BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 4. 460–494 doi:10.1093/bib/bbs018

Mining the pharmacogenomics literature—a survey of the state of the art

Udo Hahn, K. Bretonnel Cohen, Yael Garten and Nigam H. Shah

Submitted: 18th November 2011; Received (in revised form): 23rd March 2012

Abstract

This article surveys efforts on text mining of the pharmacogenomics literature, mainly from the period 2008 to 2011. Pharmacogenomics (or pharmacogenetics) is the field that studies how human genetic variation impacts drug response. Therefore, publications span the intersection of research in genotypes, phenotypes and pharmacology, a topic that has increasingly become a focus of active research in recent years. This survey covers efforts dealing with the automatic recognition of relevant named entities (e.g. genes, gene variants and proteins, diseases and other pathological phenomena, drugs and other chemicals relevant for medical treatment), as well as various forms of relations between them. A wide range of text genres is considered, such as scientific publications (abstracts, as well as full texts), patent texts and clinical narratives. We also discuss infrastructure and resources needed for advanced text analytics, e.g. document corpora annotated with corresponding semantic metadata (gold standards and training data), biomedical terminologies and ontologies providing domain-specific background knowledge at different levels of formality and specificity, software architectures for building complex and scalable text analytics pipelines and Web services grounded to them, as well as comprehensive ways to disseminate and interact with the typically huge amounts of semiformal knowledge structures extracted by text mining tools. Finally, we consider some of the novel applications that have already been developed in the field of pharmacogenomic text mining and point out perspectives for future research.

Keywords: *text mining; information extraction; knowledge discovery from texts; text analytics; biomedical natural language processing; pharmacogenomics; pharmacogenetics*

INTRODUCTION

Among the many promises of the *Human Genome Project* (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml) was the prospect of treating human diseases in a much more focused manner and targeting specific parameters and conditions based on finely grained phenotype or even individual genetic profiles. Such novel opportunities have shaped the idea of 'individualized medicine' and have led to the emergence of a field called

Corresponding author. Udo Hahn. Jena University Language and Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany. Phone: +49 3641 944 320; Fax: +49 3641 944 321; E-mail: udo.hahn@uni-jena.de

Udo Hahn is a Full Professor of Computational Linguistics and Language Technology at Friedrich-Schiller-Universität Jena, Germany, where he leads the Jena University Language and Information Engineering (JULIE) Lab. His research interests cover statistical and machine learning-based methodologies for natural language processing, as well as knowledge representation and formal reasoning, including ontology engineering. His system-oriented work focuses on text analytics (such as information extraction, text mining, document retrieval and text summarization), with particular emphasis on applications in the life sciences.

Kevin Bretonnel Cohen leads the Biomedical Text Mining Group in the Computational Bioscience Program at the University of Colorado School of Medicine. He is the chairman of the Association for Computational Linguistics Special Interest Group on biomedical natural language processing. His research covers most areas of text mining in the biomedical domain.

Yael Garten completed her PhD in Biomedical Informatics at Stanford University, with dissertation work on 'Text Mining of the scientific literature to identify pharmacogenomic interactions'. Following a year as a Research Scientist working with PharmGKB, she is now a Senior Data Scientist at LinkedIn Corporation, USA.

Nigam H. Shah is an Assistant Professor of Medicine (Biomedical Informatics) at the Stanford School of Medicine. His research is focused on developing applications of bio-ontologies, specifically building novel approaches to annotate, index and analyze diverse information types available in biomedicine. Shah holds an MBBS from Baroda Medical College, India, a PhD from Penn State University, USA, and completed postdoctoral training at the Stanford Medical School.

'pharmacogenomics' [1]. It can be defined as 'the study of how an individual's genetic variation impacts (originally: genetic inheritance affects) the body's response to drugs. The term comes from the words pharmacology and genomics and is thus the intersection of pharmaceuticals and genetics' (http://www.ornl.gov/sci/techresources/Human_Genome/medicine/pharma.shtml).

This combination of hitherto almost unrelated scientific areas, on the one hand, opens up a plethora of challenging research questions. On the other hand, researchers in each of the camps involved often face serious information deficits regarding what concerns previous work and progress in the discipline(s) other than their own research speciality. The tremendous heterogeneity of the relevant themes to be covered here—e.g. information about genes, gene variants and proteins, diseases and other pathological phenomena, drugs and other chemicals, as well as various forms of semantic relations between them, constitute an information acquisition and knowledge management problem that gave rise to the emerging field of 'pharmacogenomic text mining' (see Garten *et al.* [2], for a recent survey). Work in this field builds on and complements research previously carried out in closely related areas such as biomedical text mining (see e.g. Krallinger *et al.* [3], for a survey, and Rodriguez-Esteban [4], for a basic tutorial) and clinical text mining (see e.g. Meystre *et al.* [5] and Demner-Fushman *et al.* [6], for surveys). Pharmacogenomic text mining is concerned with automatically locating and integrating biomedical knowledge that is currently scattered between millions of scientific publications and thousands of highly specialized, mostly disconnected, public, as well as proprietary (e.g. clinical) databases, often making use of a large variety of biomedical lexicons, terminologies and ontologies.

Text mining is a field of applied natural language processing that deals with automatically extracting relevant information (single facts and assertions, complex propositional or even hypothetical statements, etc.) from written texts. There are excellent and comprehensive text books available on the methodological foundations of natural language processing (NLP), information retrieval (IR), text mining (TM) and information extraction (IE), such as those by Manning and Schütze [7], Jurafsky and Martin [8], Manning *et al.* [9], Jackson and Moulinier [10], Feldman and Sanger [11] and Weiss *et al.* [12].

The TM task is much more ambitious than finding relevant documents, the challenge that characterizes IR systems such as Google and (life science-centric) PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). The latter is based on MEDLINE (http://www.nlm.nih.gov/bsd/pmresources.html), the largest document repository for textual information in the life sciences worldwide, with currently more than 21 M bibliographic units, most of them accompanied by scientific abstracts, i.e. the author-supplied surrogate of the original full text. It is complemented by PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/), a continuously growing full text archive for a subset of PubMed documents that currently comprises more than 2.3 M documents. When applied within the domains of medicine, biology and pharmacology, TM primarily deals with scientific publications (historically in the form of abstracts supplied by PubMed, but increasingly in the form of full texts as from PubMed Central or Open Access publishers such as BioMed Central (http://www.biomedcentral.com/), the latter currently has 220 biomedical online journals in its portfolio). Clinical narratives (such as discharge summaries, admission, X-ray, surgery or pathology reports), patent texts, drug adverse effect reports (including drug package inserts) and quite recent contributions to collaboratively written and maintained document collections on the Web (such as Wikipedia-type documents, see, e.g. [13]), blogs, mailing lists or other sorts of electronic discussion fora, many of them dealing with public health-related or health consumer issues, are also increasingly being considered for analytic purposes (see, e.g. [14]).

Text mining (when defined as an 'information extraction' task proper) identifies

- mentions of named entity types, considered relevant—in this case for the life sciences (genes, gene variants, proteins, protein mutations, diseases, treatments, drugs, chemicals, tissues, species, etc.) or
- relationships among those named entities, either explicitly lexicalized, and thus 'semantically typed relations' holding between named entities [e.g. various sorts of protein–protein interactions (PPIs), drug–drug interactions (DDIs), locational (is-contained-in, has-location, etc.) or functional relations (is-caused-by, is-treatment-for, etc.)], or purely associative, and thus semantically underspecified relations.

For example, given a sentence such as 'NF-kappa B may activate the production of TNF-α,' an ideal TM system would characterize 'NF-kappa B' as

being an instance of the named entity type (or class) 'TranscriptionFactor' and 'TNF-α' as being an instance of the class 'Gene'. In addition to named entity recognition, the text miner should also identify two semantic relations, one embedded in the other. First, a binary relation of the type 'PositiveRegulation' should be determined, consisting of two arguments, namely 'NF-kappa B', the driver (or agent) of that regulation process and second, something that is acted upon—in our example not just 'TNF-α' but 'the expression of TNF-α', which we might encode as the unary relation 'GeneExpression(TNF-α)'. These pieces can be combined in an expression that reads as 'PositiveRegulation [NF-kappa B, GeneExpression (TNF-α)]', the final outcome of relation extraction. Note that many linguistic subtleties (such as the fact that this event might only potentially occur, indicated by 'may') are discarded from these two tasks in their standard setting. However, there is also research going on that tries to incorporate indications of certainty, plausibility, believability, trustability and speculation, which modify the strength of confidence in an extracted relation (see, e.g. [15–17]), and further work reported, e.g. at the CoNLL 2010 Shared Task Learning to Detect Hedges and their Scope in Natural Language Text (http://www.inf.u-szeged.hu/rgai/conll2010st/), the 2010 ACL Workshop on Negation and Speculation in Natural Language Processing (http://www.clips.ua.ac.be/NeSpNLP2010/), and in the 2009 (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/) and 2011 (http://2011.bionlp-st.org/) BioNLP Shared Tasks on Event Extraction (see [18] and [19], respectively) associated with the Association for Computational Linguistics BioNLP workshops (SIGBioMed is the Special Interest Group for Biomedical Natural Language Processing of the Association for Computational Linguistics, which among other activities, hosts workshops at ACL annual meetings. It is dedicated to NLP in the biological, biomedical, and clinical domain; their website is located at http://www.sigbiomed.org/r02.01.11/index.php?title=Main_Page).

Text mining also has a second definition, where the focus is on the recognition of implicit rather than explicit (as above) relations among instances of named entities. Typically, direct lexicalizations of these relations are missing in the underlying documents. Work within this framework aims fundamentally at finding hitherto unknown, i.e. 'new' knowledge in documents, which includes the generation of hypotheses that have to be experimentally tested. This heuristic approach to 'knowledge discovery' from texts is either based on distributional co-occurrence characteristics of terms or requires logical reasoning over text-derived knowledge representation structures (see Section 'Knowledge discovery: mining implicit and novel information'). Still, the vast majority of TM systems focus on extracting explicit information from texts, i.e. they are true IE engines.

Literature in the field of pharmacogenomics bridges at least three disciplines: pharmacology (drugs and other treatment-related chemical substances), medicine (diseases or pathological phenomena, treatments, including medication and tests) and molecular biology in all of its forms. The complexity and diversity of themes that have to be handled at the overlap of these areas is unmatched in the field of IE/TM, in particular when compared with mainstream NLP research, which focuses mostly on newspaper and newswire analysis [see classic research carried out within the frameworks of the *Message Understanding Conference* (MUC), e.g. http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html, and more recent work in the frameworks of *Automatic Content Extraction* (ACE) (http://www.itl.nist.gov/iad/mig/tests/ace/) and the *Text Analysis Conference* (TAC) (http://www.nist.gov/tac/)]. When we directly compare the task of locating person names in a newspaper article with performing the same identification task for genes or proteins in a biological research paper, we encounter an intrinsically higher level of task complexity in the life sciences (see, e.g. the discussion of special phenomena in the biological literature by Leser and Hakenberg [20]). In terms of diversity, biological TM has mainly focused on genes and proteins and their various interactions (mainly PPIs), whereas medical TM has dealt with disease–drug relations (as well as many other relations not relevant to this article). Pharmacological TM has almost exclusively dealt with the recognition of drugs and their properties (dosage, pharmacokinetics, etc.), and only recently moved towards locating drug–drug interactions (DDIs) in documents. The combined exploration of all of these efforts, e.g. looking at drug–protein, gene–disease, or even more complex gene–disease–drug relations, is new and opens up an exciting opportunity for TM research and applications—the unifying topics of pharmacogenomic TM.

Being aware of this challenge, the authors of this survey, together with associated colleagues, created a forum for this kind of research at the Pacific Symposium on Biocomputing (PSB; http://psb.stanford.edu/index.html) in 2010, where the first workshop on *'Genotype-Phenotype-Drug Relationship Extraction from Text'* took place (http://psb.stanford.edu/psb10/gpdrxn-workshop2.pdf), followed by the PSB 2011 workshop on *'Mining the Pharmacogenomics Literature'* (http://psb.stanford.edu/psb-online/proceedings/psb11/wkshop-pharma.pdf). At PSB 2012, the growing importance of this area was recognized by the conference organizers, who assigned a special conference track to this topic, *'Text and Knowledge Mining for Pharmacogenomics: Genotype-Phenotype-Drug Relationships'* (http://psb.stanford.edu/psb-online/proceedings/psb12/intro-textmining.pdf). Following up on these meetings, this survey captures relevant work that has already been done in this exciting new field, tries to identify missing links and research desiderata and concludes with some challenges for future research in this area.

MINING THE PHARMACOGENOMICS LITERATURE

In this section, we will review the most recent results on automatically mining the pharmacogenomics literature. A distinction will be made in Sections 'Genotype mining' through 'Mining pharmacological information' between the identification of relevant named entities and the identification of relations linking them at the genotype level (mainly covering genes, genomic variations, proteins and protein mutations), the phenotype level (mainly covering diseases and other pathological phenomena, as well as treatment efforts, i.e. therapies, in the context of the biomedical literature of interest to this survey, or animal models of these diseases and pathological phenomena) and the pharmacological level (drugs and other sorts of treatment-related chemicals). In Sections 'Genotype–phenotype mining' through 'Genotype–phenotype–drug mining', we will then discuss increasingly complex relation patterns that cross the borders of genotype, phenotype and pharmacological IE as they are investigated in the literature. In Section 'Knowledge discovery: mining implicit and novel information', we will then consider the efforts that view TM as a knowledge discovery task, with two different strands of methodologies, while Section 'Summarizing remarks on text mining approaches' will wrap up the entire discussion.

Besides discussing methodological aspects of these research efforts, we also attempt to contextualize the quality of the results achieved up until now by presenting the outcomes of various evaluation experiments. Very often, the 'F-measure' will be mentioned, one of the most prominent and widely used performance metrics for TM systems. It is defined (see, e.g. [21]) as the harmonic mean of precision (the proportion of true positive outputs to total system outputs, or $TP/TP + FP$, where TP denotes true positives and FP stands for false positives; this measure comes close to specificity and is most similar to positive predictive value) and recall (the proportion of correct system outputs to total correct answers in the gold standard, or $TP/TP + FN$, where FN denotes false negatives; also called sensitivity). All values are here normalized to the interval [0,100] where the value '100' stands for the ideal, best conceivable performance, whereas '0' stands for the worst conceivable performance. Note that we face a kind of 'natural law' in such evaluations, where gains in precision come almost inevitably with losses in recall, and vice versa [22]. The *F*-measure thus provides a reasonable compromise between both evaluation dimensions.

Genotype mining

The crucial named entity types at the level of genotypes are genes and proteins (whose correct distinction in lots of papers is, interestingly enough, hard to make even for biological experts [23], so that this class commonly occurs in disjunctive combination). Usually, a division is made between the task of recognizing gene/protein name mentions in text and, even more ambitious, normalizing these mentions in the sense that database identifiers corresponding to these literal mentions are identified. The latter approach provides valuable links to authoritative gene/protein databases such as EntrezGene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) or UniProt (http://www.uniprot.org/), which allow researchers to easily switch between literature- and database-oriented retrieval.

The current landmark performance for these entities has been established in various iterations of *BioCreAtIvE* competitions which, among other classification tasks, have dealt primarily with recognizing

and normalizing gene/protein mentions, as well as interactions between proteins (PPIs) for abstracts, as well as for full texts (see the surveys of BioCreAtIvE I [24], BioCreAtIvE II [25], BioCreAtIvE II.5 [26], and BioCreAtIvE III [27]) (http://www.biocreative.org/events/biocreative-iii/). No steady increase of performance for the normalization task can be observed due to the fact that substantial parameters of the task were changed from year to year, most notably changing the organisms, moving from abstracts to full texts, providing species a priori or not, and changing performance metrics. However, gene mention and normalization peak, using combined systems, on *F*-scores in the low 90 s [the best single system achieved, on abstracts (BioCreAtIvE II), an *F*-score in the high 80s for gene recognition (87), whereas the highest *F*-score for gene normalization was in the low 80s (81)]. The top-scoring GNAT system for gene normalization in BioCreAtIvE II, in a follow-up study, achieves an *F*-score of 86.4 [28], a result that could exactly be replicated by the GeNo system [29] and seems to constitute the current upper ceiling for this task, not considering system ensembles, which usually perform better than any single system alone.

Although BioCreAtIvE was also concerned with PPI extraction, the current benchmarks for this task have been determined within two consecutive rounds of the BioNLP Shared Task on Event Extraction [18, 19]. A particular to greater specificity of PPIs has been made here—rather than BioCreAtIvE's focus on coarse-grained PPI, the BioNLP Event Extraction competition came up with a set of highly specific PPIs, such as Binding, Phosphorylation, Transcription, ProteinCatabolism, PositiveRegulation and NegativeRegulation, etc. [18]. In 2009, the best system achieved 52.0 *F*-score [30], but was outperformed in the BioNLP 2011 Shared Task with 56.0 *F*-score [31]. The 2011 Shared Task also dealt with an interesting subtask, namely the biomolecular mechanisms of infectious diseases [32], where the over-all winner system of the BioNLP 2011 Shared Task performed best as well with 55.6 *F*-score. This task was construed as an application and extension of the BioNLP 2009 Shared Task event extraction approach to 30 full papers on infectious diseases that represented biomolecular events relating to transcription factors in human blood cells, and its adaptation to a domain that centrally concerns both bacteria and their hosts. That task involves a variety of novel aspects, such as events concerning whole organisms, the chemical environment of bacteria, prokaryote-specific concepts (e.g. regulons as elements of gene expression), as well as the effects of biomolecules on large-scale processes involving hosts such as virulence.

Besides these fundamental named entities and PPI-style relations, pharmacogenomics needs to deal with genetic variants and protein mutations as well. Although sequence data about genetic variation is found at databases such as dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), clues about the functional and phenotypic consequences of the gene variations are generally found in the biomedical literature. In order to find citations of allelic variants of genes in biomedical texts, Furlong *et al*. [33] developed an extension of OSIRIS (http://ibi.imim.es/osirisform.html) which incorporates a specialized named entity recognition module and is built on top of a local mirror of the MEDLINE collection and HgenetInfoDB (http://ibi.imim.es/Hgenesform.html), a database that collects data on human gene sequence variations. The entity recognition module is based on a pattern-based search algorithm for the identification of variation terms in the texts and their mapping (i.e. normalization) to dbSNP identifiers. The performance of OSIRISv1.2 was evaluated on a manually annotated corpus, resulting in an *F*-score of 89 (precision: 99, recall: 82). These data almost perfectly match results achieved with the pattern-based MutationFinder system [34], with recall (81.9) and precision (98.4) figures yielding an *F*-score of 89.4 for variant recognition.

