Identification of new drug classification terms in textual resources

Corinna Kolárik1,*, Martin Hofmann-Apitius1,2, Marc Zimmermann1 and Juliane Fluck1

1Fraunhofer Institute SCAI, Schloß Birlinghoven, 53754 Sankt Augustin and 2Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Dahlmannstrasse 2, 53113 Bonn, Germany

ABSTRACT

Knowledge about biological effects of small molecules helps in the understanding of biological processes and supports the development of new therapeutic agents. DrugBank is a high quality database providing such information about drugs that contains annotation of drug effects and classification of therapeutic effects. However, to broaden the scope of such a database in classifying and annotating drugs, systems for automatic extraction of classification terms and the corresponding annotation of drugs are needed. We have developed an approach for the identification of new terms used in unstructured text that provide information about drug properties. It is based on the identification and extraction of phrases corresponding to lexico-syntactic patterns - so-called Hearst patterns that contain drug names and directly related drug annotation terms. Such phrases could be identified with a high performance in DrugBank text (0.89 F-score) and in Medline abstracts (0.83 F-score). In comparison to DrugBank annotation terminology, a huge amount of new drug annotation terms could be found. The evaluation of terms extracted from Medline showed that 29–53% of them are new valid drug property terms. They could be assigned to existing and new drug property classes not provided by the DrugBank drug annotation. We come to the conclusion that our system can support database content update by providing additionally drug descriptions of pharmacological effects not yet found in databases like DrugBank. Moreover, we propose that automatic normalization of terms improves the annotation and the retrieval of relevant database entries.

Contact: corinna.kolarik@scai.fraunhofer.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The study of molecular and systemic effects caused by biologically active small molecules helps in the understanding of biochemical and molecular processes in the cell. The knowledge created in such studies forms the core of pharmacology. Moreover, the identification of the principles underlying the pharmacological action of small molecules provides the basis for the development of new therapeutic agents. A classification of pharmaceutical effects is useful for the identification of such principles as it allows the establishment of relationships between classes of chemicals defined by chemical features and classes of biological entities defined by biological structures or processes. Such classification of pharmacological effects requires a substantial understanding of the concepts and their relationships used in the domain of pharmacology.

A database containing this information is DrugBank1 (Wishart, 2005; Wishart et al., 2006), an open access, web-enabled database that contains structural, physicochemical, pharmacological and target information of approximately 4300 substances, of which 1177 are approved drugs.

However, most of our knowledge about pharmaceutical effects is communicated through scientific text. In fact, most of the research results concerning drug properties are available only as unstructured text in scientific publications, patents and drug safety reports. This information comprises the entire wealth of factual statements, assumptions, hypotheses and conclusions. While the high expressivity of natural language used in scientific text allows for the representation of complex information, this textual information is hardly accessible for automated data analysis. Nevertheless, advances in text-mining research have significantly improved the automated recognition of named entities relevant for biology, medicine and chemistry, such as proteins and genes which can be identified in text with reasonably high accuracy (e.g. Ananiadou et al., 2004; Hanisch et al., 2005; Krauthammer and Nenadic, 2004; Spasic et al., 2005; Zimmermann et al., 2005). Furthermore, efforts are being made to extract relationships between proteins, drugs, genes or diseases from scientific text (e.g. Chun et al., 2006; Plake et al., 2006; Rindflesh and Fiszman, 2003).

The aim of the work presented here was to automatically identify terms in text that describe pharmacological, systemic, chemical or biological properties of drugs that can be used for their annotation. The drugs themselves were identified with the help of a dictionary-based named entity recognition. Although there are already existing terminologies used for the annotation of drug effects and properties in open source chemical databases (DrugBank, PubChem2), we hypothesize that the annotation of drugs in public databases based on these terminologies is far from being complete. We therefore developed a system aimed at extracting additional annotation information from unstructured text and to compare these annotations with the annotation terms already used in DrugBank.

DrugBank contains pharmacological drug classification terms and identifiers from the Anatomical Therapeutic

*To whom correspondence should be addressed.

1http://redpoll.pharmacy.ualberta.ca/drugbank/
Chemical Classification System (ATC) in the data field ‘Drug Category’. ATC hierarchically classifies therapeutic drugs and divides them into groups according to the organ or organ system on which they act as well as their chemical, pharmacological and therapeutic properties. The system was developed by the World Health Organization (WHO). Additional annotation of drugs in DrugBank can be found in text fields describing their properties or effects in unstructured text using natural language. As a source of unstructured textual information on drugs and their pharmacological effects we used the Medline database. This largest and most comprehensive public literature database comprises the abstracts of the majority of biomedical publications from the 1970s until now.

