of the most current and valuable information is unstructured and written in natural language. Information extraction from both structured and unstructured data sources can be of great benefit in the pharmaceutical industry, allowing identification and extraction of relevant information.

In the exponentially growing number of biomedical text sources, literature-based research is crucial in turning an idea into a marketable therapeutic product 12–15 years later [1]. Industry estimates suggest that 90% of drug targets are derived from the literature [2]. One of such sources is PubMed with more than 16 million MEDLINE journal article references and abstracts going back to the mid-1960s, together with 1.5 million references back to the early 1950s. Some 900 million searches of MEDLINE are done each year by health professionals, scientists, librarians and the public.

The detection of drug interactions, as a type of adverse drug event in clinical care, is an important research area in patient safety. Patient safety has become a priority for health systems and, in fact, the Institute of Medicine (IOM), in its report *Err is Human* [3], estimated that between 44 and 98 thousand people die in US hospitals each year as the result of problems in patient safety.

In recent years, several of the major health organizations, such as the World Health Organization (WHO), the Pan American Health Organization (PAHO) and the European Environment and Health Committee (EEHC) have developed strategies to propose plans, actions and legislative measures to control avoidable adverse effects in clinical practice. However, while some progress has been made in the area of patient safety, there is still much room for improvement. Thus, though some areas have effective safety systems, the area of drugs does not appear to have reached the level of development that initially was expected [4,5].

Among the advice given for patient safety, major health organizations recommend promoting communication of incidents in electronic medical records. For this reason, we believe that patient

Corresponding author: Segura-Bedmar, I. (isegura@inf.uc3m.es)

1359-6446/06/$ - see front matter © 2008 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2008.06.001

safety will be greatly enhanced by developing a valuable tool for efficient information access about drug interactions for the healthcare professionals, as well as facilitating the semi-automatic updating of drug knowledge databases.

Focusing on this topic of automatic discovery of drug interactions from biomedical texts, the recognition of pharmacological substances is not only an essential prerequisite step, but also required in other kinds of applications, such as information retrieval, information extraction, information management and new knowledge discovery in the pharmacological domain.

Drug name recognition aims to find drugs in biomedical texts and classify them into predefined categories, such as analgesics, antihistamines, antivirals and so on. This process is a challenging task, given the difficulties implied in biomedical text processing. First, with the rapidly changing vocabulary, new drugs are introduced on an almost daily basis, while old ones are made obsolete. Second, it is difficult to maintain and update terminological resources constantly. Although frequently updated, such resources cannot keep up with the accelerated pace of the changing terminology. Thus, systems capable of automatically detecting candidate terms for augmenting these ontologies would help in speeding up the time-consuming task of maintenance. Third, naming conventions are available for a variety of domains in the Biomedicine field; however, these conventions are not strictly followed. Despite this fact, integrating this type of information can help in gaining basic insights into the underlying meanings of the terms in question and, therefore, help in the classification of the terms.

We present a system for drug name recognition and classification from texts, necessary as a first step in automatic detection of drug interaction in biomedical texts. A collection of abstracts from MedLine are processed by Unified Medical Language System (UMLS) MetaMap Transfer (MMTx) [6], a highly configurable program that maps biomedical texts to concepts in the UMLS[3] Metathesaurus. MMTx allows identification of the pharmacological substances as well as other biomedical concepts in the text. Subsequently, the pharmacological substances are classified by a second module, on the basis of nomenclature rules recommended by the WHO International Nonproprietary Names (INNs)[4] Program. These nomenclature rules classify pharmacological substances according to pharmacological or chemical groups. In addition, they help in the identification of drugs that have not been detected by the MMTx program.

Drug families may represent a valuable clue for the detection of drug interaction in biomedical texts. In the vast majority of cases, drugs that belong to the same family usually share the basic chemical structure and mechanism of action [7,8], though there are exceptions. Therefore, if the interaction of a particular drug is known, there is a reasonable probability that another drug with similar chemical structure and metabolic pathway will exhibit similar interaction [9].

Identifying gene and protein names, chemical compounds and drugs and diseases is crucial for facilitating the retrieval of relevant documents and the identification of relationships between those entities (e.g. protein interactions, drug interactions and so on).

Biomedical named entity recognition (BNER) is defined as: the task of recognizing and categorizing entity names in biomedical domains. Different approaches for handling the problem of BNER have been developed: rules based approaches; dictionary-based approaches; machine learning techniques; statistical methods and hybrid approaches combining different techniques. Numerous studies have tackled this topic in the fields of genes and proteins [10–15]. Nevertheless, the field of drugs has not been widely addressed [16,17].