As to the automatic detection and extraction of mutation impacts, Yeniterzi and Sezerman [35] developed the EnzyMiner system with the aim of the automatic classification of MEDLINE abstracts based on the impact of a protein level mutation on the stability and the activity of a given enzyme. EnzyMiner extracts the mutations and disambiguates the cell line names and strain names from mutations. Using a document classifier, the abstracts containing mutations without any impacts are removed and the remaining abstracts are classified into two groups of disease-related and nondisease-related documents, after which extracted mutations are listed for each group. In the case of the nondisease-related abstracts, the documents are further classified into two groups: Documents containing impacts on stability; and documents containing impacts on functionality. Accuracy rates (accuracy is defined by the proportion of correct versus all classification decisions,

(TP + TN)/(TP + TN + FP + FN), using the short cuts from the definition of precision and recall from above) ranging from 93.3 (for mutation extraction) to 85 (for stability/catalytic classification) are reported. Laurila *et al.* [36] continue on this work and present a rule-based approach for the extraction of mutation impacts on protein properties, categorizing their directionality (positive, negative or neutral) and grounding these entities to their respective UniProtKB IDs and selected protein properties, namely protein functions to concepts found in the Gene Ontology (GO; http://www.geneontology.org/). TM is performed within the GATE software framework [37] (http://gate.ac.uk/), a mutation gazetteer list builds on the MutationFinder [34]. The extracted entities are populated to a formalized OWL-DL (http://www.w3.org/2004/OWL/) Mutation Impact ontology, thus not only allowing sophisticated access to mutation impacts using the SPARQL (http://www.w3.org/TR/rdf-sparql-query/) query language, but also facilitating the deployment of novel semantic web services based on the *Semantic Automated Discovery and Integration* (SADI) framework (http://sadiframework.org).

Phenotype mining

Phenotype is the set of observable (molecular or gross) characteristics of an individual resulting from the interaction of its genotype with the environment. Special emphasis is put on pathological phenomena—diseases in particular, as well as their anatomical sites, conditions and treatment. Because of the availability of a large variety of authoritative disease terminologies (most important are the Medical Subject Headings (MeSH; http://www.ncbi.nlm.nih.gov/mesh), the Unified Medical Language System (UMLS; http://www.nlm.nih.gov/research/umls/), the International Classification of Diseases (ICD-10; http://www.who.int/classifications/icd/en/), the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT; http://www.ihtsdo.org/snomed-ct/), and the Medical Dictionary for Regulatory Activities (MedDRA; http://www.meddramsso.com/); see also Section 'Terminologies and ontologies' below), dictionary-based methods are very much favored for this task because term variability is lower for standardized phenotype terms (for instance, disease names) than it is for genotype terms. However, the dictionary-based methods are susceptible to definite weaknesses, because nonstandardized phenotype descriptions employ complex and highly variable linguistic utterances (spanning long phrases, even entire sentences) and are therefore, hard to locate in texts.

Jimeno *et al.* [38], e.g. report the highest recognition accuracy for dictionary look-up (68.4 *F*-score) in comparison with a statistical, information theory-inspired approach (66.6) and MetaMap (65.4), (http://metamap.nlm.nih.gov/), a program that maps text mentions to the UMLS Meta Thesaurus (http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html) or finds MetaThesaurus concepts in text [39]. The reported *F*-scores can further be boosted (up to 83.0) using a parameterized voting schema that takes the results of all three different approaches into account.

Whereas Jimeno's focus is on PubMed abstracts, another crucial text genre for phenotype recognition is the clinical report. Such documents present entirely new challenges—for instance, de-identification of patient names (see the survey by Meystre *et al.* [40]), the presence of large amounts of fragmented and ungrammatical utterances, including misspellings, an abundance of short forms (such as acronyms or other types of abbreviations), and frequent spelling variation (see the survey by Meystre *et al.* [5], for approaches dealing with these matters). Kipper-Schuler *et al.* [41] present an evaluation of the dictionary look-up component of Mayo Clinic's Information Extraction system cTAKES [42]. It was tested on a corpus of 160 free-text clinical notes that were manually annotated with the named entity type 'Disease'. The dictionary used for this evaluation was a subset of SNOMED-CT, with semantic types corresponding to diseases/disorders without any augmentation. The recognizer achieves an *F*-score of 56 for exact matches and *F*-scores of 76 and 62 for including right- and left-partial matches, respectively. The over-all *F*-score for all match types peaks at 81. Using the same corpus, machine learning (ML) methodology has also been tested in a comparison between conditional random fields (CRFs) and support vector machines (SVMs) by Li *et al.* [43], yielding 86 and 64 *F*-scores, respectively, with a dictionary look-up baseline of 60 (using SNMOD-CT). There is a clear advantage using specific ML approaches (here, CRFs) rather than dictionary-based approaches alone. Based on UMLS as a terminological source, Roberts *et al.* [44] describe an SVM recognizer for five named entity types (Condition, Drug or Device,

Intervention, Investigation, and Locus). This entity set is much more diverse and closer to clinical reality. Using a corpus of 77 clinical documents, an average *F*-score of 70.7 (only 3% points below human inter-annotator agreement, discussed below) is reported for all five entity types (see also Section 'Phenotype-focused corpora' where we discuss the underlying CLEF initiative in more depth). This result shows that conceptually rich clinical named entity recognition is intrinsically hard—not only for machines, but for human experts as well.

A fundamentally different approach is chosen for BANNER [45]. This is an open-source biomedical named entity recognition system based on advanced ML technology (CRFs), intended to serve as a benchmark for the field. It is not restricted to a special entity recognition problem (such as capturing disease names) by maximizing on domain independence. Evaluation on different corpora for gene mention and disease/treatment recognition show comparatively favorable results in relation to alternative off-the-shelf software, but also reveal a marked decrease in performance, relative to recognizers dedicated to specific entity types, for gene mention recognition [in the high 80 s (82.0) on *F*-score] on the BioCreAtIvE II Gene Mention corpus, and disease/treatment recognition [in the mid 50 s (54.8) on *F*-score] on the BioText corpus ([46]; see also Section 'Phenotype–drug mining').

Mining pharmacological information

TM with focus on pharmacological information tries to identify drugs and other chemicals that are functionally important in treating or causing medically significant phenotypes in the course of treatments, therapies, etc. As with phenotypical entities, pharmacological entities are well represented in a variety of lists and nomenclatures that can serve as a primary resource for identification programs. Segura-Bedmar *et al.* [47] report in a ground-breaking study 78.0 precision and outstanding 99.3 recall for their DrugNER system, which combines information obtained from the UMLS, the 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. Their system is also capable of detecting possible candidates for drug names that have not been recognized by the MMTx program by applying these rules.

Kolárik *et al.* [48] start from a dictionary-based approach as well, exploiting DrugBank (http://drugbank.ca/), a resource that combines more than 13 K entries for detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure and pathway) information, and then identify drug effect expressions with the help of an extended rule system based on the so-called Hearst patterns [49]. They describe quite stable phrasal and lexical patterns that are indicative of relevant information and thus enumerate crucial linguistic structures explicitly. The system performs with 89 *F*-score on DrugBank texts, whereas on less formatted MEDLINE abstracts, an 83 *F*-score is achieved. Interestingly, 29–53% of the terms extracted from MEDLINE are new valid drug property terms—information that can then be fed back to further enhance resources like the DrugBank.

Drugs (as pharmaceutical products) are special types of chemical substances with high relevance for biomedical research. It might thus be worthwhile to look at chemical substances that have not yet found their way into the canonical vocabulary of pharmacogenomics, or are discussed on the basis of purely chemical considerations, without any immediate link to biology, pharmacology or even medicine. Since the number of chemical compounds cannot be enumerated explicitly, static dictionary-based methods are less likely to cover all required data and must be complemented or even replaced by 'generative' methods, such as rule systems or ML classifiers.

Klinger *et al.* [50] report on such experiments targeting the recognition of chemical names proper whose mentions follow the International Union of Pure and Applied Chemistry (IUPAC) rules (expressions such as *"7-ethyl-10[4-(1-piperidino)-1-piperidino]carbonyloxycamptothecin"*. The IUPAC is the body behind these standardization efforts unmatched for the biology or medicine domains; http://www.chem.qmul.ac.uk/iupac/). For the MEDLINE section of the Fraunhofer corpus (see Section 'Drug- and chemicals-focused corpora' below), they achieve an *F*-score of 85.6 by combining dictionaries and ML approaches (CRFs). For harder-to-process patents, the *F*-score drops to 81.5. Also using ML techniques, Corbett and Copestake [51] developed a system to use character-based *n*-grams, Maximum Entropy Markov Models, and re-scoring procedures to recognize chemical names, and to make confidence estimates for the extracted entities.

This system is integrated in the Oscar-3 chemical names recognizer [52] (https://sourceforge.net/projects/oscar3-chem). The Corbett-Copestake system is more general than Klinger's in that its focus is not only on chemical names that adhere to IUPAC conventions, but also on those which deviate from this norm. It is also more flexible in that an adjustable threshold allows the system either to be tuned to high precision or high recall. At a threshold set for balanced precision and recall, the Corbett-Copestake system extracts chemical named entities at an *F*-score of 80.7 from chemistry full text papers (see also Section 'Drug- and chemicals-focused corpora') and 83.2 from chemistry MEDLINE abstracts. This pattern, drops in performance of TM software running on full text documents compared with the processing of abstracts by the same program, has been observed with stunning systematicity (see also [53]) and can be explained by the increase of linguistic complexity when moving from abstracts to full texts (see also Section 'Conclusions and outlook'). Recently, a linguistically much more informed extension of OSCAR, the ChemicalTagger, has been introduced [54]. It incorporates part-of-speech and phrasal information using a standard rule-based grammar and not only recognizes chemical names, but also relations holding among them (such as Adding, Dissolving, Cooling, Purifying, Evaporating, etc.). ChemicalTagger has been deployed for over 10 K patents and has identified solvents from their linguistic context with precision peaking at 99.5.

Regarding the creation of a large dictionary of chemical names, Hettne *et al.* [55] describe a rule-based method (primarily for term filtering and disambiguation) that helps to identify names of drugs and small molecules, including metabolites and endogenous molecules, by incorporating a broad collection of other dictionaries, such as the UMLS, MeSH, Chemical Entities of Biological Interest (ChEBI; http://www.ebi.ac.uk/chebi/), DrugBank, KEGG (http://www.genome.jp/kegg/), HMDB (http://www.hmdb.ca/) and ChemIDplus (http://chem.sis.nlm.nih.gov/chemidplus/). They report an overall performance of the combined dictionary on the Fraunhofer corpus (see Section 'Drug- and chemicals-focused corpora'below) of 67 precision and of 40 recall (80 for trivial names), which seems modest in comparison with the approaches described above using ML techniques for recognition. Still, the combined dictionary performed better than the dictionary incorporated in the chemical names recognizer OSCAR3.

Medication information is one of the most important types of clinical data in electronic medical records (EMRs). As medication data is often expressed in verbal prose in clinical notes rather than structured coded data, they pose a particular challenge for TM. There are a couple of NLP systems around that automatically extract medication information from clinical notes. For instance, Levin *et al.* [56] used lists of abbreviations and open source software for the recognition of spelling variations and normalized drug information from free-text fields in EMRs with the help of RxNorm (http://www.nlm.nih.gov/research/umls/rxnorm/), a cross-referenced lexicon of clinical drug nomenclature that was used as the reference source for drug name verification and for mapping trade (proprietary) names to their generic equivalents. They achieved 92.2 sensitivity (recall) and 95.7 specificity (precision) for the validation set (14 655 cases), respectively.

MedEx [57], a semantic rule-based system, was evaluated on a data set of 50 discharge summaries and showed it performed well on identifying not only drug names (*F*-score 93.2), but also signature information, such as strength, route and frequency, with *F*-measures of 94.5, 93.9 and 96.0, respectively. Jagannathan *et al.* [58] compared four commercial NLP engines on this task, which lag behind these figures; they also report an identical *F*-score of 93.2 for capturing drug names, yet significantly lower *F*-scores of 85.3, 80.3 and 48.3 for retrieving strength, route and frequency, respectively. When MedEx was applied unchanged to outpatient clinic visit notes, similar *F*-scores over 90 on a set of 25 clinic visit notes were found.

As part of the 2009 i2b2 Challenge (the task required accurate recognition of medication name, dosage, mode, frequency, duration and reason for drug administration; see [59] and Section 'Phenotype-focused Corpora'), a slightly extended version of MedEx [60] achieved an overall *F*-score of 82.1 (second rank out of 20 participating teams). The best system in this competition, a hybrid incorporating ML (CRFs for named entity recognition, SVM for medication-specific relation extraction) and rule-based technology, peaked at 85.7 *F*-score [61]. In a recent follow-up study, Halgrim *et al.* [62] make use of that i2b2 Challenge corpus as well, which in their study contains the original 696 discharge summaries from which roughly 400 were gold-annotated

in the meantime. In their hybrid system, a first pass involves a cascade of statistical maximum entropy classifiers (incorporating the FDA's National Drug Code Directory list) to identify crucial medication-related named entities (name of medication/drug, dosage, mode, frequency, duration and reason), whereas a second pass uses simple heuristics to link those isolated entities into medication events (e.g. the dosage of a specific drug). While for 'name of medication/drug' an *F*-score of 89.8, and for, 'dosage', 'frequency' and 'mode' *F*-scores from 93.1 to 93.3 were determined, 'duration' and 'reason' yield bad performance figures on the test set with 51.5 and 47.1, respectively, with an over-all *F*-score of 86.9, whereas for linkage the system achieved an *F*-score of 84.1.

Interactions among drugs are as important for clinical research as interactions among proteins are for molecular biology. This is due to the fact that multiple drug prescriptions are the norm for patients rather than the exception. The various interactions among drugs in multi-drug therapy have fuelled research on DDIs. Early work on DDIs was reported by Mille *et al.* [63]. More recently, a special DDI Extraction Challenge (http://labda.inf.uc3m.es/DDIExtraction2011/) was established to create a benchmark data set (DrugDDI, see also Section 'Drug- and chemicals-focused corpora') and evaluation task that will enable researchers to compare their algorithms when applied to the extraction of DDIs from textual descriptions in DrugBank (see [64] for the challenge description). The best result achieved in the first shared task, in which drug names were already identified, is reported by Thomas *et al.* [65], who use state-of-the-art dependency parsing (a type of in-depth analysis of the syntactic structure of a sentence) and combine this type of linguistic information in an ensemble-based ML approach where the best single classifier achieves 63.4 *F*-score and the best ensemble yields 65.7 on the test set. These results differ from the best PPI extraction results by almost +10% (note that in the Shared Task on Event Extraction [18] gene/protein names were also prespecified).

Previously, Segura-Bedmar *et al.* [66] reported on experiments that combined shallow parsing (a type of syntactic analysis which finds word groups within the sentence, but does not provide a full syntactic analysis of the entire sentence) and simplification of complex syntactic structures using pattern matching. This system under-performed by all measures, with a precision of 48.7, a recall of 25.7 and an *F*-score of only 33.6, whereas an ML-based SVM approach peaked at a considerably higher *F*-score of 66.0 (55.1 precision, 82.3 recall) on DrugDDI [67]. In another system configuration, Segura-Bedmar *et al.* [68] employ a supervised, linguistically shallow kernel-based technique, with which 51.0 precision, 72.8 recall and an *F*-score of 60.0 were achieved on DrugDDI. It appears that an *F*-score of 66 defines the current landmark result for the DDI extraction task.

Tari *et al.* [69] propose an entirely different methodological approach we have not discussed before in this article. They integrate biological domain knowledge with biological facts that are extracted from applying TM to MEDLINE abstracts and other curated sources (such as UniProt and GO) to automatically derive enzyme-based DDIs on the basis of automated reasoning. The authors thus distinguish between explicit extraction of DDIs (basically, information that can be recognized in their system by dependency-based parsing and querying a corresponding parse tree database) and implicit extraction, the latter requiring logical inferences based on various properties of drug metabolism, in their system by using an AnsProlog-based reasoning engine (Prolog is a logic programming language). With a DrugBank-elicited gold standard of 494 DDIs, 77.7 precision is reported for explicit and 81.3 precision for implicit extraction, thus revealing the added value of inference-derived biomedical knowledge.

Also capitalizing on the prospects of combining TM with formal reasoning, Percha *et al.* [70] describe an ML-based approach (using random forests), which builds on the previous work by Coulet *et al.* [71] in extracting and normalizing gene–drug relationships from MEDLINE abstracts, to infer DDIs possibly interacting via a common genetic pathway. The classifier recognizes the combinations of relationships, drugs and genes that are most associated with the gold standard DDIs, correctly identifying 79.8% of assertions relating interacting drug pairs. The methodology enables the creation of novel predictions of interacting drug-pairs, while maintaining links back to the original evidence and sentences supporting these predictions.

Inferring novel facts related to drug metabolism and DDIs that are not explicitly mentioned in text but have to be inferred by means of formal reasoning is certainly an area of ground-breaking research; we will discuss it further in Section 'Knowledge

discovery: mining implicit and novel information'. In Sections 'Genotype–phenotype mining' through 'Genotype-phenotype-drug mining', we will deal with several entity type-mixed relationships that highlight research with a particular focus on pharmacogenomics.

Genotype–phenotype mining

The first set of mixed relations we consider deals with the role that genotype data (i.e. genes and proteins, genetic variations, etc.) plays for phenotypic phenomena (among them diseases and pathological phenomena, but also nonpathological processes such as aging). Krallinger *et al.* [3] surveyed the recent approaches and systems combing gene-centric and disease-centric TM, with focus on the molecular oncology domain. Most of the literature deals with disease-type phenotypic phenomena. Since the types of semantic relations holding between these entities are less clear than in the previous sections, unsupervised ML methods, clustering algorithms in particular and similarity scores between data sets increasingly play a role here. Sophisticated NLP methods for relation extraction are mostly lacking because usually the lexicalization of the target relations remains unclear or is even unknown; rather, the computation of association strengths between terms (entities) prevails. Also, hybrid use of resources (databases, terminologies and text collections) is a crucial issue.

As an example of sophisticated resource-juggling, we consider the work of Butte and Kohane [72], who try to identify genotype–phenotype and environmental/experimental condition-genotype relations by using publically available sample annotations from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/), a repository for microarray, next-generation sequencing and other forms of high-throughput functional genomic data, as well as vocabulary from the UMLS. The annotations covering phenotype data and environmental/experimental context data are processed using MetaMap, leading to a mapping to UMLS concepts. In order to extract genotype data, GEO identifiers are manually linked to LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink) identifiers, allowing nearly 55 M expression measurements to be referenced. Gene expression data is then hierarchically clustered by phenotypical, environmental and experimental context (based on the UMLS mappings). A ranking of gene expression measurements is computed and those measurements that show significant differential expression form part of the network of relations. This procedure has identified novel genes related to concepts such as aging, among others.