The technical approach we took in this study is based on work published by Rindflesh (Rindflesh and Fiszman, 2003). This approach identifies phrases representing hypernymic propositional text structures and containing drug information. In this work the UMLS terminology and the tool MetaMap (MMTx) were used to interpret the semantics of identified simple noun phrases. The first formal description of patterns encoding hypernymic propositions was published by Hearst (Hearst, 1992), and since then these patterns are therefore also referred to as Hearst patterns in the literature. Hypernymic propositions are relations that explicitly relate a term to another by an is-a relation. They are encoded by humans by specific lexi-co-syntactic structures. A formal linguistic description defines them as e.g. ‘Noun phrase 1 is (an) Noun phrase 2’, whereas the is-a relation can be represented by various structures (cf 2.3). An example for a typical hypernymic proposition used in the field of pharmacology is ‘Adinazolam is a benzodiazepine derivative’.

We used the Hearst-pattern identification approach to extract terms from text of DrugBank free text fields as well as Medline abstracts that describe drug properties directly related to drugs. In contrast to Rindflesh’s work, we did not utilize MMTx and UMLS terminology because of the minor coverage of the UMLS terms in comparison to the terms found in Medline text, especially those that describe a drug effect onto a protein. We demonstrate that the system we developed can add further useful information to a database. A huge amount of new classification terminology could be found in scientific literature, which has not been used for the annotation of drugs in DrugBank. For eleven drugs we evaluated the extracted terms and identified additional new drug property classes for these drugs not provided by DrugBank or ATC so far. We discuss these findings and propose a generalized approach to literature-based terminology and classification extraction as an automatic pre-processing for a faster construction of terminology resources for pharmacological property classes and richer annotations of drugs.

2 METHODS

2.1 Processing of data from DrugBank and Medline

We extracted the names of all approved drugs (1177) and their synonyms provided by DrugBank to generate a Drug Name Dictionary. It was used to retrieve a drug-specific text corpus from Medline (Medline Text Corpus) with the named entity recognition system ProMiner (Hanisch, 2005).

Another text corpus was extracted from DrugBank (DrugBank Text Corpus), which contains unstructured pharmacological and biological property information for each drug. The following database fields were used: ‘Indication’, ‘Pharmacology’, ‘Mechanism of Action’, ‘Absorption’, ‘Toxicity’, ‘Protein Binding’ and ‘Biotransformation’. The entire corpus consists of 12846 sentences. Since a huge amount of them contain phrases corresponding to Hearst patterns, they were applied for the pattern creation and evaluation studies.

For the comparison of identified textual terms with an existing terminology, the drug classification terms and ATC identifiers used by DrugBank (provided in the ‘Drug Category’ data fields) were extracted. The ATC terms corresponding to the ATC identifiers were obtained from the weblink provided by DrugBank. All classification terms annotated to a certain drug were stored as DrugBank Annotation Terms.

2.2 Noun phrase recognition

Biological and pharmacological descriptions of drug effects are usually represented by nested multi-word terms of complex structure. Such terms often contain protein names that are usually complex by themselves. Furthermore, they comprise inserted numbers, name abbreviations or adjectives written inside or outside of parentheses as can be seen in the following examples:

- ‘competitive beta (1)-selective adrenergic antagonist’
- ‘angiotensin-converting enzyme (ACE) inhibitor’
- ‘inhibitor of the cyclooxygenase pathway of arachidonic acid metabolism’

A classical chunker of base noun phrases (NP) which recognizes series of adjectives and nouns often extracts only part of the complex noun phrase and is in most instances not able to extract such terms completely. Since there is no training corpus for the recognition of these complex noun phrases available, we used a grammar-based system to define pattern rules for the extraction of complex noun phrases. As a basic system we made use of a commercial predefined noun phrase chunker Analytics (provided by TEMIS). It has previously been shown that the software of TEMIS performs similar to other machine-learning-based methods in recognizing base NPs in biomedical domain texts (Wermter et al., 2005). Analytics extracts noun phrases like ‘treatment of insomnia’, ‘alpha receptor subtype’ or ‘GABA-BZ receptor complex’. Furthermore, the software allows a user-specific creation of new grammar rules which can easily be incorporated into the already existing system. We extended Analytics by combining the existing noun phrase patterns with additional rules to enable the identification of complex noun phrases. This extension forms the ExtAnalytics chunker. Some examples of patterns and corresponding noun phrases extracted with the original Analytics chunker (Analytics) or with the extended Analytics chunker (ExtAnalytics) are shown in Table 1.