Kolarik *et al*. [17] have described an approach for the identification of new terms used in unstructured text that provide information about drug properties. They propose using lexico-syntactic patterns for identifying and extracting drug property information. The evaluation of terms extracted from Medline showed that 29–53% of them are new valid drug property terms not included in the Drugbank database.[5]

Description of resources and methodological approach

The Unified Medical Language System (UMLS)

Due to the increasing amounts of biomedical information available in electronic form, the National Library of Medicine (NLM) has developed the UMLS to retrieve and integrate information from multiple sources, such as bibliographic databases, patient record systems, factual databanks and knowledge bases [18]. In addition, UMLS can help to develop natural language technology for biomedical texts.

The UMLS has three major knowledge sources: the Metathesaurus; the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus is a very large, multipurpose and multilingual database that contains information about biomedical and health-related concepts, their various names and the relationships among them. It is built from more than 100 biomedical vocabularies and classifications, including Alcohol and Other Drugs Thesaurus, Master Drug Data Base, Micromedex DRUGDEX, Metathesaurus Forms of FDA National Drug Code Directory, National Drug Data File Plus Source Vocabulary, among others. The organization of Metathesaurus is based on the concept or meaning that links alternatives of the same concept together and identifies useful relationships between different concepts.

The current release of the Semantic Network contains 135 semantic types and 54 relationships. All concepts in the Metathesaurus are assigned to at least one semantic type from the Semantic Network, providing a consistent categorization of all concepts.

Figure 1 includes a portion of the UMLS Semantic Network. In this work, we mainly focus on 'Pharmacological Substance' (PHSU) and 'Antibiotics' (ANTB) semantic types.[6]

The UMLS Semantic Network allows the semantic categorization of a wide range of terminology in multiple domains as a result of its broad coverage. However, we find that the UMLS semantic types are too broad to classify the concepts of specific domains such as pharmacological ones. For example, all pharmacological substances are categorized as PHSU, or, as ANTB, a subtype of the former. A more fine-grained classification is necessary to develop

[3] http://www.nlm.nih.gov/research/umls/.

[4] http://www.who.int/medicines/services/inn/en/.

[5] http://www.drugbank.ca/.

[6] The pharmaceutical preparations, which are produced by manufacturers, are classified by 'Clinical Drug' semantic type, and actually, are outside the scope of our study.

methods for the automatic extraction of common relations in pharmacological texts such as drug interactions. The pharmacological and/or chemical group to which a drug belongs can be an essential clue to detect information automatically regarding its interactions or adverse effects.

Finally, SPECIALIST is a lexicon that contains syntactic, morphological and orthographic information for biomedical and common words in the English language.

UML MetaMap Transfer (MMTx)

The MMTx is a program developed at the NLM to map text to concepts in UMLS Metathesaurus or, equivalently, to discover concepts in text. MMTx uses a knowledge-intensive approach on the basis of symbolic, natural language processing and computational linguistic techniques [19].

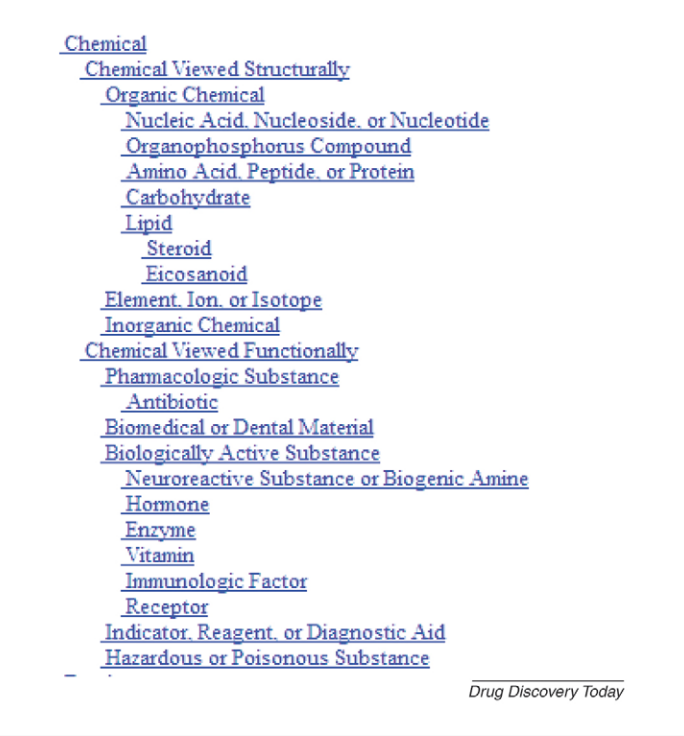
Nomenclature rules recommended by WHOINN

Once abstracts have been processed by MMTx and terms occurring in the text annotated and related to concepts of the UMLS Metathesaurus, a second rule-based module classifies the pharmacological substances occurring in texts in pharmacological or chemical families.