As a second example from this stream of work, van Driel *et al.* [73] classify over 5000 human phenotypes contained in the Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/omim) database [using the MeSH hierarchies for anatomy (A) and disease (C) sections for extraction from OMIM] and, using MeSH concepts as features, find that vector-based similarity between phenotype records reflects biological knowledge of interacting functionally related genes. These similarities are positively correlated with a number of measures of gene function, including relatedness at the level of protein sequence, protein motifs, functional annotation and direct PPI. Since phenotype grouping reflects the modular nature of human disease genetics, phenotype mapping may be used to predict candidate genes for diseases, as well as functional relations between genes and proteins.

As a third example, Gonzalez *et al.* [74] combine gene–disease relationships extracted from MEDLINE abstracts and further augmented by consulting the vocabulary services from resources such as the HUGO Gene Nomenclature database (http://www.genenames.org/), using the IE system IntEX [75], with protein interaction networks extracted from curated databases, such as the Biomolecular Interaction Database (BIND; http://bond.unleashedinformatics.com/Action?), the Molecular Interaction Database (MINT; http://mint.bio.uniroma2.it/mint/Welcome.do), or the Database of Interacting Proteins (DIP; http://dip.doe-mbi.ucla.edu/dip/Main.cgi). The augmented list of genes and gene products is then ranked, combining a score that reflects the strength of and confidence in the relationship with the initial set of genes and another score that reflects the importance of the gene in maintaining the connectivity of the network. This scoring is used to predict the proteins most likely to be related to the disease under scrutiny. Top-ranked proteins are related to the evaluation example, atherosclerosis, with accuracy between 85–100 for the top 20 and 64–80 for the top 90, if duplicates are ignored. Similar resource-heavy work with emphasis on using OMIM and other databases has been reported by Chen *et al.* [76].

As a last example, Bundschus *et al.* [77] apply CRFs to extract relations between genes and diseases

from GeneRIF (Gene Reference Into Function) concise functional description phrases (available from the EntrezGene database), where five relation types (among them, AlteredExpression, Genetic Variation, RegulatoryModification, etc.) were considered. In order to demonstrate the scalability of their approach, the whole GeneRIF database is processed, resulting in a gene–diseases network that contains 34 758 semantic associations between 4939 genes and 1745 diseases. Disease recognition peaks at an *F*-score of 78.0 (the gene entity is identical to the EntrezGene ID, and thus trivial to get), whereas relation recognition ranges between 80.0 and 78.9 *F*-score for GeneticVariation and AlteredExpression, RegulatoryModification drops to 71.2 *F*-score (over-all *F*-score is 78.0).

Tiffin *et al*. [78] stress the pivotal role of ontologies by using the eVOC Anatomical System ontology (http://www.obofoundry.org/cgi-bin/detail.cgi?id=evoc) as a bridging vocabulary that integrates clinical and molecular data through a combination of text and data mining. They then select candidate disease genes according to their expression profiles within tissues (taken from the Ensembl database; http://www.ensembl.org/index.html) affected by the disease of interest. First, the association between each eVOC anatomy term and disease name is found by co-occurrence in MEDLINE abstracts. Then the identified anatomy terms are ranked and candidate genes annotated with the top-ranking terms are selected. The system succeeds in selecting the correct disease gene amongst other candidate genes in 15 out of 17 diseases in the training data set (88.2% success rate). Whereas the previous study considers ontologies only as a terminological resource, Coulet *et al*. [79] use OWL-based ontologies by exploiting their formal specification, relying on subsumption, properties and class descriptions, for the purpose of knowledge discovery of genotype–phenotype relationships.

More sophisticated NLP methods are used by Chun *et al*. [80]. They describe a system for gene–disease relation extraction that is based on the co-occurrence of gene and disease name mentions (found via dictionary look-up) and additional filtering of false positives with a Maximum Entropy-based named entity classifier accounting for gene and disease name entities. In the filtering mode, the system achieves 78.5 precision and 87.1 recall on a manually annotated corpus with 1000 co-occurrences of gene and disease names. Taking exploitation of linguistics further, Masseroli *et al*. [81] build on the output of SemGen [82], a system that extracts semantic predications about the etiology of genetic diseases. They apply phrase-based distance heuristics to the argument and its predicate, based on the intuition that arguments which occur close to their predicate are easier to identify than those at a distance, to filter the extracted semantic relations according to their likelihood of being correct. Considering distance criteria (or shortest path counts in parse tree structures, i.e. formal representations of the syntactic structure of a sentence) has turned out to be an important idea for any sort of relation extraction relying on linguistically informed analytics (see, e.g. the top-scoring systems from the Shared Tasks on Event Extraction; Section 'Genotype Mining'). They compare relations extracted in this way to those identified with co-occurrence processing only. Postprocessed SemGen predications are then used to investigate the genetic basis of Parkinson's disease. Two of the genes extracted by postprocessing are likely to be relevant to Parkinson's disease, but were previously not associated with this disease in several important databases of genetic disorders. Finally, an interesting use of an EMR system to conduct genome-wide association studies is reported by Kullo *et al*. [83]. Trying to find genomic indicators for peripheral arterial disease (PAD) they collected demographic data and laboratory values from EMR, medication use and smoking status from clinical notes (using cTAKES [42]), as well as other cardiovascular risk factors and co-morbidities based on, e.g. ICD-9-CM codes, and linked this information to a clinical biorepository of DNA and plasma.

Genotype–drug mining

Much of the current research on the genotype–drug connection considers the potential to read from genetic data how drugs can be effectively tailored to a given genetic context. Under the heading of individualized medicine, personalized drug dosage tuning, adverse effect prediction, etc. are among the primary goals of this research. It is this kind of information that TM systems try to harvest from source documents. Methodologically, many studies start from the co-occurrence of genotype and drug entities in some formal text segment (usually, sentences) and then try to filter out false positives using linguistic or statistical criteria, or other kinds of heuristics.

Chang and Altman's system [84] recognizes such relations between genes and drugs in MEDLINE

abstracts with a co-occurrence-based approach. Subsequently, relations are classified into five categories, as specified by PharmGKB (see Section 'Drug- and chemicals-focused corpora' below), using a Maximum Entropy-based ML approach. The relation recognition step is evaluated against a small data set of 215 gene–drug relations manually extracted from a review article, whereas the classification step is assessed against human-curated articles from PharmGKB. Evaluation results for all five categories range from 88.0 recall with 75.0 precision for predictions of pharmacokinetics, to 9.0 recall with 27.0 precision for the ClinicalOutcome category. The authors concede that the selected PharmGKB data set is really small, including only 325 gene–drug pairs, and that the evaluation results depend heavily on the size of the training data.

More recent systems for the extraction of gene–drug relationships are Pharmspresso and GenDrux. Pharmspresso [85] builds on the well known Textpresso tool [86], a full-text search engine for biological entities and facts such as PPIs. Pharmspresso has been extensively evaluated concerning the detection of gene and drug names. With respect to relationships, it yields only 50.0 recall on gene–drug 'association' relations when evaluated against 45 full-text articles that contain 178 gene name instances and 142 drug name instances. GenDrux [87] is a Web-based retrieval tool whose document collection consists of 4 K MEDLINE abstracts collected using gene and drug name filters. There is a focus on gene and drug names related to breast cancer, while relation extraction is based only on the co-occurrence of relevant terms in the titles of documents. Yet, no evaluation is provided for the system. The first large-scale evaluation study in this field was carried out by Coulet *et al.* [71], who extracted PharmGKB-relevant relationships from 17 M MEDLINE abstracts. The extracted relationships are reported to have a precision up to 87.7. However, this work does not evaluate the recall of the relationship extraction approach.

In a comparatively large-scale experiment, yet with an entirely different approach which resembles the resource-juggling activities reported in Section 'Genotype–phenotype mining', Kuhn *et al.* [88] developed a Search Tool for InteracTions of CHemicals (STITCH) which integrates information for over 68 K chemicals, including 2200 drugs, and 1.5 M genes. They mine both MEDLINE and OMIM for term co-occurrence and then apply NLP-based relation extraction machinery [89]. The resulting relations (as well as metrics accounting for chemical structure similarity) are used as evidence to predict relationships between chemicals, drugs and genes in particular.

A particularly innovative combination of resources is reported by Xu *et al.* [90]. They linked the DNA data bank at Vanderbilt University Hospital, which contains over 100 K DNA samples to de-identified EMRs from their hospital in order to identify associations between genetic variations and drug efficacy and toxicity. In manual experiments they had already investigated associations between steady-state Warfarin weekly dose and variants in *VKORC1* and *CYP2C9* in the biobank. Since these experiments were overly time-consuming, they develop an automated weekly dose calculation system based on an existing medication-IE system, MedEx [57], and applied it to data sets from the aforementioned Warfarin pharmacogenetic study. Using automatically extracted Warfarin weekly doses, they achieved similar *P*-values for genetic associations to those from manual data extraction, indicating that such EMR-based pharmacogenetic studies could be done in an *in silico* fashion.

An issue often raised is whether TM technology can compete in qualitative terms with human efforts on the same text sources—quantity-wise, computers have an undisputable advantage already. Using a gene–drug network automatically created from sentence-level co-occurrence data in the full text of scientific articles (1731 documents taken from PharmGKB), Garten *et al.* [91] compare the performance with that of a network created by manual curation of those articles. Under a wide range of conditions, they show that a knowledge base derived from text-mining the literature (using a combination of Pharmspresso [85] and PGxPipeline [92]) performs, as well as and sometimes even better than, a high-quality, manually curated knowledge base. The authors conclude that the use of relationships mined automatically from the literature as a knowledge base for pharmacogenomics is both reasonable and empirically justified. Additionally, their system can accurately extrapolate new relationships with 77.4 precision.

Phenotype–drug mining

Another major challenge for pharmacogenomic TM is concerned with determining in which ways drugs

affect certain phenotypic states—in particular, what the effects of (particular dosages of) drugs are on patients, what side or even adverse effects might occur, etc. These issues move the field from the biomolecular arena proper to clinical settings. From a methodological perspective, NLP and ML methods for relation extraction from MEDLINE or clinical document sets, EMRs in particular [93], prevail in this area. (see Warrer *et al*. [94] for a comparative survey of TM technology for EMRs currently in use.) Recently, the integration of disparate resources, including terminologies, databases and clinical document collections, to automatically generate an executable and publically available drug-indication knowledge base has become a major concern [95].

Early and ground-breaking work was performed in the context of the BioText project (http://biotext.berkeley.edu/ The corresponding corpus is available at http://biotext.berkeley.edu/data/dis_treat_data.html) [46], where a 7-way classification of disease–treatment relations was developed (treatments include the application of drugs, and relations incorporate Treatment-Cures-Disease, Treatment-Causes-Disease, Treatment-Prevents-Disease, etc.). Based on suitable MeSH filtering of MEDLINE documents, generative (maximum likelihood-based) and discriminative (neural network-based) ML models are used for relation extraction. The best generative classifier achieves 74.9 accuracy, while the neural network classifier does much better (79.6).

In a follow-up study using the same MEDLINE corpus, Bundschus *et al*. [77] apply CRFs to extract relations between diseases and treatments, where seven relation types (such as Cures, DoesNotCure, Prevents, SideEffect, Vague) are considered. They get an *F*-score of 72.0 for disease/treatment recognition (Disease recognition yields an *F*-score of 77.2, while Treatments have a considerably lower *F*-score of 64.6), and achieve 96.9 accuracy for relation extraction when the entities are known, whereas 79.5 accuracy (almost identical with the results for neural networks from the study discussed above) were attained when the entities are unknown.

Based on the set of 826 patient records (349 documents for training, 477 for testing) from the 2010 i2b2 challenge [96] (see also Section 'Phenotype-focused corpora'), Doğan *et al*. [97] consider the extraction of three major entity types, namely MedicalProblem, Treatment (including procedures and medications) and Test (lab procedures and measurements prescribed to a patient), linked by eight relationship types such as Treatment-Improves/Worsens/Causes-MedicalProblem, or MedicalProblem-Reveals/Conducted-Test. They combine the outcome of a statistical model for concept recognition (based on CRFs) with a linear SVM approach for relation extraction and achieve for the named entity task, for partial span match, *F*-scores of 91.2, 96.0 and 95.4 for Problem, Treatment, and Test, respectively (for exact span, the *F*-scores are 81.0, 90.9 and 91.7, respectively), with an over-all *F*-score of 93.9 (87.0). In their end-to-end system (i.e. starting with the automatic recognition of named entities and then predicting possible relationships between two entities found in the same sentence), they end up for relationship extraction using automatically extracted concepts with an *F*-score of 70.4, which is comparable to that obtained using manually annotated concepts (*F*-score 71.1), the difference not being statistically significant.

Chen *et al*. [98] report on the comparative evaluation of two IE systems, namely MedLEE [99, 100] and BioMedLEE [101], on 81 828 randomized controlled trial (RCT) articles from PubMed and 48 360 hospital discharge summaries. Disease and drug entities are identified using the two NLP systems, in addition to MeSH annotations for the PubMed articles. Focusing on eight diseases, co-occurrence statistics are computed and evaluated concerning the strength of the association between each disease and relevant drugs. Ranked lists of disease–drug pairs are generated, and cut-offs are calculated for identifying stronger associations among these pairs for further analysis.

A very specific, yet clinically highly relevant theme addressed in this area are unwarranted side effects (or adverse drug reactions, ADRs) of prescribed drugs, also referred to as pharmacovigilance. Gysbers *et al*. [102] optimized the Cancer Text Information Extraction System (CaTIES; https://cabig.nci.nih.gov/community/tools/caties) [103] for the identification of terms suggestive of ADRs in text. Wang *et al*. [104, 105] applied the MedLEE system [99] for the identification of side effect profiles from discharge summaries for selected drugs/drug classes on the market. Based on 437 (Japanese) discharge summaries, Aramaki *et al*. [106] achieve 59 recall and 30 precision for ADR extraction. Leaman *et al*. [107] developed a framework for extracting relationships between drugs and adverse effects from user posts in health-related social networks. The most recent work by Gurulingappa

et al. [108] in this context deals with the classification of sentences from clinical records that indicate ADRs in patients, e.g. due to dosage errors. The authors use a hybrid approach combining dictionary-based information with a Maximum Entropy classifier, which achieves 77.0 *F*-score.

In an attempt to extract such side effect information from the literature, as well as from labels of FDA-approved drugs without using heavy NLP machinery, Kuhn *et al*. [109] developed the SIDER resource. The system connects drugs to their phenotypic effect by extracting side effects from drug labels.In total, 62 269 drug–side effect pairs covering a total of 888 drugs and 1450 distinct side effects were extracted.

Genotype–phenotype–drug mining

Binding everything together, in this section the focus is on the interrelations between genotypes, phenotypes and drugs. The interrelationships that occur at the phenotype–drug level are traced back to possible genetic traits here. Work in this area is at the heart of pharmacogenomic TM. Typically, studies in this area are hybrid in the sense that not only TM machinery is used, but other types of resources such as interaction network (pathway) databases are also integrated. There is some preliminary evidence that for this kind of research, formal reasoning is particularly helpful to put all of the different knowledge threads together. Furthermore, a transition from explicit IE to discovering implicit knowledge in structured (database) and unstructured (document) resources can be observed in some publications (see also Section 'Knowledge discovery: mining implicit and novel information').

Rindflesch *et al*. [110] carried out a classical study when developing the EDGAR system for the extraction of information about genes, cell types and drugs, as well as gene–drug relations relevant for cancer. EDGAR exploits underspecified syntactic parse trees (similar to the shallow parses discussed above) and applies manually specified syntactico–semantic rules for the extraction of relationships. A background knowledge representation composed of gene–drug–cell relationships is used both to constrain the extraction of explicitly stated relationships (e.g. Drug–Suppress–GeneExpression), as well as the inference of new ones. The EDGAR system was also the first to extract nonbinary relationships between these three entities. However, EDGAR was not systematically evaluated.

Ahlers *et al*. [111] discuss the rule-based Enhanced SemRep system, which extracts a range of gene–disease and drug–disease relations (such as Stimulates, Disrupts, or Causes) from approximately 1 K MEDLINE abstracts on pharmacogenomics. The UMLS Metathesaurus and Semantic Network is used to enforce semantic constraints on the extraction procedure that yields a recall of 55 and a precision of 73.

Roberts *et al*. [112] describe an ML-based system for relation extraction, using SVMs, and trained and tested it on a clinical subcorpus of CLEF (see Section 'Phenotype-focused corpora') dealing with 77 oncology narratives hand-annotated with clinically important relationships. Over a class of seven relation types (among them HasTarget, HasLocation, HasIndication and HasFinding), the system achieves an average *F*-score of 72, only slightly behind an indicative measure of human inter-annotator agreement (75) on the same task.

Whereas this research adheres to textual sources only, Li *et al*. [113] integrate gene/protein and drug connectivity information based on protein interaction networks with literature mining. Taking Alzheimer's Disease (AD) as an example, they first incorporate molecular interaction networks taken from OMIM and OPHID (http://ophid.scholarsportal.info/) to reduce bias and improve the relevance of AD seed proteins. Then MEDLINE abstracts are used to retrieve enriched drug terms that are indirectly associated with AD through molecular studies. Term frequency statistical methods (the *P*-value of each term's significance) are applied that take advantage of term statistical distributions from the entire MEDLINE collection; no further linguistic processing is required here. Finally, a comprehensive AD connectivity map is created by relating enriched drugs and related proteins in the literature. They show that their approach outperforms both curated drug target databases and conventional IR systems. Furthermore, initial explorations of the AD connectivity map yielded new hypotheses regarding candidate drugs for AD treatment.

Linkage of heterogeneous resources is a crucial concern behind the SNPshot system developed by Hakenberg *et al*. [114] (http://bioai4core.fulton.asu.edu/snpshot). It contains information on phenotypic effects of genetic variants, focusing on effects on drug response selected from close to 180 K MEDLINE abstracts. They make available summarized

information linking genes, gene variants, diseases, drug efficacy, ADRs, populations and allele frequencies, with cross-references to the literature, EntrezGene, PharmGKB, DrugBank and dbSNP. SNPshot achieves an impressive performance of 85–92 precision for the recognition of the main entity types (Genes, Drugs and Diseases) and 79–83 for relationships involving these types.