2.3 Hearst phrase extraction

Common lexi-co-syntactic structures defining Hearst patterns (HP) were taken from (Cimiano et al., 2005; Hearst, 1992 and Rindflesh and Fiszman, 2003) and are described in the following lines:

http://umlsinfo.nlm.nih.gov/
http://mmtx.nlm.nih.gov/docs.shtml

www.temis-group.com
that are described by NP0. In our case these are drugs.

One chunkers were created according to the pattern described above. One noun phrases and Hearst patterns were established and two Hearst phrase (hypernym)—e.g. a drug class. NP1, NP2,..., and NPn are NP0.

Table 2. Term variation types and term variation examples normalized with UMLS lexical tool Lexical Variant Generator (lvg2006)8

Propositions involving verbs:

NP1 is (a | an) NP0
NP1 is one of (the | a | an) NP0
NP1, NP2,..., and NPn are NP0

Propositions of appositive structure:

NP0 such as NP1, NP2,..., NPn-1 (and | or) NPn
such NP0 as NP1, NP2,..., NPn-1 (and | or) NPn
NP0 (including | especially | like) NP1
NP0 for example NP1, NP2,..., NPn-1 (and | or) NPn
NP1, NP2,..., NPn (and | or) other NP0
NP1, (a | an) NP0

NP0 stands for a noun phrase that represents a general term (hypernym)—e.g. a drug class. NP1,..., NPn stand for terms (hyponyms) that are described by NP0. In our case these are drugs.

Rules set for the identification of lexico-syntactic patterns for complex noun phrases and Hearst patterns were established and two Hearst phrase chunckers were created according to the pattern described above. One incorporates the original Analytics (Analytics-HP chunker) and the other uses the extended NP chunker (ExtAnalytics-HP chunker). Hearst phrases contained in the DrugBank text corpus were used as standard for the pattern evaluation.

The following criteria were applied for the annotation of the extracted Hearst phrases:

- A true positive phrase needs to fit syntactically to the given Hearst patterns.
- Semantically the phrase content needs to make sense in a pharmacological way, i.e. it should be a part of an explicit description providing some biological, chemical or pharmacological properties of a drug or an enumeration of drugs. It does not need

Examples of extracted Hearst phrases and their annotations are given in Tables 3 and 4. For the evaluation of the automatically extracted Hearst phrases we calculated precision, recall and F-score (1).

\[
F–score = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]

2.4 Hearst phrase fragmentation

Extracted Hearst phrases semantically not containing drug-specific information were automatically filtered out after their automatic extraction, i.e. all phrases that contain drug names covered by the Drug Name Dictionary were further processed. Terms describing drug properties were extracted from the remaining phrases in the next step. The phrases were automatically split and assigned to their meaning parts i.e. drug names (the hyponyms)—NP1, NP2,..., NPn and terms describing drug properties or effects (hypernym)—NP0. Partioning of the phrase ‘Adinazolam is a benzodiazepine derivative’ given as an example would result in: ‘Adinazolam’—a drug and ‘benzodiazepine derivative’—a drug property term. The latter one—the NP0 of the Hearst phrase—is the term we are interested in and that is used for further processing and analyses.

2.5 Term normalization

In texts as well as in terminologies, different variants of terms, representing one concept are used. They can occur as orthographical, morphological, syntactic and lexico-semantic term variations or term abbreviations (Nenadic et al., 2004; Savary and Jacquemin, 2003). To compare terms between the two examined resources—text and terminology—we needed to address this problem.