This module implements the naming convention rules defined by the WHOINNs[7] Program to facilitate the identification and classification of pharmaceutical substances or active ingredients. Each INN is a unique name that is globally recognized and is public property. A nonproprietary name is also known as a generic name.

The rules are based on the common stems selected and defined by WHOINN. These common stems, currently in use, represent classes of substances that are pharmacologically or chemically related [20]. By using common stems the medical practitioner, the pharmacist, or anyone dealing with pharmaceutical products can recognize that the substance belongs to a group of substances having similar pharmacological activity or chemical structure. Table 1 shows some of the stems used in the classification of drug names. The full list and the stem classification can be found in the document [21]. This classification, proposed by WHOINN, consists of two different types of categorizations: pharmacological or chemical. Thus, it is more inconsistent than other classifications such as the Anatomical Therapeutic Chemical (ATC)[8] classification system in which the drugs are divided into different groups according to the organ or system on which they act and their chemical, pharmacological and therapeutic properties. Despite the inconsistencies of the WHOINN classification, we have decided to use it because it provides the stems and pharmacological, as well as chemical, information that could also be very useful in predicting drug interactions occurring in the texts.

The term 'drug family' is frequently used and can be interpreted as 'pharmacological group', though it is also often used to designate 'chemical group'. We have decided to use the term 'drug family' instead of 'group', since this term is broader and more general and because it allows us to include both the pharmacological and chemical groups.



- Chemical
  - Chemical Viewed Structurally
    - Organic Chemical
      - Nucleic Acid, Nucleoside, or Nucleotide
      - Organophosphorus Compound
      - Amino Acid, Peptide, or Protein
      - Carbohydrate
      - Lipid
        - Steroid
        - Eicosanoid
    - Element, Ion, or Isotope
    - Inorganic Chemical
  - Chemical Viewed Functionally
    - Pharmacologic Substance
      - Antibiotic
    - Biomedical or Dental Material
    - Biologically Active Substance
      - Neuroreactive Substance or Biogenic Amine
      - Hormone
      - Enzyme
      - Vitamin
      - Immunologic Factor
      - Receptor
    - Indicator, Reagent, or Diagnostic Aid
    - Hazardous or Poisonous Substance

Drug Discovery Today

FIGURE 1

Some UMLS Semantic Types

The system architecture

In the current implementation PubMed [2] was used as the main data source because of its wide coverage of biomedical sciences and its public availability.

A corpus of medical abstracts was compiled with the help of a web crawler, 849 abstracts were downloaded from PubMed, using the keyword 'drug interaction'. In this corpus, DrugDDI, attempted to identify and classify the drug names and other biomedical concepts through different processing stages (Fig. 2) that output the information in XML format.

First, MMTx analyzes the text syntactically to split it into components including sentences, paragraphs, phrases, lexical elements and tokens. Secondly, the UMLS SPECIALIST Lexicon generates variants from each phrase to look up the concepts in the Metathesaurus, which contain one or more of these variants. A set of candidate concepts are retrieved from the UMLS Metathesaurus and then are evaluated against the phrases using a linguistically rigorous metric. Those candidates that best fit the text are selected and organized into a final mapping.

Although MMTx data is updated each year, the latest available release of MetaMap, MMTx 2.4.C, is on the basis of the 2006AA UMLS knowledge sources. Thus, MMTx cannot detect those new concepts that have been included in UMLS over the past year (2007). We believe that the stems recommended by the WHOINN not only allow the classification of the pharmacologic substances, but also help to find possible new pharmacological substance candidates that have not been detected by MMTx.

Once the UMLS concepts have been detected by MMTx, the rule-based module classifies the generic drugs according to the stems recommended by the WHOINN Program. These stems

[7] http://www.who.int/medicines/services/inn/en/.

[8] http://www.whocc.no/atcddd/.