Coulet *et al.* [71] describe an ontology of pharmacogenomic relationships built starting from a lexicon of key pharmacogenomic entities and a syntactic parse of more than 87 M sentences from 17 M MEDLINE abstracts. The syntactic dependency structure of pharmacogenomic statements is used to systematically extract commonly occurring relationships and to map them to a common schema. The extracted relationships have 70.0–87.7 precision and involve not only key pharmacogenomic entities such as Genes, Drugs and Phenotypes (e.g. VKORC1, Warfarin and clotting disorder), but also critical entities that are frequently modified by these key entities (e.g. VKORC1 polymorphism, Warfarin response and clotting disorder treatment). The result of this analysis is a network with clear semantics of 40 K relationships between >200 entity types. A significant innovation in this work is the unbiased extraction of the relationships themselves, between subject and object of the sentences. Whereas previous methods had prespecified relationships of interest (e.g. Bind, Inhibit, Metabolize), Coulet *et al.* extract all relationships commonly occurring in the literature, connecting a drug-related entity to a gene-related one.

Rather than only dealing with sets of single relations lacking further integration, Tari *et al.* [115] go one significant step further and propose a novel approach to automated 'pathway synthesis'. Facts are acquired from hand-curated knowledge bases (such as DrugBank or PharmGKB), as well as through automated extraction from MEDLINE abstracts. The text analytics component contains a syntactic parse tree database, while semantic analytics are provided by MetaMap and the gene/protein normalizer GNAT [28]. A flexible parse tree query language was developed to perform IE at the parse database level. An essential novel aspect of that approach is to apply 'logical reasoning' to the acquired facts based on biological knowledge about pathways. By representing such biological knowledge as clauses, an AnsProlog-based reasoning engine is capable of assigning ordering to the acquired facts and interactions that is crucial for pathway synthesis. As an example, 20 pharmacokinetic pathways were synthesized and evaluated by reconstructing the existing pharmacokinetic pathways available in PharmGKB. The results show that this approach not only is capable of synthesizing these pathways, but also of uncovering information that is not available in the manually annotated pathways, a pharmacologically relevant use case of knowledge discovery.

As an alternative to logical reasoning for knowledge discovery, Frijters *et al.* [116] combine a co-occurrence-based recognition approach with Swanson-style knowledge discovery in their CoPub discovery system to search for new relationships between genes, drugs, pathways and diseases in MEDLINE abstracts. Several of the newly found relationships were validated using independent literature sources. In addition, newly predicted relationships between compounds and cell proliferation were validated and confirmed experimentally in an *in vitro* cell proliferation assay. Also using a Swanson-style approach, Baker and Hemminger [117] infer drug–disease associations by typing B-terms to be proteins only, thus minimizing the usually exploding number of B-terms (the notion of B-term is explained in Section 'Knowledge discovery: mining implicit and novel information' below). As with the previous system, a number of well known but also entirely new hypothetical relations could be found. These results show that the co-occurrence approach is also capable of identifying novel associations between genes, drugs, pathways and diseases that have a high probability of being pharmacologically valid.

Knowledge discovery: mining implicit and novel information

Work within the knowledge discovery framework aims fundamentally at finding relations between entities that are not explicitly spelled out in the underlying documents. This area of research, rather than locating already known facts literally mentioned in the document, is primarily devoted to mine novel and often hypothetical knowledge given appropriate evidence from text. Two completely different approaches prevail (much in the spirit of the systems described in the last two paragraphs of the previous section)—one has its roots in information retrieval and capitalizes on distributional and statistical characteristics of terms, the other derives from artificial

intelligence and employs logical reasoning on text-derived knowledge bases.

The so-called B-list paradigm (or ABC model) originating from the seminal work of Swanson (for a survey, see [118]) has been the most influential distributional model. Its rationale can be described as follows: if terms A and C (both being highly relevant for the problem under scrutiny) do not co-occur in the same text, then, first, consider all nontrivial terms B_i and B_j that co-occur with A and C, respectively. Those terms B_k that lie in the set-theoretical intersection of B_i and B_j, the so-called B-list, might be evidence for reasonable associative, and thus implicit, bridges between A and C. However, each of the AB_kC hypotheses is kind of speculative, an educated guess, and therefore, must be empirically tested to determine whether or not it is valid in the corresponding domain of discourse. See also the reviews of earlier work that follows this paradigm in [119, 120], whereas more recent activities are portrayed in [116, 121, 122].

Creating knowledge from logical inferences is increasingly becoming an issue for the life sciences; see, e.g. [69, 79, 115, 123]. It requires, however, a formal logical specification of domain knowledge, usually a time-consuming and theoretically challenging (research) activity. Although strictly deductive reasoning has its inherent limits for finding 'novel' knowledge, nondeductive reasoning modes (for instance, inductive or abductive reasoning; see, e.g. [124]), can indeed find truly 'new' knowledge. However, this comes at the price of not being able to guarantee (unlike the case of deductive reasoning) whether or not the newly derived assertions are logically true (valid assertions). From a life science perspective, such considerations might by an exaggerated argument that is outweighed by the heuristic potential for finding and much-wanted guidance for focusing on 'interesting' claims.

Summarizing remarks on text mining approaches

Text mining software builds on, and often combines, three different types of methodological approaches: lexical resources, rule systems and machine learning.

Lexical resources incorporate a wide range of term repositories—simple nomenclatures (term lists, no additional information), linguistic and domain-sensitive dictionaries (which add morphological, syntactic or semantic knowledge, e.g. synonyms, to terms), terminologies (which add conceptual, i.e. taxonomic or partonomic, knowledge to terms) and ontologies (which are based on some formal knowledge representation format, e.g. logics or graphs and thus, in principle, allow formal reasoning). Fortunately, the life sciences are incredibly rich in these kinds of resources [see the multitude of lexical repositories assembled at sites such as the UMLS, the Open Biological Ontologies (OBO; http://www.obofoundry.org/) or the NCBO BioPortal (http://bioportal.bioontology.org/)] and thus, provide TM systems with richly structured background knowledge of the underlying domain of discourse, if coverage permits (see also Section 'Terminologies and ontologies'). On the flipside of this diversity of resources lies the lack of interoperability of different terminologies that TM systems have to account for [125, 126].

Rule-based systems are used to specify systematic extraction patterns usually observable at the surface level of literal mentions in a text in order to capture relevant structural configurations, as well as syntactic or semantic variants in natural language. These patterns either reflect regularities of lexical strings (i.e. sequences of words or tokens) or graphs, mostly parse trees or directed acyclic graphs (DAGs), which represent syntactic structures of natural language. There is a long tradition in biomedical NLP using regular expressions (RegExs) to specify relevant patterns, or using corresponding finite-state automata (FSAs) for pattern recognition. Formally more expressive are context-free rule systems (CFGs) or computationally more flexible production rules, which are typically used to formulate elaborated grammar systems for natural (sub)languages, either subscribing to a dependency-based or a constituency-based representation format (both are, to a certain degree at least, translatable in both directions). Earlier systems contained mostly manually created rules and were also maintained manually; increasingly, rules are automatically learnt or adjusted.

Machine learning is today's dominant paradigm of research in NLP and human language technology. At the core of ML technology lies the identification of distinctive features relevant for classification, i.e. deciding on class membership (e.g. for each word in a text a classifier decides whether or not it belongs to the class Gene or Disease), and the induction of statistical decision models (classifiers) from some positive ground truth (usually, relying on annotated corpora; see Section 'Annotated corpora'). ML methodology covers a wide variety of techniques—

Hidden Markov Models (HMMs), Maximum Entropy (MaxEnt) models, Conditional Random Fields (CRFs) and different kernel parameterizations of Support Vector Machines (SVMs) currently constitute the most successful approaches; for textbooks on ML, see, e.g. [127, 128], a tutorial perspective on ML methods for biomedical NLP is provided in [129]. As with lexical resources, researchers can choose among many alternatives, in this case algorithms implementing a large variety of ML techniques, which can be accessed at some general purpose ML site such as WEKA (http://sourceforge.net/projects/weka/), or NLP-oriented sites such as Mallet (http://mallet.cs.umass.edu/), OpenNLP (http://incubator.apache.org/opennlp/) and LingPipe (http://alias-i.com/lingpipe/) (see also Section 'Software infrastructure').

The ML approach reflects the commonly shared insight in the NLP community that natural languages constitute an extremely complex system of regularities at many different, yet highly interrelated levels (words/lexicology, sentences/syntax, meaning/semantics, situatedness and intentions behind utterances/pragmatics, discourse structures/text linguistics) that has resisted manual descriptive efforts for almost a century in modern linguistics. Note that even after decades of intensive linguistic research, a complete grammar and lexicon have not been worked out manually for any of the natural languages, including English. Rather than further continuing to specify these regularities manually, as a way to get out of this daunting dilemma, NLP has moved to just providing reliable linguistic metadata manually (i.e. from an expert perspective, these are valid assertions about linguistic structures and regularities in terms of annotations for the different layers mentioned before, so-called ground truth; e.g. 'horse' is a NOUN, 'a horse' is a NOUN PHRASE, 'horse' denotes a type of ANIMAL, etc.). The task on which humans have failed for a long time is delegated to the machine (i.e. ML algorithms), namely to automatically find the essential linguistic regularities, usually a statistical model, underlying TM, e.g. named entity recognition or relation extraction. Because the provision of this ground truth for supervised ML is costly and time-consuming (note the quite limited sizes of the annotated document collections mentioned so far), efforts are also under way to learn these regularities from the scratch, i.e. from raw, nonannotated data in an unsupervised ML mode. As can be expected though, supervised ML in the vast majority of applications outperforms unsupervised ML because of the surplus of discriminative knowledge contained in the metadata.

For assessing the usefulness of such methodologies, evaluation experiments are run where some baseline system is usually compared with the particular approach under scrutiny. For example, for named entity recognition, often a hybrid solution involving lexical resources, rules or ML approaches is compared against a lexical off-the-shelves baseline (often including some simple forms of variant normalization, e.g. morphological stemming, easily available synonym lists, etc.). The hypothesis being that the new approach outperforms lexical matching based on an already known term list. For relation extraction, (automatically learnt) rule systems and ML approaches co-exist, with emphasis on the latter though. A typical baseline for evaluating relation extraction builds on co-occurrence where n terms constituting such a relation are required to co-occur in some formal text segment, usually at the sentence level, without any further structural constraint. The hypothesis being that the new approach outperforms simple co-occurrence because of the incorporation of additional structural constraints. Note that co-occurrence relations are un-typed, i.e. their semantics is merely association-based. Only typed relation extraction typically provides a meaningful linkage between the entities involved in a relation (e.g. PPI relations indicate that proteins or genes are supposed to interact; hence, 'Interact' can be taken as the specific semantic type (ako predicate symbol) linking the gene/protein arguments within its scope).

Both the accuracy of the recognition (and normalization to identifiers in standard gene database resources such as UniProt or EntrezGene) of the basic named entities, as well as relation types still fall short of performance figures available for nonbiological entities and relations (e.g. Person, Organization or Location, as well as Employee-of, is-Successor-of, etc. typical of the newspaper domain) by rates from 10% to 15%. Although we presented lots of measurement efforts (mostly F-scores) for the different TM tasks, the concrete values are by no means directly comparable and thus should be considered with utmost care. The parameters that influence the reported performance data are manifold. Most important are the following distinctions and methodological considerations:

- Named entity recognition versus relation extraction (recognition).

- Arity (i.e. the number of arguments) of relations (binary versus ternary, quadruple, etc.).
- Recognition versus normalization (to some already established, community-wide used identifier system, e.g. UniProt, MeSH).
- Single-type versus multi-type entity recognition or relation extraction (e.g. Disease only, versus Disease, Treatment and Test).
- Feature population of ML algorithms—although we often mention, for instance, CRFs or SVMs as ML algorithms of choice for certain applications, the performance we report for them is mostly dependent on the feature set with which they are equipped (this is optimized in the training and development phases); hence, any discussion of results for some ML algorithm must be seen in the light of the feature set selected for a specific experiment—there is no superiority of a particular ML technique as such. Although some standard feature subsets are applied in the vast majority of cases (e.g. lexical features extracted from some canonical lexical resource such as the MeSH), top performance can only be achieved by introducing innovative (combinations of) feature classes.
- Text genre of the corpus which includes condensed texts, such as text snippets from biological or clinical databases or abstracts from scientific publications, versus full texts of scientific publications, clinical notes or patent documents versus social media texts from blogs, mailing lists or other formats where threading and sender–recipient relations are crucial.
- Size of the corpus—with commonly low sizes (the number of documents ranging between 100 and 300, and token sizes ranging between 30 K and 100 K; see also Section 'Annotated corpora') the representativeness of a corpus for the underlying text genre, and hence the generalizability of the results for the chosen task (are the results statistically significant?), is somewhat hard to claim.
- Sampling method for the corpus—often content-directed PubMed queries are run (e.g. 'human' and 'disease') to assemble a focused document collection, which is then used for training and testing purposes; even the replication of such queries at different time points does not necessarily lead to the same corpus because PubMed is dynamically changing its MEDLINE data.
- Metrics for the evaluation of the quality of a text mining system and its results—although we have predominantly been dealing with *F*-scores here, precision@*k*, recall@*k* (*k* stands for a rank position in an ordered list), accuracy, as well as the Relative Operating Characteristic or Receiver Operating Characteristic (ROC) curve and others are reasonable alternative metrics for proper evaluation, though they are not directly comparable.
- Quality of the training material and gold standard for testing—this relates to the authors of the data set in terms of their domain expertise, education level (e.g. graduates versus PhD students, experienced versus nonexperienced physicians) and mutual agreement on their classification decisions for different recognition tasks [measured by the inter-annotator agreement (IAA), see also Section 'Annotated corpora'].

Under these premises and caveats, an *F*-score of 80.0–85.0 (±5.0) for life science entity types (recognition, as well as grounding in companion databases) and 55.0–65.0 *F*-score for life sciences relations (PPIs versus DDIs, respectively) constitute the current state of the art. Overall, recognition rates vary substantially among different biomedical entity types (see the results reported in Sections 'Genotype mining' to 'Mining pharmacological information'). One reason for the comparatively lower level of performance may be that biomedical TM still suffers from a certain poverty of reference data, since really large annotated document sets (comparable with those available for newspaper analytics) are still missing. The CALBC silver standard initiative [130, 131] can be considered as a step in the direction of addressing this problem; see the discussion of CALBC in the final paragraph of Section 'Annotated corpora'.

From a text genre perspective, it turns out that preselected snippets from free-text fields in databases are easier to deal with than scientific abstracts, the currently still prevailing resource for pharmacogenomic text analytics. Scientific full texts are harder to analyze than their associated abstracts. Clinical reports and patents—highly relevant for pharmacogenomics—present the strongest challenge for TM because they usually even exceed the level of linguistic complexity found in scientific full text publications.

While in the biological TM community, the emphasis of work on named entity recognizers and relation extractors is devoted to genes or proteins and their interactions, the medical TM community

focuses on disease and drug recognition, as special cases of phenotype- and pharmacology-oriented TM, respectively. Both camps are actively pursuing their research agenda and benefit a lot from the achievements made within competition-based challenges—as evidenced, e.g., by BioCreAtIvE for the biological community and by the i2b2 Challenge or the DDI Extraction Challenge for the medical community—see also Sections 'Phenotype-focused corpora' and 'Mining pharmacological information', respectively.

INFRASTRUCTURE RELEVANT FOR MINING THE PHARMACOGENOMICS LITERATURE

TM relies on the availability of a considerable amount and variety of resources—for setting up a system (e.g. training data for ML-based systems), for testing it (gold standard data), for implementing it according to good software engineering standards (e.g. middleware frameworks), for providing knowledge of the underlying domain (in terms of terminologies and ontologies) and for reporting the acquired knowledge to the biomedical and pharmaceutical users in a comprehensible way (e.g. visualization tools). This infrastructure will be described in the subsequent subsections.

Annotated corpora

In recent years, we have seen an enormous growth of document corpora annotated with relevant biomedical named entities and relations. Typically, human experts (annotators) while reading (snippets of) raw text data assign special kinds of metadata, so-called tags, from a predefined tag set to relevant stretches of text based on coding conventions that are laid down in annotation guidelines (on which the annotators were trained). Assuming we were given the sentence *'NF-kappa B may activate the production of TNF-α'* and the task would be to assign biological named entity tags of the type TranscriptionFactor *(TF)* and Gene to this sentence. Then an annotation would encapsulate the relevant text tokens and look like the following: '<TF>*NF-kappa B*<\TF> *may activate the production of* <GENE>*TNF-α*<\GENE>'.

TM relies on the availability of such metadata for at least two reasons. First, in order to evaluate TM systems, some undisputed ground truth must be provided against which system performance can be measured. Second, the development of a large proportion of TM systems relies on some sort of supervised ML approach for their named entity and relation extraction classifiers and annotations implement that supervision. In order to induce the models of these classifiers, informative input data has to be provided for model generation. Even for setting up and maintaining manually created rule systems, some credible and diverse data resource is needed to formulate rules that cover a maximal variety of linguistic phenomena.

The creation of corpora, i.e. annotation with metadata, is a time-consuming and intellectually demanding task—both for those who plan and manage such an annotation project, as well as for the staff involved in actually generating the annotations. Wilbur *et al*. [132] define five qualitative dimensions along which this process can be structured, including the design of adequate annotation guidelines and measurement of inter-annotator agreement. Appropriate software tools for supporting the annotation process are also important. For example, the Knowtator system [133] is a general-purpose text annotation tool that is integrated with the Protégé knowledge representation system (http://protege.stanford.edu/). Knowtator facilitates the creation of training and evaluation corpora for a variety of biomedical language processing tasks (http://knowtator.sourceforge.net/). It may also be used to view text-mined relationships, if stored in the appropriate Protégé frames format.

Similar to the importance of the *F*-measure (and its constituent precision and recall metrics) as a means to quantify system performance, the measurement of agreement among several annotators (inter-annotator agreement, IAA) allows to assess the quality of an annotated corpus. It captures the overall consistency of the judgments of a group of usually human annotators as an indicator of shared understanding of the contents of a document (for a survey of various metrics, see [134]). Depending on the complexity of the task, it is common (although often challenged) received wisdom to consider different values as acceptable based on community consensus. One of the most widely used IAA metrics is the kappa statistic, which measures the agreement of annotators adjusted for chance agreement (see, e.g. [135, 136]). Kappa values lie in the interval [0,1], with '0' indicating complete lack of agreement and '1' indicating perfect agreement between the annotators. Since some of the statistical assumptions holding

for kappa might not carry over to corpus annotation in the biomedical domain (e.g. the non-independence of terms occurring in biomedical taxonomies), the *F*-measure using one annotator as the gold standard for the other(s) can be a more adequate alternative [137].