In (Sarkar et al., 2003) four strategies—exact string match, normalization, the tool MetaMap and Blast-based matching—were investigated for their potential to map terms of Gene Ontology8 to

Table 3. Examples of extracted Hearst phrases (HP) from Medline, potential annotation terms contained within the phrases and their evaluation

<table>
<thead>
<tr>
<th>Hearst phrase</th>
<th>Drug name</th>
<th>Potential classification term</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘NSAIDs like ibuprofen**</td>
<td>Ibuprofen</td>
<td>NSAIDs</td>
<td>Already used in DrugBank Annotation Terms</td>
</tr>
<tr>
<td>ketoprofen, naproxen and other allied drugs**</td>
<td>Ketoprofen, naproxen</td>
<td>Allied drugs</td>
<td>Not relevant</td>
</tr>
<tr>
<td>ibuprofen, a preferential excretion*</td>
<td>Ibuprofen</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>ibuprofen, a prostaglandin synthase inhibitor**</td>
<td>Ibuprofen</td>
<td>Prostaglandin synthase inhibitor</td>
<td>New drug classification term</td>
</tr>
</tbody>
</table>

*False positive Hearst phrase, **True positive Hearst phrase.

terms of UMLS. It was shown that the term normalization approach performed best, both in recall and precision. In a second study (Nenadic et al., 2004) normalization of biomedical terms was also applied successfully to map between different surface realizations belonging to one concept.

Since we wanted to compare terms with each other we decided to normalize all terms—those extracted from text as well as the drug DrugBank Annotation Terms. The following processing steps were applied to each term: first it was tokenized with a Genia tagger-based tokenizer and POS-tagged. Nouns and adjectives lemmatized by the Genia tagger (Tsuruoka et al., 2005) were transformed into a canonical representative form with the UMLS lexical tool Lexical Variant Generator (lvg2006) provided by the National Library of Medicine (NLM). The following term variations were normalized using lvg2006:

Syntactic term variants like ‘inhibitor of protein synthesis’ were automatically normalized to ‘protein synthesis inhibitor’ by a heuristic.

Furthermore, we manually generated a dictionary of synonymous expressions and synonyms for automated mapping to a canonical term form. Synonymous head nouns are for example ‘agent’, ‘drug’ or ‘compound’ which can be automatically mapped to each other because they share an equivalent meaning in pharmacological descriptions. Examples for semantic equivalent term parts that can differ in the number of words are ‘blocker’ versus. ‘blocking agent’.

2.6 Annotation comparison and evaluation

The normalized terms extracted from DrugBank Hearst phrases or from Medline Hearst phrases were compared with the normalized DrugBank Annotation Terms. To allow further spelling variants, this step was not done by an exact string matching but with the named entity recognition tool ProMiner (Hanisch et al., 2005). ProMiner uses term lists (dictionaries) in an approximate search and is able to cope with permutations and nested terms.

For this purpose, the normalized terms extracted from DrugBank text and Medline text corpora were incorporated into ProMiner dictionaries. These dictionaries were then compared to the normalized DrugBank Annotation Terms of DrugBank using ProMiner technology. Those terms, extracted from text, that were not found by the system were analyzed and evaluated manually to consider their usage as new drug classification terms. The constraints that must be fulfilled by a term to be considered as a new drug annotation term are as follows:

1. The term was not found in DrugBank Annotation Terms.
2. It should contain relevant pharmacological, biological effect or chemical property specific information about a drug.

Examples of extracted Hearst phrases from Medline abstracts with the extracted classification terms and the explanation of their evaluation are given in Table 3.

3 RESULTS

3.1 Workflow for the extraction of drug annotation terms

Our aim was the extraction of drug property or effect-describing terms from text that could be used as drug classification terms to support the extension of terminologies and ontologies in the area of pharmacology. Text-mining methods were utilized for the extraction of such terms that are directly related to drugs. We developed a generic system with which drug classification terms from different resources can be extracted and compared. A schematic representation of the workflow of the approach we developed is shown in Figure 2.

The DrugBank text corpus derived from text fields and the drug-specific Medline corpus were used as input for the extraction of textual drug annotations. The DrugBank text corpus was used as standard for the Hearst-pattern definition. Hearst phrases with different structural forms of an is-a relation between drug names and drug property-describing terms were extracted from the two text corpora (Fig. 2, Step 1). In these phrases drug property terms were identified specifically for each drug (Fig. 2, Step 2). They were normalized and compared with the normalized DrugBank annotation terms (Fig. 2, Steps 3 and 4). New terms not yet present in the annotation term list are evaluated for both text corpora and can be used as input for new classifications of drugs.

3.2 Evaluation of two Hearst phrase chunkers on the DrugBank corpus

As a prerequisite to the term extraction and the subsequent term comparison, the performance of the two Hearst phrase (HP) chunkers, one using the Analytics chunker (Analytics-HP) and the other using ExtAnalytics-HP chunker, on the DrugBank corpus was tested.