TABLE 1

Some stems recommended by WHOINN

Stems	Family
-flurane	General anaesthetics, volatile
-bersat, -toin	Anticonvulsants
-adol-, -azocine, -eridine, -ethidine, -fentanil, nal-	Narcotic analgesics
-ac, -adol-, -arit, -bufen, -butazone, -coxib, -icam, -fenamate, -nixin, -profen, -metacin, -adom, -fenine	Analgesics–antipyretics
-fylline, -racetam, -vin-	Analeptics
-azenil, -azepam, -bamate, -carnil, -peridone, -perone, -pidem, -plon, -pride, -quinil, -spirone, -zafone	Anxiolytic sedatives
-perone	Antipsychotics (neuroleptics)
-oxetine	Antidepressants
-giline*, *-moxin	MAO inhbitors
-pin(e), -pramine, -triptyline	Tricyclic antidepressants
-anserin, -setron	Serotonin receptor antagonists
-caine	Local anaesthetics
-curium, -ium	Neuromuscular blocking agents
-azoline, -drine, -frine, -terol	Adrenergic agents
-serpine	Adrenergic neurone blocking agents
-verine	Spasmolytics, general
-afil, -dil, -entan	Vasodilators
-dipine, -fradil, -pamil, -tiazem	Coronary vasodilators, also calcium channel blockers
-nicate	Peripheral vasodilators
-astine	Antihistaminics
-tadine, -tidine	Histamine H1, H2 receptor antagonists
-bradine, -denoson, -vaptan	Cardiovascular agents
-dan, -rinone, -afenone	Cardiac glycosides and drugs with similar action
-afenone, -aj-, -cain-, -ilide, -isomide, -kalant	Agents influencing heart muscle excitability and conductivity
-azosin, -dralazine, guan-, -kalim, -kiren, -(o)nidine, pril(at), -sartan	Antihypertensives
-fibrate, -nicate, -vastatin	Antihyperlipidaemic drugs
-cog, -cogin, -fiban, -gatran, -parin	Agents influencing blood coagulation
-arol, -grel-, -irudin, -pafant, -troban	Anticoagulants

together with their corresponding pharmacological or chemical groups are compiled in a list. This information can be obtained from part III of document [21] that presents the stem classification system used by the INN Program to categorize the main activity of pharmaceutical substances. This list is scanned in order to build the suitable regular expression for each stem. For example, for the stem –*flurane,* the regular expression should be [A-Za-z0-9]*flurane, so any alphanumeric string which ends with the suffix -*flurane* is recognized by this regular expression. Similarly, for the stem -*adol-*, the regular expression should be [A-Za-z0-9]*adol[A-Za-z0-9]*. Once the regular expressions have been built, the module tries to match the text of each phrase with the regular expressions in order to detect the possible stems, which can classify the phrase. In the case in which several regular expressions can be matched with the text of the phrase, the module selects the longest stem. Table 2 shows some examples. However, the module could be configured to keep all the candidate stems, and the selection of the most suitable stem could be made by the final user of the DrugNer tool. When a correct stem is found, appropriate information about the pharmacologic or chemical family and the definition associated with the stem is added to the phrase.

The rules are not only applied to the phrases that have been classified as pharmacological substances or as antibiotics by MMTx program, but also to those for which MMTx did not found any candidate concept in UMLS. Thus, these phrases are possible new candidates for drug names that are not included in UMLS Metathesaurus.

Study outcome

Table 3 presents the main characteristics of the DrugDDI corpus created for the evaluation of the proposal. MMTx is able to find 8093 phrases (7.5% of all phrases) that are categorized as pharmacological substances (7691) or as antibiotics (402) in UMLS. Of those phrases, 49.8% (53,037) belong to other semantic types such as organic chemical, lipid, carbohydrate and so on. This subset is out of the scope of the present study. For the rest of the phrases, 45,449, MMTx did not find any concept in UMLS that covered them.