Particularly interesting are several attempts at saving annotation efforts (in terms of time and in terms of the number of decisions to be made) in the course of the annotation process. A 'weak' approach to tackle time-consuming human annotation was proposed by Craven and Kumlien early on [138], who re-used database information for building training material for machine learners. A considerably 'stronger' approach relies on Active Learning, where random selection of examples to be annotated is replaced by an intentional sampling bias that favors the elicitation of human classification decisions on 'hard' (i.e. difficult to interpret) annotation instances, whereas the easier ones are already dealt with automatically using the model learnt thus far. This procedure resulted in almost dramatic cost savings for coding biological named entities along the above mentioned dimensions, ranging from 48% to 72% of the number of tokens to be considered, without seriously sacrificing annotation quality (see, e.g. [139]).

A totally different idea underlies the CALBC annotation approach. Rather than eliciting human judgments to build a gold standard manually, the organizers of the CALBC initiative have constructed a so-called 'silver standard' from the contributions of several automatic named entity taggers only. This ensemble-style procedure requires some harmonization and voting efforts to create a consensus annotation. There is a large variety of named entity types being covered, with emphasis on Genes/Proteins, Diseases, Drugs and Species. For Genes/Proteins, recognition numbers for entity taggers trained on CALBC data of around 60 *F*-score are reported. For Diseases, an *F*-score of 79 is reached [130, 131]. CALBC features not only an unusual diversity of entity types, but is also unmatched with respect to the sheer number of documents being annotated. Around 1 M (automatically) annotated MEDLINE abstracts constitute the largest annotated corpus ever created in the biomedical community. This research, however, needs further validation, since large-scale usage of that corpus and comparisons with much smaller corpora and with human annotations are still lacking.

Genotype-focused corpora

In the past years, the BioNLP community has generated a plethora of PPI-annotated corpora. Whereas earlier attempts [e.g. LLL (http://genome.jouy.inra.fr/texte/LLLchallenge/), AIMed (ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/) and BioInfer (http://mars.cs.utu.fi/BioInfer/), as well as several iterations of BioCreAtIvE] dealt with this issue at a fairly general level of genes and proteins and binary PPIs (see also a quantitative comparison of five PPI corpora, including LLL, AIMed and BioInfer, based on PPI extraction performance [140]), requests for biologically finer-grained distinctions of the complex interactions between genes and proteins were pronounced. This led, for example, to the development of the GENIA event corpus [141] and the GENIA event corpus-derived BioNLP 2009 Shared Task data [18], which consist of 1 K abstracts (more than 9 K sentences and 36 K event annotations) and 1.2 K abstracts (more than 11 K sentences and 14 K event annotations) from MEDLINE, respectively, and contain detailed annotations of PPIs (amongst others; see also [142] for related efforts). The BioNLP Shared Task was a first step toward the extraction of specific pathways with precise information about the molecular events involved. This level of specificity is certainly needed to account for the analysis of the pharmacogenomic literature, particularly when interfaced with pathway databases such as KEGG.

The GeneReg (Gene Regulation) corpus [143] consists of 314 abstracts dealing with the regulation of gene expression in the model organism *E. coli*. The emphasis of this work is on the compatibility and thus, linkage of the GeneReg corpus with the GENIA event corpus and with several in-domain and out-of-domain lexical resources, e.g. the biomedical sublanguage Specialist Lexicon (http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html), as well as general-language lexicons such as WordNet (http://wordnet.princeton.edu/) or FrameNet (https://framenet.icsi.berkeley.edu/fndrupal/). Such links are crucial to broadening the lexical coverage of TM systems. More recent work aims at the annotation of full texts (complementing the annotation of abstracts), as well as the broadening of the number of entity types under scrutiny, e.g. including compounds, biochemical reactions, physiological states and laboratory techniques as well [144].

Phenotype-focused corpora

While genotype-focused corpora deal almost exclusively with scientific publications, phenotype-focused ones are mainly concerned with clinical notes. They vary considerably in structure, genre and the type of jargon (technical language); see also Section 'Phenotype mining' and the survey by Meystre *et al.* [5]. Consequently, the types of entities and relations being used also differ markedly from genotype-focused corpora. A true representative of this stream of work is the CLEF (CLinical E-Science Framework) corpus [145–147], which is composed of clinical narratives, histopathology reports and imaging reports from 20 K cancer patients. For each of these three genres, 50 documents were meticulously annotated with several disease-specific types of clinical entities, namely Condition (including symptom, diagnosis, complication, conditions, problems, functions, processes and injury), Result (the numeric or qualitative finding of an investigation, excluding Condition) and Locus (the anatomical structure or location, body substance or physiological function, typically the locus of a Condition). Very often, Conditions are mentioned in relation to Locus as, for example, in '$[\text{melanoma}]_{\text{Condition}}$ located in $[\text{groin}]_{\text{Locus}}$' or '$[\text{left breast}]_{\text{Locus}}$ $[\text{cancer}]_{\text{Condition}}$'. Furthermore, several relation types are annotated, including HasFinding, HasIndication, HasLocation, HasTarget and Modifies. For time-sensitive named entities even temporal annotations (such as Before, After, Overlap, Includes) based on the TimeML TIMEX3 standard [148] are provided (http://timeml.org/site/timebank/documentation-1.2.html). Thus, the annotation process for diseases is broken down into the annotation of many diverse fundamental clinical and anatomical entities and their relationships. A wide range of IAA scores are reported for such a relational decomposition of annotation, with kappas ranging widely (depending on the type of entity and relation and the way how IAAs were measured—strict or lenient (partial) match), which indicates that this fine-grained relationship annotation for clinical entities is an extremely difficult task.

On a similar quantitative scale, Ogren *et al.* [149] report on a corpus which contains 1556 annotations on 160 clinical notes using 658 unique concept codes from SNOMED-CT corresponding to human disorders. IAA for four annotators is reported, among others, for span (0.91) and concept code (0.82). In earlier work, Pestian *et al.* [150] describe one of the rare clinical notes corpora, composed of almost 2 K documents annotated at the document level for billing codes (45 categories taken from the disease classification ICD-9CM; http://www.cdc.gov/nchs/icd9.htm).

Completely different text genres were considered for two other disease-centric corpora, namely the Disease Corpus from the EBI [38] and the Arizona Disease Corpus (AZDC) [151]. Both deal with disease annotations only, a proper subset of pathological phenomena. The EBI corpus contains 600 sentences from OMIM, for which an IAA of 0.51 kappa (which is low, even by biomedical standards) is reported for two annotators. AZDC provides 3228 disease mention annotations (1202 unique disease names) for 2856 MEDLINE abstracts. Mentions of organisms and species were explicitly excluded from the disease annotation span. So, for 'human insulin-dependent diabetes mellitus', the span 'insulin-dependent diabetes mellitus' would be annotated as Disease.

Furthermore, there exist highly specialized corpora that deal with particular disease types only. For example, Cano *et al.* [152] discuss the development of annotation guidelines for PPIs and gene–disease relations and report agreement rates of 75% for a small autism MEDLINE collection made of 139 snippets. The PennBioIE corpus [153] comes with an oncology portion made of 1414 MEDLINE abstracts annotated for the molecular genetics of cancer, as well as a genotype portion made of 1100 MEDLINE abstracts annotated for the inhibition of cytochrome P-450 enzymes (http://lists.ccs.neu.edu/pipermail/bionlp/2008-November/001208.html). The oncology portion, finally evolved into the PennBioIE Oncology 1.0 corpus (http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T21).

The most diverse corpus construction efforts for the medical informatics community are currently centred around various rounds of the i2b2 initiative (https://www.i2b2.org/). It has turned into a series of competitions (comparable with BioCreAtIvE) that provides (similar in spirit with CLEF) clinical document resources. In the first contest, 502 de-identified discharge summaries were provided with smoking status annotations [154]. For the next challenges, annotations for obesity [155] and medication data [59] were provided. For instance, the classification task for the smoking challenge—identifying each patient either as a smoker, current smoker, past smoker,

nonsmoker or unknown—was solved by the best-performing system with accuracy peaking at 93.6 using an SVM approach [156]. In the latest 2010 Challenge, i2b2 has moved its thematic scope to account for clinical named entity recognition and relation extraction tasks [96]. The winner system achieved for concept extraction (finding entities related to Problems, Tests and Treatments) an *F*-score of 85.2, for assertion detection (asserting for each Problem whether the context describes it as 'present', 'absent', 'possible', 'conditional', 'hypothetical' or 'associated with someone else'.) 93.6, and for relationship detection (entities co-occurring with Problems) 73.1 [157]. The different i2b2 corpora not only contribute lots of diverse annotation types to the community but also extend the thematic scope beyond well defined diseases into less discrete types of pathological phenomena. Chapman *et al*. [158] discuss the value of i2b2-style challenges for the biomedical NLP community but also point out special problems encountered in trying to adopt results from this research in the clinical domain.

Drug- and chemicals-focused corpora

Corpora focusing on the annotation of drugs and chemicals not only consider MEDLINE abstracts but also involve full-text journal articles, patent texts and free-text fields from relevant databases. Corbett *et al*. [159] report on a corpus built from 42 full-text chemistry papers annotated with chemical names from five categories (among them Compounds and Enzymes), while Kolárik *et al*. [160] discuss the annotation of the Fraunhofer chemical names corpus, which contains about 1500 MEDLINE abstracts (The corpus is available at http://www.scai.fraunhofer.de/chem-corpora.html).

The DISAE corpus constructed by a Fraunhofer team [161] originally contained 400 MEDLINE abstracts randomly selected from the PubMed query 'Disease or Adverse effect'. Two annotators (with 0.84 and 0.89 kappa) provided 1428 Disease and 813 Adverse effect annotations, although information on the drugs involved in adverse effects was not annotated. Since the original publication, the corpus has grown to almost 3000 MEDLINE documents containing approximately 5000 Drug and 5800 drug Adverse effect annotations [108].

DrugDDI, the corpus constructed for the DDI Extraction Challenge [64], is based on 1000 randomly chosen drug names selected from the DrugBank database, for which links to documents describing DDIs in unstructured texts were extracted. These documents were then analyzed with the UMLS MetaMap Transfer (MMTx) tool, leading to a linking of phrases with UMLS Metathesaurus concepts. This corpus contains 579 documents describing 3160 positive DDIs.

The PharmGKB repository (http://www.pharmgkb.org/) [162] comes perhaps closest to the vision of an all-embracing pharmacogenomics corpus. It represents a major step towards an interdisciplinary biomedical information store. PharmGKB incorporates data on genetic variations and associated phenotypic manifestations [163]. The resource covers information on the pharmacokinetics of therapeutic drugs (how drugs are absorbed, metabolized and excreted by an organism) and the pharmacodynamics of drugs (how drugs act in an organism). It also covers certain nonpharmacological aspects of phenotypes, including susceptibility to disease. Currently (as of December 2011), PharmGKB contains information about 1164 pharmacogenes and 1753 drugs and the relations between them; and it continues to grow.

The crucial role it might play as a future benchmark standard for pharmacogenomic TM has just been demonstrated in a recent study by Buyko *et al*. [164]. They used PharmGKB as a resource for retraining the JReX system (which was second-best in the 2009 Shared Task on Event Extraction, with 46.7 *F*-score) and adapting it to account for Gene–Drug, Gene–Disease and Drug–Disease relations. Data are presented for an internal 10-fold cross-validation on the PharmGKB corpus [Gene–Drug (*F*-score: 82.3), Gene–Disease (*F*-score: 76.0) and Drug–Disease relations (*F*-score: 79.0), with an overall *F*-score of 80.1], as well as for an external evaluation incorporating PharmGKB relation test sets [Gene–Drug (*F*-score: 73.6), Gene–Disease (*F*-score: 68.8) and Drug–Disease relations (*F*-score: 77.5), with an overall *F*-score of 73.3].

Terminologies and ontologies

Terminologies and ontologies have a variety of uses in pharmacogenomic TM, ranging from providing targets for the grounding of concepts mentioned in text to addressing the need for a structured definition of the domain for the purpose of annotating corpora. The latter use is not typically seen, and when it is used, *ad hoc* ontologies are typically applied, as in the case of the GENIA and BioInfer corpora. An exception to this is the CRAFT corpus [165], which was

annotated with reference to several OBO ontologies, such as the GO and ChEBI (http://bionlp-corpora.sourceforge.net/CRAFT/index.shtml).

Genotype-focused resources

In pharmacogenomics, a primary need is to unambiguously identify and refer to genes, diseases and drugs. Contrary to popular misunderstanding, the GO (http://www.geneontology.org/), currently containing 35 500 fully defined concepts, is not about genes, but rather gene functions, broadly construed. For obtaining unambiguous names for referring to genes, there are several options, such as using EntrezGene identifiers from NCBI. For human readable names, the HUGO Gene Nomenclature Committee (HGNC) has assigned unique gene symbols and names to more than 32 000 human loci (http://www.genenames.org). It is a curated online repository of HGNC-approved gene nomenclature and associated resources, including links to genomic, proteomic and phenotypic information, as well as dedicated gene family pages.

Phenotype-focused resources

Disease name standardization is important in pharmacogenomics for the purpose of combining observations from different studies, databases or texts. One of the most elaborate ontologies for diseases [we here use the term 'ontology' to refer to artifacts—(informal) terminologies, as well as true (i.e. formalized) ontologies—that can provide a hierarchy of parent–child terms for disease conditions] is the Systematized Nomenclature for Medicine-Clinical Terms (SNOMED CT; http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html), which is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. Currently, SNOMED CT contains more than 311 K active concepts with unique meanings and formal logic-based definitions organized into multiple hierarchies. The disease hierarchy is available under the 'Clinical Finding' root node (analogous to the 'Biological Process' root node in GO).

Another widely used disease ontology is the National Cancer Institute thesaurus (NCIt; http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources; the NCIt thesaurus can be browsed via the NCI Term Browser available at http://ncit.nci.nih.gov/ncitbrowser/). The NCIt provides definitions, synonyms and other information about nearly 10 000 cancers and related diseases, 8000 single agents and combination therapies and a wide range of other topics related to cancer and biomedical research. NCIt is a recognized standard for biomedical coding and reference, used by a variety of public and private institutions, including the Clinical Data Interchange Standards Consortium Terminology (CDISC), the US Food and Drug Administration (FDA), the Federal Medication Terminologies (FMT) and the National Council for Prescription Drug Programs (NCPDP). The disease hierarchy is available under the root node 'Diseases, Disorders and Findings'.

The most widely used disease ontology, which also includes a wide variety of signs, symptoms, abnormal findings, complaints, etc. is the ICD, which is part of the WHO Family of International Classifications. Version 10 of ICD contains 155 K different codes (http://www.cdc.gov/nchs/icd/icd10.htm), Version 9 of ICD (http://www.cdc.gov/nchs/icd/icd9.htm) is widely used in the United States for billing purposes in the health care system, although a conversion to 10 is mandated in the near future. Finally, there is an effort to create an ontology of human diseases that conforms to the principles of the Open Biomedical Ontologies Foundry (http://www.obofoundry.org/). This Human Disease ontology has been under review by the OBO Foundry since 2006.

Drug- and chemicals-focused resources

Drugs are another entity type that needs to be consistently named and referred to in pharmacogenomics TM. Analogous to the different disease ontologies, there are several alternative ontologies for drugs. For researchers interested in a consistent way of naming clinical drugs, *RxNorm* is an excellent resource, containing around 190 K terms and >800 K relationships (http://www.nlm.nih.gov/research/umls/rxnorm/; statistics are available at http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/RXNORM/sourcemetastats.html). Over the past 6 years, RxNorm has become a central resource for communicating about clinical drugs and supporting interoperation between drug vocabularies. It provides names and relationships for medications at a given level of abstraction. The current version is based on 11 different source vocabularies, including the National Drug File Reference Terminology (NDF-RT) from the Veterans Administration [166].

ChEBI (http://www.ebi.ac.uk/chebi/) is a freely available ontology of molecular entities focused on 'small' chemical compounds, with almost 27 K entries. The molecular entities in question are either natural products or synthetic products used to intervene in the processes of living organisms. The term 'molecular entity' refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc. identifiable as a separately distinguishable entity. ChEBI includes an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents/children are specified. ChEBI uses nomenclature, symbolism and terminology endorsed by the IUPAC and Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). All data in the CheBI database are nonproprietary or are derived from a nonproprietary source [167] (The text on CheBI is from http://www.ebi.ac.uk/chebi/aboutChebiForward.do and is used under http://creativecommons.org/licenses/by/3.0).

Software infrastructure

Recent years have seen considerably increased attention to software engineering and infrastructure issues, even in the domain of academic software construction. This has led to the development of end-to-end systems that realize fully-functioning, high-throughput-scaled TM (e.g. capable of processing the whole of MEDLINE, currently comprising 21 M documents) that can be easily combined with other TM components to create full-text processing pipelines.

Much current attention has been focused on the Unstructured Information Management Architecture (UIMA; http://uima.apache.org/) [168]. UIMA was originally developed by IBM to facilitate the processing of any sort of unstructured data, ranging from free text to audio and video. It has since been released Open Source and has seen its greatest use by far in the context of TM and NLP. UIMA is based on the simple premise that all TM components should have an interface that consists of passing along 'annotations' of the input text in the form of character offsets into the original document. This simple contract allows for flexible integration of any TM component that adheres to the character offset constraint.

UIMA systems are based on semantic characterizations of the nature of the annotations that are added in the course of the document analysis by TM components. The limiting factor in interoperability of UIMA-based text analytics so far has been the lack of a community-wide consensus on these semantic characterizations, known in UIMA parlance as 'type systems.' Influential efforts in the direction of UIMA TM components and type systems have included work from the JULIE Lab [169, 170], the U-Compare project [171] for genomics publications and the Open Health Natural Language Processing (OHNLP) Consortium for clinical documents (http://www.ohnlp.org/), which among other systems includes the Mayo Clinic cTAKES system [42].

A striking example of what can be achieved with the UIMA architecture is the U-Compare project [171]. It offers a fully-featured platform for evaluating TM workflows and components (http://u-compare.org/). It makes available corpora, tools and evaluation metrics, and allows users to assemble their own workflows through a simple drag-and-drop interface, or to carry out TM tasks without assembling workflows on their own by using predefined workflows. Performance statistics are generated automatically, making U-Compare a powerful tool for evaluating the contribution of different TM components to overall system performance. For example, a user might want to compare the performance of a PPI extraction system when different gene mention taggers are used; U-Compare makes such a comparison simple. Quite recently, nine event extraction systems have been integrated in the U-Compare framework, making them intercompatible and interoperable with other U-Compare components [172].