Table 4. Results of the Hearst phrase extraction with 'Analytics-HP' and 'ExtAnalytics-HP', true positives (TP), false negatives (FN), false positive (FP) variants (phrase examples are shown—underlined text corresponds to standard phrase and bold text in italics belongs to automatically found phrases)

<table>
<thead>
<tr>
<th>Phrase examples</th>
<th>Analytics-HP</th>
<th>ExtAnalytics-HP</th>
<th>Phrase matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>clavithromycin, a macrolide antibiotic</td>
<td>34</td>
<td>500</td>
<td>(1) FP partial Phrase is too short compared to standard</td>
</tr>
<tr>
<td>availability of dopamine, a brain chemical</td>
<td>65</td>
<td>26</td>
<td>(2) FP too long Phrase is too long, at the beginning or at the end</td>
</tr>
<tr>
<td>ibuprofen, a 61-year-old woman</td>
<td>417</td>
<td>572</td>
<td>(3) FP wrong content Phrase matches Hearst-pattern, content makes no sense</td>
</tr>
<tr>
<td>FN</td>
<td>155</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.69</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.83</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>F-score</td>
<td>0.75</td>
<td>0.89</td>
<td></td>
</tr>
</tbody>
</table>
and the second containing the ExtAnalytics chunker (ExtAnalytics-HP), was evaluated.

Since the DrugBank text corpus contains many Hearst phrases comprising drug effect or property descriptions, half of the corpus was taken for manual annotation of text phrases corresponding to Hearst-patterns. A total of 572 phrases containing drug names from DrugBank were defined as a standard Hearst phrase corpus. The analysis of the extracted phrases was done semi-automatically. All those not exactly matching to phrases in the standard corpus were inspected by hand and were classified into false positives (FP) and true positives (TP). For annotation rules see Section 2.3. The result is shown in Table 4.

As can be seen in Table 4 the application of the ExtAnalytics-Hearst Phrase chunker increased precision, recall and F-score significantly in comparison to the Analytics-Hearst Phrase chunker. We can demonstrate that the extension of the Analytics noun phrase chunker leads to an improvement in the identification correctness of the complete Hearst phrase. This has also an impact on the extraction quality of complex terms describing drug properties. Even complex and long terms were extracted correctly. The extraction of phrases that are too short is based on errors of the Part-of-speech tagging.

Since we retrieved better results with the ExtAnalytics-Hearst Phrase chunker it was applied for all further extraction tasks and analyses.

3.3 Information content of DrugBank: drug category versus textual ‘Hearst-Pattern’ information

3.3.1 Structured information—DrugBank For all drugs we could extract 1073 DrugBank Annotation Terms from the Drug Category fields. ATC identifiers were automatically translated into the corresponding classification term. Figure 3 shows the distribution of the number of DrugBank Annotation Terms assigned to drugs.

A large number of drugs were annotated with four or five drug category terms. We found out that the term variation problem also occurs in databases. Here different term variants for the same drug category were assigned to different drugs.
Identification of new drug classification terms

Fig. 3. Distribution of the number of annotation terms assigned to drugs in DrugBank—dark columns: DrugBank Annotation Terms; light columns: Terms extracted from DrugBank text.

Table 5. List of numbers of drug entries corresponding to different search term variants of the concept ‘nonsteroidal anti inflammatory agent’, term differences are marked in bold

<table>
<thead>
<tr>
<th>Term variants of the concept ‘nonsteroidal anti inflammatory agent’</th>
<th>Query result: number of entry matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-steroidal Anti-inflammatory agent</td>
<td>0</td>
</tr>
<tr>
<td>Nonsteroidal Anti-inflammatory agents</td>
<td>2</td>
</tr>
<tr>
<td>Nonsteroidal Anti-inflammatory agent</td>
<td>3</td>
</tr>
<tr>
<td>Nonsteroidal Anti-inflammatory drugs</td>
<td>4</td>
</tr>
<tr>
<td>Nonsteroidal Anti-inflammatory agents</td>
<td>9</td>
</tr>
<tr>
<td>Nonsteroidal Anti-inflammatory drug</td>
<td>12</td>
</tr>
<tr>
<td>Nonsteroidal Anti-Inflammatory</td>
<td>15</td>
</tr>
<tr>
<td>NSAIDs</td>
<td>22</td>
</tr>
<tr>
<td>NSAID</td>
<td>29</td>
</tr>
<tr>
<td>AntiInflammatory agents, Nonsteroidal</td>
<td>30</td>
</tr>
<tr>
<td>Anti-inflammatory agents, Non-steroidal</td>
<td>35</td>
</tr>
</tbody>
</table>

Through normalization and mapping of the term variants, we reduced the overall number of annotation terms to 966, which is a decrease by ~10%. The found term variant types are of morphological, orthographical and lexico-syntactic form.