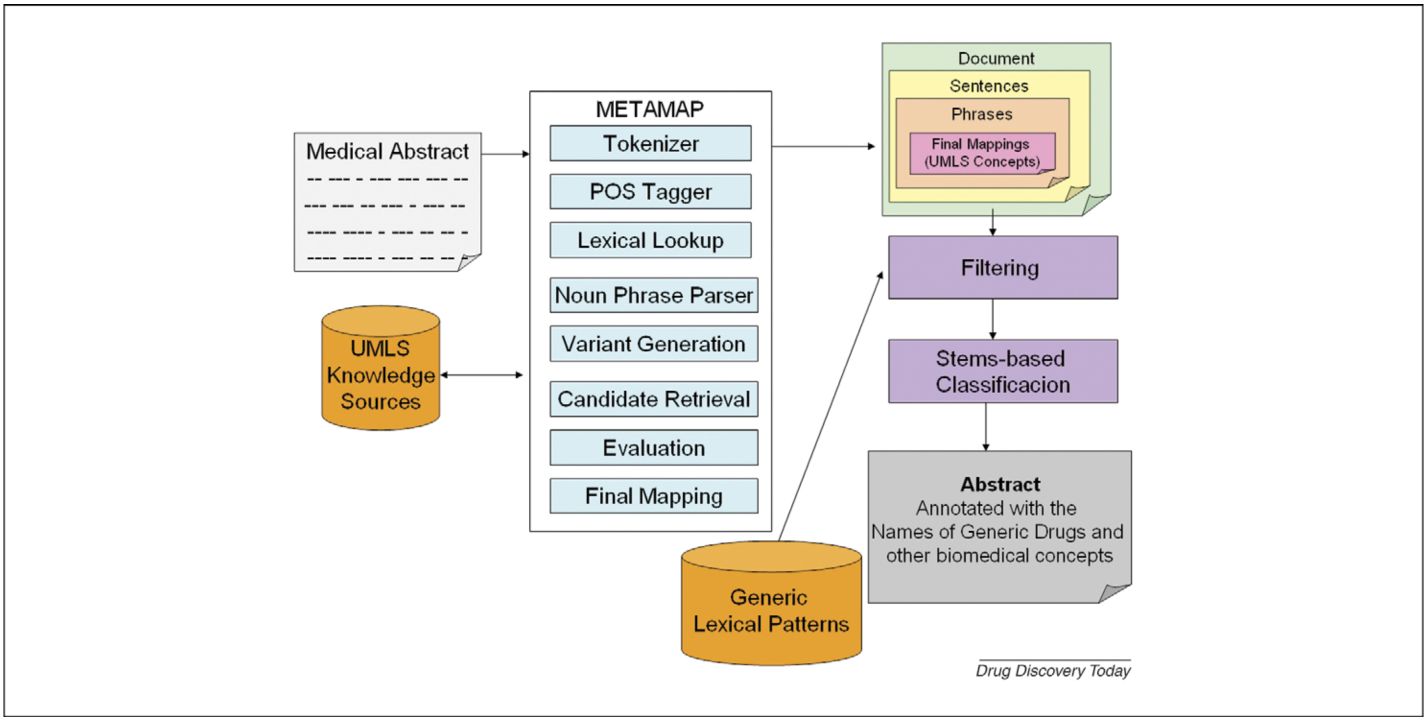


FIGURE 2

System Architecture

In order to identify new candidates of generic drugs not detected by MMTx, the stem-based module is also applied to the phrases that have not been detected by MMTx (45,449), identifying 255 initial candidates. A pharmacologic expert evaluated this set, ruling out 74 phrases and declaring the rest, 181, as generic drugs or pharmacological substances. Some of these pharmacological substances are presented in Table 4.

To calculate the total coverage of our system it is necessary to take into account those drugs that have not been detected either by MMTx or by stems. The manual evaluation of 45,194 phrases is time-consuming as well as tedious. For this reason, an automatic process (Fig. 2) to filter terms such as numeric expressions, verbs, adverbs and common nouns of biomedical domain, was applied on the subset of phrases and was able to decrease the number to 5964. Finally, a manual evaluation shows that only 20 of them are drugs. Table 5 shows some of them.

Precision and Recall are standard measures for evaluating the performance of Information Retrieval Systems [22]. Recall can be described as the ratio between a number of correctly recognized drugs and all the drugs occurring in the corpus. Precision is the ratio between the number of correctly recognized drugs and all the drugs recognized by the system (see Table 6). Table 7 shows the overall performance obtained using only MMTx and combining MMTx and the stem-based classification.

An important contribution of this work is the classification achieved by the stems recommended by the WHOINN Program that MMTx is not able to provide. Our hypothesis is on the basis of the idea that the stem-based classification could allow detection of the pharmacological or chemical family of the drugs classified as pharmacological or as antibiotics by MetaMaps, achieving, in this way, a more informative and suitable categorization of them.