One of the commonest uses of ontologies in pharmacogenomics TM is as a source of lexicons for the entities of interest (genes, diseases and drugs). Several efforts have appeared in recent years that facilitate the use of ontologies for lexicon generation. For example, using a user-provided textual corpus, the Ontology Recommender Web service from the NCBO [173] suggests which ontology to use as a source of standard terms to tag the corpus. In addition, the Lexicon Builder Web service from NCBO [174] reduces the investment of time and effort involved in lexicon creation from ontologies. The service has three components. Inclusion selects one or several ontologies (or its branches) and includes preferred names and synonym terms. Exclusion filters terms based on the term's

MEDLINE frequency, syntactic type, UMLS semantic type and match with stop words. Output aggregates information and handles compression and output formats. Similarly, the BioLexicon (http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/biolexicon.html) effort [175] gathers together different types of terms from several existing data resources into a single, unified repository, and augments them with new term variants automatically extracted from biomedical literature. The BioLexicon was built specifically for the purpose of enabling TM tools at several levels.

Disseminating and presenting the results of text mining

Disseminating the results of TM to the biomedical target community is an ongoing challenge. As a corollary to the advances in NLP and TM, the role of the World Wide Web for knowledge generation, distribution and exchange, as well as the increasing sophistication of the methods being used, the sheer size and complexity of the output has increased considerably. Different systems output different attributes of the recognized relationships between entities, and there are efforts underway to harmonize and interact with the various presentation formats to enable life scientists to deeply comprehend and flexibly use the results of TM for data analysis, model building, as well as inference and reasoning.

Accordingly, we not only witness the trend towards performing text analytics over the Web, but also presenting their results over the Web. Similar to U-compare (see the previous Subsection), Web services are available that perform both disease and drug tagging based on the UMLS and the NCBO BioPortal, and thus automatically add valuable semantic metadata to documents [176]. There are efforts underway that espouse the use of an ontology of relationships—which can be learned during the process of TM—that can aggregate and normalize alternative formulations of the same relationship and present the aggregated 'facts' in a knowledge base which is based on the Resource Description Format (RDF) recommendation by the W3C (http://www.w3.org/RDF/). For example, in the work by Coulet *et al.* [71, 177, 178], the PHArmacogenomic RElationships Ontology (PHARE), comprised of 200 curated relations from over 40 K heterogeneous relationships extracted via TM, serves as a common framework to form the foundation of a knowledge base named PHARE-KB that can be queried via SPARQL. In a similar vein, Dumontier and Villanueva-Rosales [179] populate their Pharmacogenomics Ontology (PO) with data gathered from PharmGKB Web services. They instantiated 40 core concepts incorporating drugs, genotypes, phenotypes and drug treatments, a procedure that yielded 4266 KB instances.

An entirely Web-specific form to publicize scientific data and results are so-called 'nanopublications' (http://www.nanopub.org/), which are also based on the RDF framework. At its core, a nanopublication has three basic elements: (i) An assertion whereby two entities (called the Subject and the Object) are associated (using a third entity called the Predicate); (ii) Metadata regarding conditions under which the assertion holds and (iii) Metadata regarding the provenance of the assertion, such as its author, a time-stamp marking when it was created, links to DOIs, URLs, etc. [180]. There are also efforts such as Bio2RDF [181] and Chem2Bio2RDF [182], which attempt to facilitate the conversion of diverse data sources (including text-mined relationships) into RDF for publication on the Semantic Web. Also, the Linked Open Data (LOD; http://linkeddata.org/) initiative might play an important role here in the future [183].

Other approaches focus primarily on comprehensively presenting text-mined relationships inside of existing tools (such as the visualization tool Cytoscape; http://www.cytoscape.org/) already in use by many biologists. For example, the Hanalyzer system [184, 185] can present text-mined relationships in context with other high-throughput datasets in a network view to allow interactive exploration. As the complexity of text-mined assertions increases, moving from binary to *n*-ary relations (e.g. consider the following five entities A to E that have an impact on Activates; 'gene A activates protein B in the context of disease C in organ D under the influence of chemical E'), more sophisticated visualization approaches might be required, e.g. involving hyper graphs [186].

Among the most recent developments are the DOMEO framework (http://code.google.com/p/domeo/), which allows users to annotate and then publish relationships in Web pages, and the Utopia document format (https://sites.google.com/site/beyondthepdf/utopia) and its associated annotation framework, which allows the exchange of relationship annotations between online documents and PDFs [187].

APPLICATIONS

A variety of applications are informed by the results of pharmacogenomic TM, such as discovery of interaction and potential cause–effect phenomena (e.g. candidate gene ranking, drug–drug interaction and adverse drug interaction prediction), as well as guidance in human database curation.

Crucial portions of discovery work are already enabled by TM of pharmacogenomic knowledge. Candidate gene ranking, i.e. ranking genes based on an objective relevance score, has often been used in identifying disease genes [72–74, 76, 78]. Past work in the pharmacogenomics arena includes using gene–drug relationships mined from the literature to inform a candidate gene ranking algorithm to predict pharmacogenes for a given drug, thereby replacing the need for using manually curated knowledge [91]. Recent adaptation of the GeneRanker system [74] enables prediction of genes related to ADRs. The methodology enables studying the molecular scope of ADRs, to aid in the study of known ADRs and indicate potential unknown ones.

Drug–drug interactions (DDIs) are another discovery area for pharmacogenomics TM. A DDI occurs when the use of one drug by a patient influences the effects of another drug that is concomitantly used by that patient. This has great clinical value, since the knowledge of a DDI may impact the choice of medication or dosage by the clinician. Tari *et al*. [69] have developed a method that combines TM with automated reasoning to extract novel DDIs. Comprehensive extraction of pharmacogenomic drug–gene relationships, as presented recently in [71, 177], enables inference of known and novel DDIs from existing interactions described in papers buried deep within the scientific literature, across multiple journals and fields [70]. This application area might shift gears with the outcomes of future DDI Extraction Challenges (see also Section 'Mining pharmacological information').

Literature-based ADR prediction and drug repurposing have been a recent area of focus, as surveyed by Deftereos *et al*. [188] and Andronis *et al*. [189] in extensive reviews. Agarwal and Searls [190], as well as Plake and Schroeder [191], review the latest applications of TM and ontologies suitable for target and drug–target interaction discovery, showing how drug discovery, in general, can benefit from TM.

Curation guidance has been a much more recent area of focus by a number of groups (see Winnenburg *et al*. [192] for a survey of TM approaches that are relevant to annotation and available online services analyzing biomedical literature by means of TM techniques). Wiegers *et al*. [193] performed curation guidance experiments for the Comparative Toxicogenomics Database (CTD), a publicly available resource that promotes understanding about the etiology of environmental diseases. It provides manually curated chemical–gene/protein interactions and chemical– and gene–disease relationships from the literature. Prototype TM applications were developed and evaluated using a CTD data set consisting of manually curated molecular interactions and relationships from 1600 documents. The prototype found 80% of the gene, chemical and disease terms appearing in curated interactions. These terms were used to re-rank documents for curation, resulting in increases in mean average precision (63.0 for the baseline versus 73.0 for rule-based re-ranking), and in the correlation coefficient of rank versus number of curatable interactions per document (baseline 14.0 versus 38.0 for the rule-based re-ranking).

Interest in this area of research is gaining pace both on the side of the database curation community [194] as well as on the academic side. A single track in the latest BioCreAtIvE III competition was entirely devoted to TM-based curation support [195]. PharmGKB is developing methods for assisting curators via an NLP pipeline. In the SASEBio project a version of PharmGKB has been processed to investigate the use of a text annotation interface to guide curation [196]. Further work also focuses on automatic approaches to extracting information from the biomedical literature to help expand the scope of PharmGKB (PSB 2011 Workshop; http://psb.stanford.edu/psb-online/proceedings/psb11/wkshop-pharma.pdf).

CONCLUSIONS AND OUTLOOK

The motivation for investing long-standing efforts in the development of methodologies and implementation of systems devoted to automatic TM is 2-fold. First, humans, whether in their role as database curators, as bench scientists or as clinicians, are unable to keep up with the ever-increasing flood of scientific publications [197]. Second, human authors are not good at encoding their own findings in a semiformal representation format, e.g. using controlled terminologies [198].

Fortunately, some optimistic observations can be made relating to certain problems of TM for pharmacogenomics. For example, a series of shared tasks (most notably, BioCreAtIvE) have resulted in the creation of gene mention recognition systems that function at a level nearly as high as that achieved years ago for person names, organizations, and geospatial locations for the analysis of newspapers. Mutations and variants, which are crucial for establishing individual responses to drugs, can be recognized with exceptional accuracy [33, 34]. Low-level TM tasks that appear simple but are actually highly complex, such as finding sentence boundaries and splitting input texts into 'tokens' (the basic unit of textual analysis, including words, as well as punctuation marks), can now be tackled with sophisticated tools like LingPipe that have been specialized for the biomedical domain (see also [199]). Similar tools, e.g. ARC [200] and cTAKES [42], are available for processing clinical documents.

However, other problems remain in need of significantly more progress. In general, performance figures for relation extraction always lag largely behind entity recognition, as have those for the more difficult problem of entity normalization (mapping terms to common database identifiers). Relationship normalization is a particularly difficult problem—both at the linguistic variant level (e.g. 'X activates Y' [active voice] = 'Y was activated by X' [passive voice] = 'the activation of Y (by X)' [nominalization], see, e.g. [201, 178], and at the level of mapping text terms to ontology entries (e.g. GO molecular function terms). Prosaic but important challenges such as extracting data from tables or properly converting PDFs of full text into some processable XML format continue to pose problems for high-coverage TM.

Overall, these are exciting times for TM for pharmacogenomics. NIH policies have resulted in the availability of large collections of full-text journal articles (PubMed Central), which present enormous new opportunities for TM, but significant new challenges as well.

In a recent study, Cohen *et al.* [53] focused on differences between abstracts and full texts. Content-wise they found substantial distributional differences in entity mentions, such as significantly higher frequency of mutations mentioned in the bodies of articles, which did not mention the mutations in the abstracts at all. The authors also evaluated the differential performance of TM tools, reporting, e.g. that commonly used gene taggers perform substantially better in abstracts than in article bodies. Blake [202] discusses empirical evidence for communication patterns of authors which indicate huge losses of empirical statements (e.g. explicit versus implicit claims, under-specified claims such as correlations, comparisons and observations) when comparing the information in abstracts with that in associated full texts.

From the perspective of full-text analytics, other crucial challenges must be handled (see also [203]), e.g. the increased proportion of reference relations of all sorts—pronominal, nominal and bridging anaphora [204–208]—which establish cohesion among sentences at the cost of introducing different mentions for the same entity [e.g. 'IL-7' . . . 'this protein' (=IL-7; nominal anaphora) . . . 'it . . .' (=IL-7; pronominal anaphora)], macro-level forms of text structuring (e.g. formal, layout-oriented text segmentation, but also rhetorical and argumentative zoning [209, 210]; see also recent efforts to provide a biomedical discourse relation bank [211] along the lines of the Penn Discourse Relation Treebank; http://www.seas.upenn.edu/~pdtb/), as well as the problem of interactions between text passages, tables, graphics and other forms of nontextual data [212].

As already mentioned, 'deep' text analytics involving some form of formal reasoning have long been missing in the field, but recently have begun to see some adoption (see, e.g. [79, 69, 115]). Although deductive reasoning might be fully appropriate for taxonomic inferences in OWL-style ontologies (see, e.g. [79, 177, 179, 213] for examples from pharmacogenomics), nonstandard forms of formal reasoning might be more adequate for knowledge discovery tasks (for instance, involving inductive or abductive inferencing; see, e.g. [124]) and for grading the credibility of extracted knowledge (using probabilistic inferencing techniques, e.g. based on Bayesian Network [214, 215]). Furthermore, the inherent 'glue' of complex events, their temporal and causal structure, remains largely underexplored territory up to the present (see the reviews [216, 217, respectively] and as concrete examples, e.g. [218, 219]). We claim that such novel methodologies will have a high potential for advanced text analytics that go beyond merely extracting simple unconnected pieces of explicit information from full-text documents.

This leads us, finally, to the Holy Grail of TM—knowledge discovery. In this area, researchers aim at

finding new pieces of knowledge; knowledge that, unlike in the IE scenario, is not already explicitly stated in natural language documents. Knowledge discovery systems unveil associations between relevant entities, hint at implicit assertions, help find new speculative hypotheses (which then have to be experimentally validated) and assist in shaping assumptions and claims *in silico*. Some implemented systems, such as BITOLA [220], LitLinker [221], the GeneWays project [222] and FACTA + [223], already focus on some aspects of these hard problems, but a stable methodology is not in sight. Certainly, these challenges need years of basic research not only on TM but also on NLP and text understanding proper and its many links to an amazing variety of formal reasoning styles.

Key Points

- We argue that automatic text mining has turned into a viable alternative to human database curation and literature indexing because natural language processing technologies have become robust and mature, are scalable on very large document collections (e.g. millions of Medline documents) and work on a variety of genres in the life sciences (scientific publications, clinical notes, patents, blogs, etc.).
- We review the increasing diversity of named entities and semantic relations among entities that are crucial for text mining in the field of pharmacogenomics, with particular emphasis on genotypes (genes/proteins), phenotypes (diseases, disorders, tests, treatments, etc.) and drugs/chemicals.
- We review the underlying infrastructure for text mining—in terms of annotated document corpora, domain knowledge resources (terminologies, ontologies, etc.), as well as software available open source.
- We point out current applications and opportunities of future research, with emphasis on full text analytics, credibility of automatically harvested data, visualization of large knowledge repositories harvested by text mining systems, automatic reasoning for deeper text analytics and text-based knowledge discovery.

FUNDING

Udo Hahn was partially funded by the German Ministry of Education and Research (BMBF) as part of the National Research Core within the Jena Centre of Systems Biology of Ageing (JENAGE) (grant no. 0315581D). K. Bretonnel Cohen was supported via NIH 5R01 LM009254-06 (Hunter, PI), NIH 5R01 LM008111-07 (Hunter, PI) and NIH 5R01 GM083649-04 (Hunter, PI) grants. Yael Garten received funding from the National Institutes of Health (grant GM61374) and the National Library of Medicine (contract HHSN276201000025C). Nigam H. Shah acknowledges support from NIH grant U54-HG004028.

References

1. Licinio J, Wong ML, (eds). *Pharmacogenomics: The Search for Individualized Therapies*. Weinheim: Wiley-VCH, 2002.
2. Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics* 2010;**11**:1467–89.
3. Krallinger M, Leitner F, Valencia A. Analysis of biological processes and diseases using text mining approaches. In: Matthiesen R, (ed). *Bioinformatics Methods in Clinical Research*. New York (NY): Humana Press, a part of Springer Science+Business Media, 2010;341–82.
4. Rodriguez-Esteban R. Biomedical text mining and its applications. *PLoS Comput Biol* 2009;**5**:e1000597.
5. Meystre SM, Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. In: Geissbühler A, Kulikowski C, (eds). *IMIA Yearbook of Medical Informatics 2008*. Stuttgart: Schattauer, 2008;128–44.
6. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009;**42**:760–72.
7. Manning CD, Schütze H. *Foundations of Statistical Natural Language Processing*. Cambridge (MA): MIT Press, 1999.
8. Jurafsky D, Martin JH. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd edn. Upper Saddle River (NJ): Prentice Hall, 2009.
9. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge (U.K.): Cambridge University Press, 2008.
10. Jackson P, Moulinier I. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization —2nd Revised Edition*. Amsterdam, Philadelphia: John Benjamins, 2007.
11. Feldman R, Sanger J. *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data*. Cambridge (U.K.): Cambridge University Press, 2007.
12. Weiss SM, Indurkhya N, Zhang T, *et al*. *Text Mining. Predictive Methods for Analyzing Unstructured Information*. New York (NY): Springer, 2005.
13. Good BM, Howe DG, Lin SM, *et al*. Mining the Gene Wiki for functional genomic knowledge. *BMC Genomics* 2011;**12**:603.
14. Malouf R, Davidson B, Sherman A. Mining web texts for brand associations. In: *Proceedings of the AAAI–2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*. 2006. Menlo Park (CA): American Association for Artificial Intelligence (AAAI);125–6.
15. Kilicoglu H, Bergler S. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics* 2008;**9(Suppl 11)**: S10.
16. Vincze V, Szarvas G, Farkas R, *et al*. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. *BMC Bioinformatics* 2008; **9(Suppl 11)**:S9.
17. Özgür A, Radev DR. Detecting speculations and their scopes in scientific text. In: *EMNLP 2009—Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. 2009. Stroudsburg (PA): Association for Computational Linguistics (ACL);1398–407.