Terms used in databases and that are accessible via a database text query have a high impact on the retrieval of information. To demonstrate this effect, DrugBank was queried using the plural and singular forms as well as different spelling or lexico-semantical term variants of the classification concept ‘nonsteroidal anti inflammatory agent’ (see Table 5).

As Table 5 shows, the result of the database query depends strongly on the applied search term. Term variants seem not to be mapped to a controlled vocabulary in DrugBank. The terms of the ATC system were not used within DrugBank, so that only their identifiers are accessible via a database query.

3.3.2 The impact of textual DrugBank information A total of 1164 Hearst phrases containing drug names were automatically obtained with the ExtAnalytics Hearst Phrase chunker from the DrugBank text corpus. They comprise 860 terms, which were reduced to 829 after normalization. This shows that, even in a more structured database text, different term variants are used for the description of the same concept.

Figure 3 shows that for most drugs one or two terms were extracted from DrugBank text. The automatic comparison between DrugBank Annotation Terms and drug classification terms derived from the DrugBank text corpus showed that 84% (694) of the latter terms have not been used as drug annotation terms so far. They contain new relevant information about drugs. The terms give more detailed information, e.g. about the protein that the drug influences, the mechanism of the drug effect (e.g. ‘irreversible proton pump inhibitor’), the natural resource of the drug or even describe a new reaction mechanism of drugs onto protein targets contained in the DrugBank Annotation Terms (histamine h2 agonist).

This experiment shows that even database text contains drug classification terms that have the potential of being annotation terms and that can add additional information to a terminology. The advantage of using them as normalized annotation terms is to provide the database user with better and more structured information. The usage of controlled terms would lead to a higher degree of data consistency and improved querying capabilities.

3.4 Drug classification information in Medline Abstracts

The main information source beside databases is scientific text. We therefore used our Hearst phrase chunker to retrieve drug classification information from Medline abstracts. The Hearst phrase and term extraction process was the same as for the DrugBank text corpus. The Drug Name Dictionary generated from DrugBank entries was used to retrieve drug-specific Medline abstracts and drug-relevant Hearst phrases. We obtained 2231823 drug-specific abstracts and ~6% of these abstracts contained Hearst phrases describing drug effects or properties.

3.4.1 Hearst Phrase chunker evaluation on a Medline text corpus In Section 3.2, we showed that the ExtAnalytics Hearst phrase chunker performed well on the standard corpus of DrugBank. To assess the extraction quality of the ExtAnalytics Hearst phrase chunker applied on the Medline abstract corpora, it was evaluated on an abstract corpus for the selected drug Ibuprofen. It contains 1089 arbitrarily chosen abstracts in which 101 Hearst phrases containing pharmacological information about Ibuprofen were manually annotated and compared with the automatic extraction result (see Table 6).

Compared to the standard corpus of DrugBank, we achieved a higher recall but a lower precision with the Medline corpus. The proportion of true positives compared to the phrase number of the Medline standard is similar (~78%) to that obtained for DrugBank. We got a lower fraction of FP partial, but a higher relative amount of FP too long and FP with wrong content.

3.4.2 Term extraction All extracted Hearst phrases were filtered. Only those were processed that contain drug names and for which their synonyms are provided by the Drug Name Dictionary.
Table 6. Evaluation of the ExtAnalytics Hearst Phrase chunker on an abstract corpus for Ibuprofen from Medline (for descriptions of false positives (FP) variants see Table 4)

<table>
<thead>
<tr>
<th></th>
<th>Medline ibuprofen corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of manually annotated HP phrases (Standard corpus)</td>
<td>101</td>
</tr>
<tr>
<td>Total number of automatically extracted HP phrases</td>
<td>108</td>
</tr>
<tr>
<td>TP</td>
<td>79</td>
</tr>
<tr>
<td>(1) FP partial</td>
<td>2</td>
</tr>
<tr>
<td>(2) FP too long</td>
<td>18</td>
</tr>
<tr>
<td>(3) FP wrong content</td>
<td>9</td>
</tr>
<tr>
<td>FN</td>
<td>2</td>
</tr>
<tr>
<td>Recall</td>
<td>0.73</td>
</tr>
<tr>
<td>Precision</td>
<td>0.97</td>
</tr>
<tr>
<td>F-score</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 7. Number of extracted Hearst phrases (HPs), unique identified terms and unique normalized terms