TABLE 2

Examples of matching phrases and stems

Drug	Suitable stems	The most suitable stem
Azelnidipine	-dipine, -pine, -ine, -ni-	-dipine
Lopinavir	-navir, -vir-	-navir
Amiodarone	-arone, -one, -io-	-arone
Minocycline	-cycline, -ine	-cycline
Sulfinpyrazone	-azone, -zone, -one	-azone
Aripiprazole	-piprazole, -prazole	-piprazole
Furafylline	-fylline, -ine	-fylline
Gemcitabine	-citabine, -abine, -ine	-citabine
Mometasone	-metasone, -one	-metasone
Simvastatin	-vastatin, -stat-	-vastatin

TABLE 3

Characteristics of DrugDDI corpus

Characteritics	
Number of abstracts	849
Number of sentences	10,146
Number of phrases	106,579
Number of phrases detected by MMTx (i.e. phrases which are concepts in UMLS)	61,130
Phrases which are classified as PHSU or as ANTB in UMLS	8,093
Phrases which are classified with other types in UMLS	53,037
Number of phrases not detected by MMTx	45,449

TABLE 4

Some pharmacological substances detected only by the stem-based classification

Name	Stem	Family	Num
Ciclofenac	-ac	Antiinflammatory	5
Efepristin	-pristin	Antibacterial	7
Armodafinil	-nil	Anxiolytic sedatives	10
Dabigatran	-gatran	Antithrombotic agents	3
Aplaviroc	-vir-	Antivirals	1
Maraviroc	-vir-	Antivirals	5
Vicriviroc	-vir-	Antivirals	3
Darunavir	-navir	Antivirals	39
Dasatinib	-tinib	Antineoplastic agents	7
Sunitinib	-tinib	Antineoplastic agents	28
Nilotinib	-tinib	Antineoplastic agents	2
Vorinostat	-inostat	Histone deacetylase inhibitors	7
Sitagliptin	-gli-	Oral antidiabetics	7
Tanespimycin	-mycin	Antibiotics	5

TABLE 5

Some drugs detected neither by MMTx nor by stem-based classification

Name	Family
Posaconazale	Triazole drug
Rapamcyin	Immune suppression drug
Gadobenate dimeglumine	Contrast agent for magnetic resonance
Riluzole	Nervous system drugs
2-Methoxyoestradiol	Angiogenesis inhibitors

Initially, the stems were able to classify 48.5% (3926) of the phrases that were detected and categorized as drugs by MMTx (8093).

In order to assess the accuracy of the stem-based classification, the pharmacologic expert manually evaluated the phrases. The ATC classification system and other drug information resources were used to assist in this evaluation. Accuracy can be defined as the ratio between the number of correctly classified drugs and all the classified drugs by the stem-based module. The evaluation shows that 2941 have been correctly classified by the stems as opposed to 355 that have been wrongly classified. In other words, the stem-based classification obtains an accuracy rate of 74.9%. Short stems such as *-pin* (*tricyclic antidepressants*), *-ol* (*alcohols and phenols*), *-ox* (*oral antiacids*), *-ni-* (*nicotinic acid or nicotinoyl alcohol derivatives*) are responsible for the incorrect classifications. Thus, additional clues are necessary to detect these drug families.

TABLE 6

Number of pharmacological substances in the corpus

Numbers of drugs occurring in the corpus	
Detected by MMTx	8093 (97.6%)
Only detected by stems	181 (2.2%)
Detected neither by MMTx nor by stems	20 (0.2%)
Total	8294

TABLE 7

Overall performance of the system

	Recall (%)	Precision (%)
MMTx	97.5	100
MMTx + stems	99.8	99.1

DrugNer: a visual tool for the automatic extraction of named drugs

A prototype, called DrugNer, has been developed; it is a visual tool that highlights the generic drugs and pharmacological substances in texts. DrugNer allows exploration of folders and selection of text files. Once the user has selected a file, the system processes it to detect and classify the drugs and pharmacological substances. Finally, DrugNer shows the content of file and highlights the drugs occurring in the text (Fig. 3). In addition, the user can select any of the highlighted drugs, and then, DrugNer shows information concerning the selected drug (pharmacological drug, definition, stems used in its classification and so on).

Currently, we are working to extend this prototype with a module to extract information about drug interactions occurring in the texts. We believe that DrugNer can become a valuable tool for healthcare professionals and scientists, allowing them easy and rapid access to relevant information about drugs.