18. Kim JD, Ohta T, Pyysalo S, *et al*. Overview of BioNLP'09 shared task on event extraction. In: *BioNLP 2009—Companion Volume: Shared Task on Event Extraction*. 2009. Stroudsburg (PA): Association for Computational Linguistics (ACL);1–9.
19. Kim JD, Pyysalo S, Ohta T, *et al*. Overview of BioNLP Shared Task 2011. In: *Proceedings of the BioNLP Shared Task 2011 Workshop*. 2011. Stroudsburg (PA): Association for Computational Linguistics (ACL);1–6.
20. Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinformatics* 2005;**6**:357–69.
21. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: *AI 2006—Proceedings of the Australian Conference on Artificial Intelligence*. 2006. Berlin, Heidelberg: Springer;1015–21 (LNCS 4304).
22. Cleverdon C. On the inverse relationship between recall and precision. *J Documentation* 1972;**28**:195–201.
23. Hatzivassiloglou V, Duboué PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics* 2001;**17(Suppl 1)**:S97–106.
24. Hirschman L, Yeh A, Blaschke C, *et al*. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* 2005;**6(Suppl 1)**:S1.
25. Krallinger M, Morgan A, Smith L, *et al*. Evaluation of text-mining systems for biology: overview of the Second BioCreAtIvE community challenge. *Genome Biol* 2008; **9(Suppl 2)**:S1.
26. Leitner F, Mardis SA, Krallinger M, *et al*. An overview of BioCreAtIvE II.5. *Trans Comput Biol Bioinform* 2009;**7**: 385–99.
27. Arighi CN, Lu Z, Krallinger M, *et al*. Overview of the BioCreAtIvE III Workshop. *BMC Bioinformatics* 2011; **12(Suppl 8)**:S1.
28. Hakenberg J, Plake C, Royer L, *et al*. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biol* 2008;**9(Suppl 2)**:S14.
29. Wermter J, Tomanek K, Hahn U. High-performance gene name normalization with GeNo. *Bioinformatics* 2009;**25**: 815–21.
30. Björne J, Heimonen J, Ginter F, *et al*. Extracting complex biological events with rich graph-based feature sets. In: *BioNLP 2009. Companion Volume: Shared Task on Event Extraction*. 2009. Stroudsburg (PA): Association for Computational Linguistics (ACL);10–8.
31. Riedel S, McClosky D, Surdeanu M, *et al*. Model combination for event extraction in BioNLP 2011. In: *Proceedings of the BioNLP Shared Task 2011 Workshop*. 2011. Stroudsburg (PA): Association for Computational Linguistics (ACL); 46–50.
32. Pyysalo S, Ohta T, Rak R, *et al*. Overview of the Infectious Diseases (ID) task of BioNLP Shared Task 2011. In: *Proceedings of the BioNLP Shared Task 2011 Workshop*. 2011. Stroudsburg (PA): Association for Computational Linguistics (ACL);26–35.
33. Furlong LI, Dach H, Hofmann-Apitius M, *et al*. OSIRISv1.2: a named entity recognition system for sequence variants of genes in biomedical literature. *BMC Bioinformatics* 2008;**9**:84.
34. Caporaso JG, Baumgartner WA, Randolph DA, *et al*. MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics* 2007;**23**: 1862–5.
35. Yeniterzi S, Sezerman U. EnzyMiner: automatic identification of protein-level mutations and their impact on target enzymes from PubMed abstracts. *BMC Bioinformatics* 2009; **10(Suppl 8)**:S2.
36. Laurila JB, Naderi N, Witte R, *et al*. Algorithms and semantic infrastructure for mutation impact extraction and grounding. *BMC Genomics* 2010;**11(Suppl 4)**:S24.
37. Cunningham H, Maynard D, Bontcheva K, *et al*. *Text Processing with GATE (Version 6)*. Sheffield (U.K.): University of Sheffield, Department of Computer Science, 2011.
38. Jimeno A, Jimenez-Ruiz E, Lee V, *et al*. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics* 2008;**9(Suppl 3)**:S3.
39. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**:229–36.
40. Meystre SM, Friedlin FJ, South BR, *et al*. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol* 2010;**10**:70.
41. Kipper-Schuler K, Kaggal V, Masanz J, *et al*. System evaluation on a named entity corpus from clinical notes. In: *LREC 2008—Proceedings of the 6th International Conference on Language Resources and Evaluation*. 2008. Paris: European Language Resources Association (ELRA);3007–11.
42. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507–13.
43. Li D, Kipper-Schuler K, Savova G. Conditional Random Fields and Support Vector Machines for disorder named entity recognition in clinical texts. In: *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*. 2008. Stroudsburg (PA): Association for Computational Linguistics (ACL);94–5.
44. Roberts A, Gaizauskas R, Hepple M, *et al*. Combining terminology resources and statistical methods for entity recognition: an evaluation. In: *LREC 2008—Proceedings of the 6th International Conference on Language Resources and Evaluation*. 2008. Paris: European Language Resources Association (ELRA);2974–80.
45. Leaman R, Gonzalez G. Banner: an executable survey of advances in biomedical named entity recognition. In: *PSB 2008—Proceedings of the Pacific Symposium on Biocomputing*. 2008;652–63. http://psb.stanford.edu/psb-online/proceedings/psb08/ (20 Dec 2011, date last accessed).
46. Rosario B, Hearst MA. Classifying semantic relations in bioscience text. In: *ACL 2004 —Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. 2004. Stroudsburg (PA): Association for Computational Linguistics (ACL);430–7.
47. Segura-Bedmar I, Martinez P, Segura-Bedmar M. Drug name recognition and classification in biomedical texts: a case study outlining approaches underpinning automated systems. *Drug Discov Today* 2008;**13**:816–23.
48. Kolárik C, Hofmann-Apitius M, Zimmermann M, *et al*. Identification of new drug classification terms in textual resources. *Bioinformatics* 2007;**23**:i264–72.

49. Hearst MA. Automatic acquisition of hyponyms from large text corpora. In: *COLING 1992—Proceedings of the 14th International Conference on Computational Linguistics*. 1992. Sheffiled (U.K.): International Committtee on Computational Linguistics (ICCL);539–45.
50. Klinger R, Kolárik C, Fluck J, *et al*. Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics* 2008;**24**: i268–76.
51. Corbett P, Copestake A. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics* 2008;**9(Suppl 11)**:S4.
52. Jessop DM, Adams SE, Willighagen EL, *et al*. OSCAR4: a flexible architecture for chemical text-mining. *J Cheminformatics* 2011;**3**:41.
53. Cohen KB, Johnson HL, Verspoor K, *et al*. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics* 2010;**11**: 492.
54. Hawizy L, Jessop DM, Adams N, *et al*. ChemicalTagger: a tool for semantic text-mining in chemistry. *J Cheminformatics* 2011;**3**:17.
55. Hettne KM, Stierum RH, Schuemie MJ, *et al*. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* 2009;**25**:2983–91.
56. Levin MA, Krol M, Doshi AM, *et al*. Extraction and mapping of drug names from free text to a standardized nomenclature. In: *AMIA 2007—Proceedings of the American Medical Informatics Association Annual Symposium*. 2007; 438–42. http://www.ncbi.nlm.nih.gov/pmc/issues/177326/ (20 Dec 2011, date last accessed).
57. Xu H, Stenner SP, Doan S, *et al*. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;**17**:19–24.
58. Jagannathan V, Mullett CJ, Arbogast JG, *et al*. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *Int J Med Inform* 2009; **78**:284–91.
59. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**: 514–8.
60. Doan S, Bastarache L, Klimkowski S, *et al*. Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inform Assoc* 2010; **17**:528–31.
61. Patrick J, Min L. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* 2010;**17**:524–7.
62. Halgrim SR, Xia F, Solti I, *et al*. A cascade of classifiers for extracting medication information from discharge summaries. *J Biomed Semantics* 2011;**2(Suppl 3)**:S2.
63. Mille F, Degoulet P, Jaulent MC. Modeling and acquisition of drug-drug interaction. In: *MedInfo 2007–Proceedings of the 12th World Congress on Health (Medical) Informatics: Building Sustainable Health Systems*. 2007. Amsterdam, Berlin, Oxford, etc.: IOS Press;900–4. Studies in Health Technology and Informatics, Vol. 129.
64. Segura-Bedmar I, Martinez P, Sanchéz-Cisneros D. The 1st DDIExtraction-2011 challenge task: extraction of drug-drug interactions from biomedical texts. In: *DDIExtraction 2011—Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction*. 2011;1–9. http://ceur-ws.org/Vol-761/proceedings.pdf (20 Dec 2011, date last accessed).
65. Thomas P, Neves M, Solt I, *et al*. Relation extraction for drug-drug interactions using ensemble learning. In: *DDIExtraction 2011—Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction*. 2011; 11–18. http://ceur-ws.org/Vol-761/proceedings.pdf (20 Dec 2011, date last accessed).
66. Segura-Bedmar I, Martinez P, de Pablo-Sanchéz C. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC Bioinformatics* 2011; **12(Suppl 2)**:S1.
67. Segura-Bedmar I, Martinez P, de Pablo-Sanchéz C. Extracting drug-drug interactions from biomedical texts. *BMC Bioinformatics* 2010;**11(Suppl 5)**:P9.
68. Segura-Bedmar I, Martinez P, de Pablo-Sanchéz C. Using a shallow linguistic kernel for drug-drug interaction extraction. *J Biomed Inform* 2011;**44**:789–804.
69. Tari L, Anwar S, Liang S, *et al*. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics* 2010;**26**: i547–53.
70. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. In: *PSB 2012—Proceedings of the Pacific Symposium on Biocomputing*. 2012;410–21. http://psb.stanford.edu/psb-online/proceedings/psb12/ (10 Jan 2012, date last accessed).
71. Coulet A, Shah NH, Garten Y, *et al*. Using text to build semantic networks for pharmacogenomics. *J Biomed Inform* 2010;**43**:1009–19.
72. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. *Nat Biotechnol* 2006;**24**:55–62.
73. van Driel MA, Bruggeman J, Vriend G, *et al*. A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;**14**: 535–42.
74. Gonzalez G, Uribe JC, Tari L, *et al*. Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. In: *PSB 2007—Proceedings of the Pacific Symposium on Biocomputing*. 2007;28–39. http://psb.stanford.edu/psb-online/proceedings/psb07/ (20 Dec 2011, date last accessed).
75. Ahmed ST, Chidambaram D, Davulcu H, *et al*. IntEx: a syntactic role driven protein-protein interaction extractor for bio-medical text. In: *BioLink 2005: Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*. 2005. New Brunswick (NJ): Association for Computational Linguistics (ACL);54–61.
76. Chen JY, Shen C, Sivachenko AY. Mining Alzheimer disease relevant proteins from integrated protein Interactome data. In: *PSB 2006—Proceedings of the Pacific Symposium on Biocomputing*. 2006;367–78. http://psb.stanford.edu/psb-online/proceedings/psb06/ (20 Dec 2011, date last accessed).
77. Bundschus M, Dejori M, Stetter M, *et al*. Extraction of semantic biomedical relations from text using Conditional Random Fields. *BMC Bionformatics* 2008;**9**:207.
78. Tiffin N, Kelso JF, Powell AR, *et al*. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res* 2005;**33**:1544–52.
79. Coulet A, Smaïl-Tabbone M, Benlian P, *et al*. Ontology-guided data preparation for discovering

genotype-phenotype relationships. *BMC Bioinformatics* 2008;**9(Suppl 4)**:S3.

80. Chun HW, Tsuruoka Y, Kim JD, *et al*. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. In: *PSB 2006—Proceedings of the Pacific Symposium on Biocomputing*. 2006;4–15. http://psb.stanford.edu/psb-online/proceedings/psb06/ (20 Dec 2011, date last accessed).
81. Masseroli M, Kilicoglu H, Lang FM, *et al*. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. *BMC Bioinformatics* 2006;**7**:291.
82. Rindflesch TC, Libbus B, Hristovski D, *et al*. Semantic relations asserting the etiology of genetic diseases. In: *AMIA 2003—Proceedings of the American Medical Informatics Association Annual Symposium*. 2003;554–8. http://www.ncbi.nlm.nih.gov/pmc/issues/131751/ (20 Dec 2011, date last accessed).
83. Kullo IJ, Fan J, Pathak J, *et al*. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc* 2010;**17**:568–74.
84. Chang JT, Altman RB. Extracting and characterizing gene-drug relationships from the literature. *Pharmacogenetics* 2004;**14**:577–86.
85. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics* 2009; **10(Suppl 2)**:S6.
86. Müller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2004;**2**:E309.
87. Crasto C, Luo D, Yu F, *et al*. GenDrux: a biomedical literature search system to identify gene expression-based drug sensitivity in breast cancer. *BMC Med Inform and Decision Making* 2011;**11**:28.
88. Kuhn M, von Mering C, Campillos M, *et al*. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res* 2008;**36**:D684–8.
89. Saric J, Jensen LJ, Ouzounova R, *et al*. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics* 2006;**22**:645–50.
90. Xu H, Jiang M, Oetjens M, *et al*. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of Warfarin. *J Am Med Inform Assoc* 2011;**18**:387–91.
91. Garten Y, Tatonetti NP, Altman RB. Improving the recognition of pharmacogenes using text-derived drug-gene relationships. In: *PSB 2010—Proceedings of the Pacific Symposium on Biocomputing*. 2010;305–14. http://psb.stanford.edu/psb-online/proceedings/psb10/ (20 Dec 2011, date last accessed).
92. Hansen NT, Brunak S, Altman RB. Generating genome-scale candidate gene lists for pharmacogenomics. *Clin Pharmacol Ther* 2009;**86**:183–9.
93. Harpaz R, Haerian K, Chase HS, *et al*. Mining electronic health records for adverse drug effects using regression based methods. In: *ACM IHI 2010—Proceedings of the 1st ACM International Health Informatics Symposium*. 2010. New York (NY): Association for Computing Machinery (ACM);100–7.
94. Warrer P, Holme Hansen E, Juhl-Jensen L, *et al*. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *Brit J Clin Pharmacol*. doi: 10.1111/j.1365-2125.2011.04153.x.
95. Wang X, Chase HS, Li H, *et al*. Integrating heterogeneous knowledge sources to acquire executable drug-related knowledge. In: *AMIA 2010—Proceedings of the American Medical Informatics Association Annual Symposium*. 2010;852–6. http://www.ncbi.nlm.nih.gov/pmc/issues/194363/ (20 Dec 2011, date last accessed).
96. Uzuner Ö, South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**:552–6.
97. Doğan RI, Névéol A, Lu Z. A context-blocks model for identifying clinical relationships in patient records. *BMC Bioinformatics* 2011;**12(Suppl 3)**:S3.
98. Chen ES, Hripcsak G, Xu H, *et al*. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc* 2008; **15**:87–98.
99. Friedman C, Alderson PO, Austin J, *et al*. A general natural language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;**1**:161–74.
100. Friedman C. Semantic text parsing for patient records. In: Chun H, Fuller SS, Friedman C, *et al*, (eds). *Knowledge Management and Data Mining in Biomedicine*. New York (NY): Springer, 2005;423–48.
101. Lussier Y, Borlawsky T, Rappaport D, *et al*. PhenoGO: assigning phenotypic context to gene ontology annotation with natural language processing. In: *PSB 2006—Proceedings of the Pacific Symposium on Biocomputing*. 2006;64–75. http://psb.stanford.edu/psb-online/proceedings/psb06/ (20 Dec 2011, date last accessed).
102. Gysbers M, Reichley R, Kilbridge PM, *et al*. Natural language processing to identify adverse drug events. In: *AMIA 2008— Proceedings of the American Medical Informatics Association Annual Symposium*. 2008;961. http://www.ncbi.nlm.nih.gov/pmc/issues/177327/ (20 Dec 2011, date last accessed).
103. Carrell D, Miglioretti D, Smith-Bindman R. Coding free text radiology reports using the Cancer Text Information Extraction System (caTIES). In: *AMIA 2007—Proceedings of the American Medical Informatics Association Annual Symposium*. 2007;889. http://www.ncbi.nlm.nih.gov/pmc/issues/177326/ (20 Dec 2011, date last accessed).
104. Wang X, Hripcsak G, Markatou M, *et al*. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc* 2009;**16**:328–37.
105. Wang X, Chase H, Markatou M, *et al*. Selecting information in electronic health records. *J Biomed Inform* 2010;**43**:595–601.
106. Aramaki E, Miura Y, Tonoike M, *et al*. Extraction of adverse drug effects from clinical records. In: *MedInfo 2010—Proceedings of the 13th World Congress on Medical and Health Informatics*. 2010. Amsterdam, Berlin, Oxford, etc.: IOS Press;739–43. Studies in Health Technology and Informatics, Vol. 160.
107. Leaman R, Wojtulewicz L, Sullivan R, *et al*. Towards Internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: *BioNLP 2010—Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. 2010.

Stroudsburg (PA): Association for Computational Linguistics (ACL);117–25.

108. Gurulingappa H, Fluck J, Hofmann-Apitius M, *et al*. Identification of adverse drug event assertive sentences in medical case reports. In: *KDHCM 2011—Proceedings of the 1st International Workshop on Knowledge Discovery in Health Care and Medicine*. 2011;16–27. http://www.cs.gmu.edu/~hrangwal/kd-hcm/proc/KDHCM11_procs.pdf (20 Dec 2011, date last accessed).
109. Kuhn M, Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;**6**:343.
110. Rindflesch TC, Tanabe L, Weinstein JN, *et al*. (2000). Edgar: extraction of drugs, genes and relations from the biomedical literature. In: *PSB 2000—Proceedings of the Pacific Symposium on Biocomputing*. 2000;514–25. http://psb.stanford.edu/psb-online/proceedings/psb00/ (20 Dec 2011, date last accessed).
111. Ahlers CB, Fiszman M, Demner-Fushman D, *et al*. Extracting semantic predications from Medline citations for pharmacogenomics. In: *PSB 2007—Proceedings of the Pacific Symposium on Biocomputing*. 2007;209–20. http://psb.stanford.edu/psb-online/proceedings/psb07/ (20 Dec 2011, date last accessed).
112. Roberts A, Gaizauskas R, Hepple M, *et al*. Mining clinical relationships from patient narratives. *BMC Bioinformatics* 2008;**9(Suppl 11)**:S3.
113. Li J, Zhu X, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Comput Biol* 2009;**5**:e1000450.
114. Hakenberg J, Voronov D, Nguyen VH, *et al*. Taking a SNPshot of PubMed: a repository of genetic variants and their drug response phenotypes. In: *Proceedings of the GPD-Rxn Workshop: Genotype-Phenotype-Drug Relationship Extraction from Text, Pacific Symposium on Biocomputing*. 2010. http://psb.stanford.edu/psb10/gpdrxn-workshop.html (20 Dec 2011, date last accessed).
115. Tari L, Anwar S, Liang S, *et al*. Synthesis of pharmacokinetic pathways through knowledge acquisition and automated reasoning. In: *PSB 2010 – Proceedings of the Pacific Symposium on Biocomputing*. 2010;465–76. http://psb.stanford.edu/psb-online/proceedings/psb10/ (20 Dec 2011, date last accessed).
116. Frijters R, van Vugt M, Smeets R, *et al*. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol* 2010;**6**:e1000943.
117. Baker NC, Hemminger BM. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J Biomed Inform* 2010;**43**:510–9.
118. Bekhuis T. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomed Digit Libr* 2006;**3**:2.
119. Srinivasan P. Text mining: generating hypotheses from Medline. *J Am Soc for Information Science and Technology* 2004;**55**:396–413.
120. Srinivasan P, Libbus B. Mining Medline for implicit links between dietary substances and diseases. *Bioinformatics* 2004;**20(Suppl 1)**:i290–6.
121. Smalheiser NR. Literature-based discovery: beyond the ABCs. *Journal of the American Society for Information Science and Technology* 2011;**63**:218–24.
122. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. *J Biomed Inform* 2009;**42**:633–43.
123. Kim JD, Kraines S, Guo W, *et al*. Inference for Bio-IE: Genia meets EKOSS. In: *LBM 2009—Proceedings of the 3rd International Symposium on Languages in Biology and Medicine*. 2009;126–8. http://lbm2009.biopathway.org/ (20 Dec 2011, date last accessed).
124. Thagard P, Shelley C. Abductive reasoning: logic, visual thinking, and coherence. In: Chiara MLD, Doets K, Mundici D, van Benthem J, (eds). *Logic and Scientific Methods*. Dordrecht: Kluwer, 1997;413–27.
125. Taboda M, Lalin R, Martinez D. An automated approach to mapping external terminologies to the UMLS. *IEEE Trans Biomed Eng* 2009;**56**:1598–605.
126. Zhou L, Plasek JM, Mahoney LM. Using medical text extraction, reasoning and mapping system (MTERMS) to process medication information in outpatient clinical notes. In: *AMIA 2011—Proceedings of the American Medical Informatics Association Annual Symposium*. 2011;1639–48. http://www.ncbi.nlm.nih.gov/pmc/issues/203717/ (20 Dec 2011, date last accessed).
127. Alpaydin E. *Introduction to Machine Learning*. 2nd edn. Cambridge (MA): MIT Press, 2010.
128. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edn. Burlington (MA): Morgan Kaufmann, 2011.
129. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;**18**:544–51.
130. Rebholz-Schuhmann D, Yepes AJ, van Mulligen EM, *et al*. The CALBC Silver Standard Corpus: harmonizing multiple semantic annotations in a large biomedical corpus. *J Bioinf Comput Biol* 2010;**8**:163–79.
131. Rebholz-Schuhmann D, Yepes AJ, Li C, *et al*. Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *J Biomed Semantics* 2011;**2(Suppl 5)**:S11.
132. Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics* 2006;**7**:356.
133. Ogren PV. Knowtator: a Protégé plug-in for annotated corpus construction. In: *HLT 2006—Proceedings of the 2006 Human Language Technology Conference of the NAACL, Companion Volume*. 2006. Stroudsburg (PA): Association for Computational Linguistics (ACL);273–5.
134. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. *Comp Linguistics* 2008;**34**:555–96.
135. Carletta JC. Assessing agreement on classification tasks: the kappa statistic. *Comp Linguistics* 1996;**22**:249–54.
136. Blackman NJM, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. *Stat Med* 2000;**19**:723–41.
137. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: *ECIR 2005—Proceedings of the 27th European Conference on Information Retrieval*. 2005. Heidelberg, Berlin: Springer; 345–59 (LNCS 3408).
138. Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. In: *ISMB'99—Proceedings of the International Conference on Intelligent Systems for Molecular Biology*. 1999. Menlo Park (CA): AAAI Press;77–86.