<table>
<thead>
<tr>
<th>Drugs</th>
<th>HPs</th>
<th>Number of extracted unique terms</th>
<th>Number of unique terms after normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibuprofen</td>
<td>597</td>
<td>404</td>
<td>329</td>
</tr>
<tr>
<td>Diclofenac</td>
<td>309</td>
<td>203</td>
<td>156</td>
</tr>
<tr>
<td>Atenolol</td>
<td>235</td>
<td>175</td>
<td>138</td>
</tr>
<tr>
<td>Mitomycin</td>
<td>320</td>
<td>223</td>
<td>203</td>
</tr>
<tr>
<td>Metoprolol</td>
<td>207</td>
<td>152</td>
<td>125</td>
</tr>
<tr>
<td>Tamoxifen</td>
<td>1003</td>
<td>635</td>
<td>527</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>529</td>
<td>363</td>
<td>316</td>
</tr>
<tr>
<td>Nifedipine</td>
<td>975</td>
<td>601</td>
<td>507</td>
</tr>
<tr>
<td>Chlorpromacine</td>
<td>427</td>
<td>320</td>
<td>271</td>
</tr>
<tr>
<td>Phentolamine</td>
<td>321</td>
<td>186</td>
<td>154</td>
</tr>
<tr>
<td>Midazolam</td>
<td>403</td>
<td>284</td>
<td>252</td>
</tr>
</tbody>
</table>

Dictionary. In the following analysis, we evaluated the extracted terms for 11 drugs mentioned in the context of various therapeutic areas and for which a considerable number of Medline abstracts (>4200) were retrieved. For each of the 11 drugs, the number of extracted Hearst phrases, unique terms (all redundant terms were removed), and unique normalized terms are listed (see Table 7). It shows, that a huge number of Hearst phrases and potential drug annotation terms could be extracted from Medline abstracts. Normalization reduced the number of terms by a mean of 16%, which is higher than the reduction value of DrugBank Annotation Terms by normalization. This indicates that terms used in free text are more variable than terms used in the database. The most frequent variations were of morphological, orthographical and lexico-semantic form.

3.4.3 Structured versus unstructured information resources—comparison of terms from Medline and DrugBank Annotation Terminology In the previous sections, we described how terms from DrugBank text, Medline abstracts and DrugBank Annotation Terms were extracted and normalized. Our interest was to compare these terms with each other to assess their information content and to find new terms describing drug properties. The terms extracted from text were normalized and used as a dictionary to search for corresponding terms in the list of annotation terms. This procedure was done separately for each drug. Figures 4 and 5 show the result of the comparison between extracted normalized terms of DrugBank terminology, DrugBank text originating terms and Medline terms.
A high portion—29.53%—of terms extracted from Medline abstracts could serve as annotation terms and have not been used in DrugBank so far. This means that they would add new information to the database if they were used for drug annotation. The comparison of extracted terms and DrugBank Annotation Terms shows that only a limited number of terms originating from text—only 1.3–6.4%—are already in use in the database annotation field (see Fig. 5). Terms that were not considered as new originate either from false positive Hearst phrases with wrong content, from too long phrases that incorporate a term already existing in the DrugBank Annotation Terminology, or they contain non-relevant additional information about a drug.

A deeper analysis of the valid new terms shows that they can be assigned to various drug property classes. A list of classification terms and examples from DrugBank and Medline abstracts is given in Table 8.

The ATC drug classification system as well as the internal annotation types of DrugBank is restricted to some drug property classes, like pharmacological property or chemical structure class. As can be seen in Table 8, some of the new terms found in text can be assigned to new annotation categories not contained in the DrugBank annotation terminology. With that we did not only find additional drug property terms in text, but could also establish new classes of information.

## 4 DISCUSSION

In the work presented here we wanted to extract drug classification terms that describe pharmacological, chemical and biological properties of drugs for their annotation. Two resources containing drug classification terms have been used for this purpose: DrugBank and free text from Medline. Whereas drug classification terminology in databases or even the drug ATC classification system often focuses on certain aspects in drug classification, scientific text contains many more annotation term varieties in which different authors are interested in. Extracted textual annotations could enlarge the drug classification spectrum and speed up the annotation process, even if not all found terms might be useful.