Conclusions

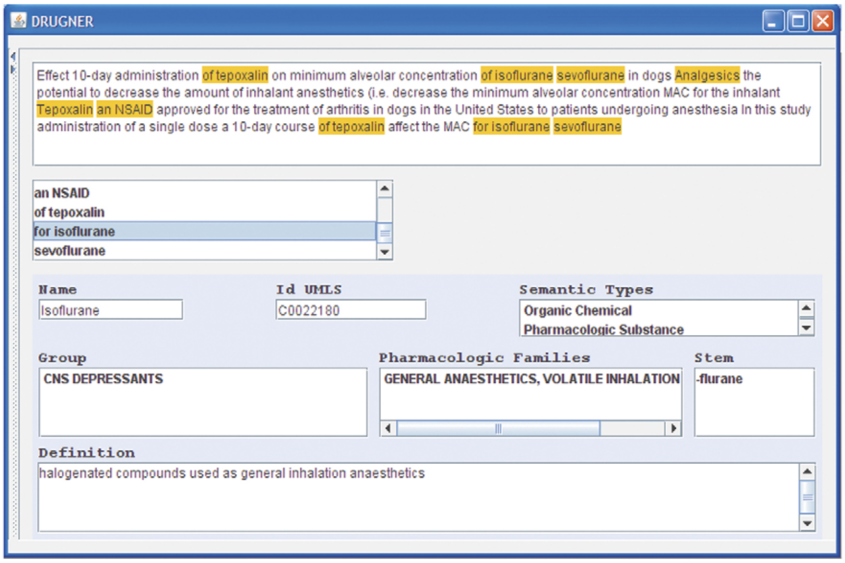
Detecting and classifying drug names occurring in biomedical text is a valuable task in drug discovery knowledge. MMTx is an effective program for the automatic processing of the biomedical texts and has been used extensively for text mining applications in the biomedical domain [23,24]. However, MMTx is not able to provide complete and useful information about pharmacological substances.

The use of stems recommended by the WHOINN Program helps to detect drugs and establish suitable information about the drugs such as the pharmacological or chemical family. Although evaluation reveals that stems alone are not feasible enough in detecting drugs they help to improve the coverage. The stem-based module identifies 255 initial candidates of whom 74 are not pharmacological substances. Most of these wrong identifications are given by the short stems that are too ambiguous to correctly detect drugs.

As outlined previously, the stems are able to correctly classify 74.9% of drugs occurring in the texts. The list of used stems is not exhaustive and does not cover all pharmacological or chemical families. Each year, the WHOINN together with other nomenclature groups[9] establish new stems and rules, in order to govern the classification of new substances and to standardize pharmaceutical nomenclature. Unfortunately, these nomenclature rules have not always been observed when naming a new drug.

On the other hand, linking the stems with the groups of the ATC classification system is an important challenge to be met in future work, because the ATC provides a global standard for classifying

[9] http://www.ama-assn.org/ama/pub/category/4769.html.



Effect 10-day administration of tepoxalin on minimum alveolar concentration of isoflurane sevoflurane in dogs Analgesics the potential to decrease the amount of inhalant anesthetics (i.e. decrease the minimum alveolar concentration MAC for the inhalant Tepoxalin an NSAID approved for the treatment of arthritis in dogs in the United States to patients undergoing anesthesia In this study administration of a single dose a 10-day course of tepoxalin affect the MAC for isoflurane sevoflurane

an NSAID
of tepoxalin
for isoflurane
sevoflurane

Name: Isoflurane
Id UMLS: C0022180
Semantic Types: Organic Chemical; Pharmacologic Substance
Group: CNS DEPRESSANTS
Pharmacologic Families: GENERAL ANAESTHETICS, VOLATILE INHALATION
Stem: -flurane
Definition: halogenated compounds used as general inhalation anaesthetics

Drug Discovery Today

FIGURE 3

DrugNer Tool interface highlights the names of pharmacological substances.

medical substances and serves as a tool for drug research. In addition, this classification system is also used for reporting adverse drug reactions.

Acronyms, frequently used in biomedical texts to rename drugs and other concepts, have not been tracked in this article. The high ambiguity of these terms and the lack of acronym dictionaries make their automatic resolution a difficult task. Nevertheless, it is essential in order to achieve complete coverage in the drug name recognition process.

Resolving drug acronyms, extending the set of stems, including additional clues for those stems that are too short and ambiguous are some challenges for our future research to improve the coverage and the accuracy of drug name recognition and classification tasks. In addition, we believe that the stems could be helpful in the classification of other types of concepts such as organic chemical, enzymes, vitamins and so on.