139. Tomanek K, Wermter J, Hahn U. An approach to text corpus construction which cuts annotation costs and maintains corpus reusability of annotated data. In: *EMNLP–CoNLL 2007—Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007. Stroudsburg (PA): Association for Computational Linguistics (ACL);486–95.
140. Pyysalo S, Airola A, Heimonen J, *et al*. A comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics* 2008;**9(Suppl 3)**:S6.
141. Kim JD, Ohta T, Tsujii Ji. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 2008; **9**:10.
142. Thompson P, Iqbal SA, McNaught J, *et al*. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics* 2009;**10**:349.
143. Buyko E, Beisswanger E, Hahn U. The GeneReg corpus for gene expression regulation events. An overview of the corpus and its in-domain and out-of-domain interoperability. In: *LREC 2010—Proceedings of the 7th International Conference on Language Resources and Evaluation*. 2010. Paris: European Language Resources Association (ELRA);2662–6.
144. Carreira R, Carneiro S, Pereira R, *et al*. Semantic annotation of biological concepts interplaying microbial cellular responses. *BMC Bioinformatics* 2011;**12**:460.
145. Roberts A, Gaizauskas R, Hepple M, *et al*. Semantic annotation of clinical text: the CLEF corpus. In: *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*. 2008. Paris: European Language Resources Association (ELRA);19–26.
146. Roberts A, Gaizauskas RJ, Hepple M, *et al*. Building a semantically annotated corpus of clinical texts. *J Biomed Inform* 2009;**42**:950–66.
147. Roberts A, Gaizauskas R, Hepple M, *et al*. The CLEF corpus: semantic annotation of clinical text. In: *AMIA 2007—Proceedings of the American Medical Informatics Association Annual Symposium*. 2007;625–9. http://www.ncbi.nlm.nih.gov/pmc/issues/177326/ (20 Dec 2011, date last accessed).
148. Pustejovsky J, Ingria R, Sauri R, *et al*. The specification language TimeML. In: Mani I, Pustejovsky J, Gaizauskas R, (eds). *The Language of Time: A Reader*. Oxford (U.K.): Oxford University Press, 2005;545–57.
149. Ogren PV, Savova GK, Chute CG. Constructing evaluation corpora for automated clinical named entity recognition. In: *LREC 2008—Proceedings of the 6th International Conference on Language Resources and Evaluation*. 2008. Paris: European Language Resources Association (ELRA);3143–50.
150. Pestian JP, Brew C, Matykiewicz PM, *et al*. A shared task involving multi-label classification of clinical free text. In: *BioNLP 2007—Proceedings of ACL 2007 Workshop on Biological, Translational, and Clinical Language Processing*. 2007. Stroudsburg (PA): Association for Computational Linguistics (ACL);97–104.
151. Leaman R, Gonzalez G, Miller C. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. In: *LBM 2009—Proceedings of the 3rd International Symposium on Languages in Biology and Medicine*. 2009;82–9. http://lbm2009.biopathway.org/ (20 Dec 2011, date last accessed).
152. Cano C, Monaghan T, Blanco A, *et al*. Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *J Biomed Inform* 2009;**42**: 967–77.
153. Kulick S, Bies A, Liberman M, *et al*. Integrated annotation for biomedical information extraction. In: *BioLINK 2004—Proceedings of the HLT-NAACL 2004 Workshop on Linking Biological Literature, Ontologies and Databases*. 2004. Stroudsburg (PA): Association for Computational Linguistics (ACL);61–8.
154. Uzuner Ö, Goldstein I, Luo Y, *et al*. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008;**15**:15–24.
155. Uzuner Ö. Recognizing obesity and co-morbidities in sparse data. *J Am Med Inform Assoc* 2009;**16**:561–70.
156. Clark C, Good K, Jezierny L, *et al*. Identifying smokers with a medical extraction system. *J Am Med Inform Assoc* 2008;**15**:36–9.
157. de Bruijn B, Cherry C, Kiritchenko S, *et al*. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 2011;**18**:557–62.
158. Chapman WW, Nadkarni PM, Hirschman L, *et al*. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011;**18**:540–3.
159. Corbett P, Batchelor C, Teufel S. Annotation of chemical named entities. In: *BioNLP 2007—Proceedings of the ACL 2007 Workshop on Biological, Translational, and Clinical Language Processing*. 2007. Stroudsburg (PA): Association for Computational Linguistics (ACL);57–64.
160. Kolárik C, Klinger R, Friedrich CM, *et al*. Chemical names: terminological resources and corpora annotation. In: *BioTxtM 2008—Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*. 2008. Paris: European Language Resources Association (ELRA);51–8.
161. Gurulingappa H, Klinger R, Hofmann-Apitius M, *et al*. An empirical evaluation of resources for the identification of diseases and adverse effects in biomedical literature. In: *BioTxtM 2010—Proceedings of the 2nd LREC 2010 Workshop on Building and Evaluating Resources for Biomedical Text Mining*. 2010. Paris: European Language Resources Association (ELRA);15–22.
162. Thorn CF, Klein TE, Altman RB. Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics* 2010;**11**: 501–5.
163. McDonagh EM, Whirl-Carrillo M, Garten Y, *et al*. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark Med* 2011;**5**:795–806.
164. Buyko E, Beisswanger E, Hahn U. The extraction of pharmacogenetic and pharmacogenomic reations: a case study using PharmGKB. In: *PSB 2012—Proceedings of the Pacific Symposium on Biocomputation*. 2012;376–87. http://psb.stanford.edu/psb-online/proceedings/psb12/ (10 Jan 2012, date last accessed).
165. Verspoor K, Cohen KB, Lanfranchi A, *et al*. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*. (to appear).
166. Nelson SJ, Zeng K, Kilbourne J, *et al*. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* 2011;**18**:441–8.

167. de Matos P, Alcántara R, Dekker A, *et al*. Chemical entities of biological interest: an update. *Nucleic Acids Res* 2010;**38**: D249–54.

168. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *J Nat Lang Engineering* 2004;**10**:327–48.

169. Hahn U, Buyko E, Landefeld R, *et al*. An overview of JCoRe, the JULIE Lab UIMA component repository. In: *Proceedings of the LREC'08 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP'*. 2008. Paris: European Language Resources Association (ELRA); 1–7.

170. Buyko E, Hahn U. Fully embedded type systems for the semantic annotation layer. In: *ICGL 2008—Proceedings of the 1st International Conf on Global Interoperability for Language Resources*. 2008. Hong Kong: SAR;26–33.

171. Kano Y, Baumgartner WA Jr, McCrohon L, *et al*. U-Compare: share and compare text mining tools with UIMA. *Bioinformatics* 2009;**25**:1997–8.

172. Kano Y, Björne J, Ginter F, *et al*. U-Compare bio-event meta-service: compatible BioNLP event extraction services. *BMC Bioinformatics* 2011;**12**:481.

173. Jonquet C, Musen MA, Shah NH. Building a biomedical Ontology Recommender Web service. *J Biomed Semantics* 2010;**1(Suppl 1)**:S1.

174. Parai GK, Jonquet C, Xu R, *et al*. The Lexicon Builder Web service: building custom lexicons from two hundred biomedical ontologies. In: *AMIA 2010 – Proceedings of the American Medical Informatics Association Annual Symposium*. 2010;587–91. http://www.ncbi.nlm.nih.gov/pmc/issues/194363/ (20 Dec 2011, date last accessed).

175. Thompson P, McNaught J, Montemagni S, *et al*. The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics* 2011;**12**: 397.

176. Jonquet C, Shah NH, Musen MA. The Open Biomedical Annotator. In: *Proceedings of the AMIA Joint Summit on Translational Bioinformatics*. 2009;56–60. http://www.ncbi.nlm.nih.gov/pmc/issues/194372/ (20 Dec 2011, date last accessed).

177. Coulet A, Smaïl-Tabbone M, Napoli A, *et al*. Ontology-based knowledge discovery in pharmacogenomics. In: Arabnia HR, Tran QN, (eds). *Software Tools and Algorithms for Biological Systems*. Berlin, Heidelberg: Springer, 2010;357–66.

178. Coulet A, Garten Y, Dumontier M, *et al*. Integration and publication of heterogeneous text-mined relationships on the Semantic Web. *J Biomed Semantics* 2011; **2(Suppl 2)**:S10.

179. Dumontier M, Villanueva-Rosales N. Towards pharmacogenomics knowledge discovery with the Semantic Web. *Brief Bioinformatics* 2009;**10**:153–63.

180. Groth P, Gibson A, Velterop J. The anatomy of a nanopublication. *Informat Serv Use* 2010;**30**:51–6.

181. Belleau F, Nolin MA, Tourigny N, *et al*. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008;**41**:706–16.

182. Chen B, Dong X, Jiao D, *et al*. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 2010;**11**:255.

183. Samwald M, Jentsch A, Bouton C, *et al*. Linked open drug data for pharmaceutical research and development. *J Cheminformatics* 2011;**3**:19.

184. Tipney HJ, Schuyler RP, Hunter L. Consistent visualizations of changing knowledge. In: *Proceedings of the AMIA Joint Summit on Translation in Bioinformatics*. 2009;129–32. http://www.ncbi.nlm.nih.gov/pmc/issues/194372/ (20 Dec 2011, date last accessed).

185. Leach SM, Tipney H, Feng W, *et al*. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput Biol* 2009;**5**:e1000215.

186. Mukhopadhyay S, Palakal M, Maddu K. Multi-way association extraction and visualization from biological text documents using hyper-graphs: applications to genetic association studies for diseases. *Artif Intell Med* 2010;**49**: 145–54.

187. Ciccarese P, Ocana M, Garcia Castro LJ, *et al*. An open annotation ontology for science on Web 3.0. *J Biomed Semantics* 2011;**2(Suppl 2)**:S4.

188. Deftereos SN, Andronis C, Friedla EJ, *et al*. Drug repurposing and adverse event prediction using high-throughput literature analysis. *Wiley Interdiscip Rev Syst Biol Med* 2011; **3**:323–34.

189. Andronis C, Sharma A, Virvilis V, *et al*. Literature mining, ontologies and information visualization for drug repurposing. *Brief Bioinformatics* 2011;**12**:357–68.

190. Agarwal P, Searls DB. Literature mining in support of drug discovery. *Brief Bioinformatics* 2008;**9**:479–92.

191. Plake C, Schroeder M. Computational polypharmacology with text mining and ontologies. *Curr Pharm Biotechnol* 2011;**12**:449–57.

192. Winnenburg R, Wächter T, Plake C, *et al*. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief Bioinformatics* 2008;**9**:466–78.

193. Wiegers TC, Davis AP, Cohen KB, *et al*. Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). *BMC Bioinformatics* 2009;**10**:326.

194. Dowell KG, McAndrews-Hill MS, Hill DP, *et al*. Integrating text mining into the MGI biocuration workflow. *Database* 2009;**2009**:bap019.

195. Arighi CN, Roberts PM, Agarwal S, *et al*. BioCreAtIvE III interactive task: an overview. *BMC Bioinformatics* 2011; **12(Suppl 8)**:S4.

196. Clematide S, Rinaldi F. Ranking interactions for a curation task. In: *ICMLA 2011-Proceedings of the 10th IEEE International Conference on Machine Learning and Applications*. 2011. Los Alamitos (CA): IEEE Computer Society Press;100–5.

197. Baumgartner WA Jr, Cohen KB, Fox LM, *et al*. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 2007;**23**:i41–8.

198. Ceol A, Chatr-Aryamontri A, Licata L, *et al*. Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS Lett* 2008;**582**:1171–7.

199. Tomanek K, Wermter J, Hahn U. A reappraisal of sentence and token splitting for life sciences documents. In: *MedInfo 2007 – Proceedings of the 12th World Congress on Health (Medical) Informatics – Building Sustainable Health Systems*. 2007. Amsterdam, Berlin, Oxford, etc.: IOS Press;524–8. Studies in Health Technology and Informatics, Vol. 129.

200. D'Avolio W, Nguyen TM, Goryachev S, *et al*. Automated concept-level information extraction to reduce the need for

custom software and rules development. *J Am Med Inform Assoc* 2011;**18**:607–13.

201. Cohen KB, Palmer M, Hunter LE. Nominalization and alternations in biomedical language. *PLoS ONE* 2008;**3**: e3158.
202. Blake C. Beyond genes, proteins, and abstracts: identifying scientific claims from full-text biomedical articles. *J Biomed Inform* 2010;**43**:173–89.
203. McIntosh T, Curran JR. Challenges for extracting biomedical knowledge from full text. In: *BioNLP 2007—Proceedings of the ACL 2007 Workshop on Biological, Translational, and Clinical Language Processing*. 2007. Stroudsburg (PA): Association for Computational Linguistics (ACL);171–178.
204. Gasparin C. Statistical anaphora resolution in biomedical texts. PhD thesis, Cambridge (U.K.): University of Cambridge, Computer Laboratory Report UCAM-CL-TR-764, 2009.
205. Segura-Bedmar I, Crespo M, de Pablo-Sánchez C, *et al*. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. *BMC Bioinformatics* 2010;**11(Suppl 2)**:S1.
206. Savova GK, Chapman WW, Zheng J, *et al*. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc* 2011;**18**:459–65.
207. Wang Y, Melton GN, Pakhomov S. It's about 'this' and 'that': a description of anaphoric expressions in clinical text. In: *AMIA 2011 Proceedings of the American Medical Informatics Association Annual Symposium*. 2011;1471–80. http://www.ncbi.nlm.nih.gov/pmc/issues/203717/ (20 Dec 2011, date last accessed).
208. Zheng J, Chapman WW, Crowley RS, *et al*. Methodological review: Coreference resolution: a review of general methodologies and applications in the clinical domain. *J Biomed Inform* 2011;**44**:1113–22.
209. Mizuta Y, Korhonen A, Mullen T, *et al*. Zone analysis in biology articles as a basis for information extraction. *International J Med Inform* 2006;**75**:468–87.
210. Teufel S, Siddharthan A, Batchelor CR. Towards domain-independent argumentative zoning: evidence from chemistry and computational linguistics. In: *EMNLP 2009—Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. 2009. Stroudsburg (PA): Association for Computational Linguistics (ACL); 1493–502.
211. Prasad R, McRoy S, Frid N, *et al*. The biomedical discourse relation bank. *BMC Bioinformatics* 2011;**12**:188.
212. Hearst MA, Divoli A, Guturu H, *et al*. BioText search engine: beyond abstract search. *Bioinformatics* 2007;**23**: 2196–7.
213. Coulet A, Smaïl-Tabbone N, Napoli A, *et al*. Suggested ontology for pharmacogenomics (SO-Pharm): modular construction and preliminary testing. In: *OTM 2006—Proceedings of On the Move to Meaningful Internet Systems 2006 Workshop on Knowledge Systems in Bioinformatics (KsinBIT)*. 2006. Berlin, Heidelberg: Springer;648–57 (LNCS 4277).
214. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo (CA): Morgan Kaufmann, 1988.
215. Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge (MA): MIT Press, 2009.
216. Zhou L, Hripcsak G. Temporal reasoning with medical data: a review with emphasis on medical natural language processing. *J Biomed Inform* 2007;**40**:183–202.
217. Kleinberg S, Hripcsak G. Methodological review: A review of causal inference for biomedical informatics. *J Biomed Inform* 2011;**44**:1102–12.
218. Zhou L, Melton GB, Parsons S, *et al*. A temporal constraint structure for extracting temporal information from clinical narrative. *J Biomed Inform* 2006;**39**:424–39.
219. Zhou L, Parsons S, Hripcsak G. The evaluation of a temporal reasoning system in processing clinical discharge summaries. *J Am Med Inform Assoc* 2008;**15**:99–106.
220. Hristovski D, Peterlin B, Mitchell JA, *et al*. Using literature-based discovery to identify disease candidate genes. *Int J Med Inform* 2005;**74**:289–98.
221. Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. *J Biomed Inform* 2006;**39**:600–11.
222. Iossifov I, Rodriguez-Esteban R, Mayzus I, *et al*. Looking at cerebellar malformations through text-mined interactomes of mice and humans. *PLoS Comput Biol* 2009;**5**:e1000559.
223. Tsuruoka Y, Miwa M, Hamamoto K, *et al*. Discovering and visualising indirect associations between biomedical concepts. *Bioinformatics* 2011;**27**:i111–9.