To identify drug annotations in text that are directly associated with drugs we utilized a NLP technique with which phrases corresponding to specific lexicoco-syntactic structures—Hearst-patterns—can be detected. Since drug property-description terms can be very complex, we successfully extended the rule pattern for noun phrases going beyond the extraction of basic noun phrases and incorporated it into the Hearst phrase chunker. Drug-specific Hearst phrases were selected with the help of a Drug Name Dictionary generated from DrugBank. We could show that the pattern extension significantly increases the F-score by 14% on the standard DrugBank corpus compared to the HP Hearst chunker applying basic noun phrases. The recognition of too long phrases is also increased but correct annotation terms could be manually extracted from such phrases and are more useful for database curators than too short phrases.

The comparison of the performance in identifying hypernymic propositions of our UMLS-independent generic system with SemSpec, developed by Rindflesh’s group, is difficult because the pattern for hypernymic propositions was not completely explained and thus may differ from our patterns. Furthermore, the two systems were evaluated on two different text corpora. We can, however, compare the precision reported by Rindflesh et al. (83% F-score), with the values we obtained on the DrugBank standard (89% F-score) and on the Medline standard corpus (83% F-score). Our generic system has a similar extraction quality on Medline abstracts as the SemSpec system Rindflesh et al. developed. For semantic term resolution and evaluation, we could not use UMLS because the UMLS terminology related to drug effects onto proteins did not match the grade of specificity and granularity of terms we extracted from text.

Beyond the mere identification of hypernymic propositions, our aim was to extract new drug descriptive terms and to compare their use in the two sources available to us on principle, namely DrugBank Annotation Terms and terms extracted from text. Since a lot of relevant concepts are represented by various term variants, we had to normalize all terms we worked with. Even in the DrugBank Annotation Terms variants have been found, which have a negative impact on the retrieval of relevant data from DrugBank. The automatic normalization of the DrugBank Annotation Terms reduced their number by 10% and in Medline text a reduction of different terms of 16% could be achieved. The term comparison between DrugBank text terms and DrugBank Term Annotation Terms showed that there is little overlap between them. Eighty-four percent of the textual terms contain additional annotations that could be used as drug classification terms and very likely could enhance the information content of annotations.

The comparison between Medline and terminology terms showed that only a small percentage of terms (1.3–6.4%)...
extracted from Medline have been used in DrugBank so far. In a further evaluation of the relevance of these annotation terms ~29–53% of the terms were identified as valid new drug classifications. Most of them are more specific and also belong to new classification types not yet used in DrugBank.

In future work, we will put effort into the automatic evaluation of the extracted terms to excerpt relevant classifications. We plan to provide these drug classifications to DrugBank or will make it available via an own website. Furthermore, we will expand our named entity recognition beyond the drug name dictionary used here for the recognition of chemical names to extract chemical classification information relevant for pharmaceutical and biological research. Moreover, we will investigate if the recognition of Hearst patterns and complex noun phrase patterns can be optimized with other technologies, such as machine-learning techniques like conditional random fields. Here we will use the annotated pharmacology domain corpus of complex noun phrases and propositions which has been created.

5 CONCLUSIONS

Apart from the difficulty in finding specific information in the increasing amount of scientific text, the explosion in the amount of data and information as well as the usage of various terminologies within one research field is a problem for information retrieval, data curation and annotation of data in repositories (Blaschke and Hirschmann, 2006).

As we show in this article, Hearst-phrase extraction in combination with the recognition of drug names can be used to find new drug annotation terms that are not yet applied for drugs in a database and that are likely to add valuable annotation information. This approach can be utilized to assist database curators to scan free text data resources which represent the most up-to-date sources of information. It is generic and can be applied also to other areas by exchanging only the named entity recognition to focus on the Hearst patterns of interest.

ACKNOWLEDGEMENTS

This work was technically supported by Temis Deutschland GmbH and financially supported by the Bonn-Aachen International Center for Information Technology (B-IT). Furthermore, we would like to thank the anonymous reviewers for their work. One very detailed review especially pinpointed critical points in our work and helped to clarify some of our statements. Many thanks also to Jonathan Sleeman for proofreading the script.

Conflict of Interest: none declared.

REFERENCES