The proposed system not only obtains automatic detection of drug names occurring in texts, but also allows the annotation of other biomedical concepts (previously detected by MetaMap) with their semantic types in UMLS. Building a manually annotated corpus is a time-consuming, labor-intensive and expensive task. Machine learning methods are not often applied in the biomedical domain because of the lack of training data. GENIA [25] and IProLINK [26] are some of the available biomedical corpora, with their main drawback being that they are limited to proteins and genes and do not contain other types of biomedical concepts. For this reason, we believe that our corpus, DrugDDI, could encourage research on automatic extraction information of drug interactions, adverse drug events and other drug information, occurring in biomedical and pharmacological texts.

Acknowledgements

This work has been partially supported by the Spanish research projects: FIT-350300-2007-75 (Semantic Interoperability in Electronic Health Care) and TIN2007-67407-C03-01 (BRAVO: Advanced Multimodal and Multilingual Question Answering).

References

1 Hale, R. (2005) Text mining: getting more value from literature resources. *Drug Discov Today*. 10, 377–379

2 J. Fickett W. Hayes, Text mining for drug discovery. European Pharmaceutical Contractor, *Autumn* 2004 (2004).

3 Kohn, L.T. *et al.* (1999) *To Err is Human: Building a Safer Health System. Committee on Health Care in America.* Institute of Medicine, National Academy Press

4 Longo, D. *et al.* (2005) The long road to patient safety. *JAMA* 294, 2858–2865

5 Leape, L. and Berwick, D. (2005) Five years alter to err is human. What have we learned? *JAMA* 293, 2384–2390

6 Aronson, A.R. *et al.* (2000) The NLM indexing initiative. *Proc. AMIA Symp.* 20, 17–21

7 Gilman, A.G. *et al.* (1992) *Goodman and Gilman's the Pharmacological Basis of Therapeutics.* Mc Graw Hill

8 Russell, R.G. (2007) Determinants of structure–function relationships among bisphosphonates. *Bone* 40, 21–25

9 Bottorff, M.B. (2006) Statin safety and drug interactions: clinical implications. *Am. J. Cardiol.* 97, 27–31

10 Franzen, K. (2002) Protein names and how to find them. *Int. J. Med. Inform.* 67, 49–61

11 Collier, N. (2000) Extracting the names of genes and gene products with a hidden Markov model. *Proc. COLING* 201–207

12 Tanabe, L. and Wilbur, W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics* 18, 1124–1132

13 Torii, M. and Liu, H. (2006) Headwords and suffixes in biomedical names. *Proc. KDLL* 3886, 29–41

14 Yang, Z. et al. (2008) Exploiting the contextual cues for bio-entity name recognition in biomedical literature. J. Biomed. Inform. In Press, Corrected Proof. Available online 11 January 2008.

15 Ponomareva, N. *et al.* (2007) Biomedical named entity recognition: a poor knowledge HMM-based approach. *NLDB* 382–387

16 Rindflesch, T.C. *et al.* (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* 5, 517–528

17 Kolarik, C. *et al.* (2007) Identification of new drug classification terms in textual resources. *Bioinformatics* 23, 264–272

18 Lindberg, D.A. *et al.* (1993) The Unified Medical Language System. *Methods Inf. Med.* 32, 281–291

19 Aronson, A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.* 17, 17–21

20 Gantner, F. *et al.* (2002) Naming, classification, and trademark selection: implications for market success of pharmaceutical products. *Drug Inf. J.* 36, 807–824

21 World Health Organization Programme on International Nonproprietary Names. (2006) *The use of stems in the selection of International Nonproprietary Names (INN) for pharmaceutical substances*. WHO Press, World Health Organization. WHO/PSM/QSM/2006.3. (http://www.who.int/medicines/services/inn/RevisedFinalStemBook2006.pdf). Accessed January 2008.

22 Baeza-Yates, R. and Ribeiro-Neto, B. (1999) Modern information retrieval, *ACM Press Series*, pp. 73–82

23 Reeve, L.H. *et al.* (2007) The use of domain-specific concepts in biomedical text summarization. *Inform. Process. Manage.* 43, 1765–1776

24 Li, Q. and Brook, W. (2006) Identifying important concepts from medical documents. *J. Biomed. Inf.* 39, 668–679

25 Kim, J.D. *et al.* (2003) GENIA corpus – a semantically annotated corpus for bio-text mining. *Bioinformatics* 19

26 Hua, Z.Z. *et al.* (2004) iProLINK: an integrated protein resource for literature mining. *Comput. Biol. Chem.* 28, 409–416