Information **2015**, *6*, 790-810; doi:10.3390/info6040790

OPEN ACCESS

information

ISSN 2078-2489

www.mdpi.com/journal/information

Review

Drug Name Recognition: Approaches and Resources

Shengyu Liu [†]**, Buzhou Tang** [†]**, Qingcai Chen * and Xiaolong Wang**

Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, 518055 Shenzhen, China; E-Mails: shengyu_liu@163.com (S.L.); tangbuzhou@gmail.com (B.T.); wangxl@insun.hit.edu.cn (X.W.)

[†] These authors contributed equally to this work.

* Author to whom correspondence should be addressed; E-Mail: qingcai.chen@gmail.com; Tel.: +86-755-2603-3475; Fax: +86-755-2603-3182.

Academic Editor: Willy Susilo

Received: 15 October 2015 / Accepted: 19 November 2015 / Published: 25 November 2015

Abstract: Drug name recognition (DNR), which seeks to recognize drug mentions in unstructured medical texts and classify them into pre-defined categories, is a fundamental task of medical information extraction, and is a key component of many medical relation extraction systems and applications. A large number of efforts have been devoted to DNR, and great progress has been made in DNR in the last several decades. We present here a comprehensive review of studies on DNR from various aspects such as the challenges of DNR, the existing approaches and resources for DNR, and possible directions.

Keywords: drug name recognition; drug information extraction; biomedical texts

1. Introduction

With the rapid development of information technology, more and more medical documents are available, which contain a great amount of medical information, such as medical entities and relations between them. In order to take full advantage of medical texts, it is necessary to extract valuable information from them. Drugs, as one type of the basic medical elements, also need to be recognized. Drug name recognition (DNR), which seeks to recognize drug mentions in unstructured medical texts and classify them into pre-defined categories, is a fundamental task of medical information extraction, and is a key component of many medical relation extraction systems (e.g., drug-drug interactions [1] and adverse drug reactions [2]) and applications (e.g., information retrieval, information management, information tracking, clinical decision support, drug discovery and drug development) [3].

Drug mentions widely exist in various types of medical texts including medical literature, electronic medical records, medical patent applications, clinical trial documents, *etc.* For different types of medical texts, a large number of efforts have been devoted to DNR in the last several decades, generating various kinds of approaches and resources. DNR has been a subtask of several public challenges in the medical domain such as the medication extraction challenge organized by the Center of Informatics for Integrating Biology and Beside (i2b2) in 2009 [4], the chemical and drug named entity recognition (CHEMDNER) challenge of the Critical Assessment of Information Extraction systems in Biology in 2013 (BioCreAtIvE IV) [5] and the drug-drug interaction (DDIExtraction) challenge in 2013 [6].

Although there are some reviews on information extraction approaches for drugs and chemical compounds [7–9], DNR is not especially discussed. In this study, we focus on DNR and present a comprehensive review including many resources, tools and approaches that are not covered by previous reviews. Moreover, we introduce the challenges and possible directions for DNR.

2. Challenges of Drug Name Recognition

DNR is a typical named entity recognition (NER) task. It is particularly challenging due to the following reasons:

- The ways of naming drugs vary greatly. For example, the drug “quetiapine” (generic name) has the brand name “Seroquel XR”, while its systematic International Union of Pure and Applied Chemistry (IUPAC) name is “2-[2-(4-dibenzo [b,f][1,4] thiazepin-11-ylpiperazin-1-yl) ethoxy] ethanol”. Furthermore, some drug names and their synonyms are the same as normal English words or phrases. For example, brand names of “oxymetazoline nasal” and “caffeine” are “Duration” and “Stay Awake”, respectively.
- The frequent occurrences of abbreviations and acronyms make it difficult to identify the concepts to which the terms refer to. For example, the abbreviation “PN” can refer to the drug “penicillin” or other concepts such as “pneumonia”, “polyarteritis nodosa” and “polyneuritis”.
- New drugs are constantly and rapidly reported in scientific publications. Moreover, drug names may be misspelled in electronic medical records such as progress notes and discharge summaries. This makes DNR systems that rely only on dictionaries of known drug names not effective.
- Drug names may contain a number of symbols mixed with common words. For example, the IUPAC name of an atypical antipsychotic is “7-{4-[4-(2,3-dichlorophenyl) piperazin-1-yl] butoxy}-3,4-dihydroquinolin-2(1H)-one”. It is difficult to determine the boundaries of such drug names in texts.
- Some drug names may correspond to non-continuous strings of text. For example, “loop diuretics” and “potassium-sparing diuretics” in the sentence “In some patients, the administration of a non-steroidal anti-inflammatory agent can reduce the diuretic, natriuretic, and antihypertensive effects of loop, potassium-sparing and thiazide diuretics”. Such examples pose great difficulties to DNR.

3. Benchmark Datasets

The development and evaluation of DNR approaches require benchmark datasets where all drug names are annotated by human experts. Benchmark datasets can be used for training machine learning-based approaches and comparing different approaches. Table 1 lists some available benchmark datasets for DNR. Some datasets in Table 1 are not developed for DNR, but drug names are annotated in them. Therefore, they can be used for DNR. For example, ADE is developed for extraction of adverse drug effects and PK, PK-DDI and DDIExtraction 2011 are developed for extraction of drug-drug interactions. Moreover, since the datasets in Table 1 are developed for different tasks, definitions of drugs vary significantly in different datasets. Datasets such as ADE, EU-ADR and DDIExtraction 2011 only define a single class of drugs, while other datasets such as PK-DDI and DDIExtraction 2013 define multiple different classes of drugs.

To evaluate the performances of DNR approaches, precision, recall and F-score of DNR approaches on the benchmark datasets are measured. Precision is the percentage of correctly recognized drug names over all recognized results by an approach. Recall is the percentage of correctly recognized drug names over all drug names annotated in the benchmark datasets. F-score is the harmonic mean of precision and recall.

Table 1. Benchmark datasets for Drug name recognition (DNR).

Dataset	Data Source	URL
ADE [10]	Medical case reports	https://sites.google.com/site/adecorpus/
PK [11]	Biomedical literature abstracts	http://rweb.compbio.iupui.edu/corpus/
PK-DDI [12]	Drug package inserts	http://purl.org/NET/nlprepository/PI-PK-DDI-Corpus
EU-ADR [13]	Biomedical literature abstracts	http://euadr.erasmusmc.nl/sda/euadr_corpus.tgz
i2b2 Medication Extraction [4]	Discharge summaries	https://www.i2b2.org/NLP/DataSets/
DrugNer [3]	Biomedical literature abstracts	http://labda.inf.uc3m.es/DrugDDI/DrugNer.html
DDIExtraction 2011 [14]	Texts selected from DrugBank	http://labda.inf.uc3m.es/ddicorpus
DDIExtraction 2013 [15]	Biomedical literature abstracts and texts selected from DrugBank	http://labda.inf.uc3m.es/ddicorpus
CHEMDNER [5]	Biomedical literature abstracts	http://www.biocreative.org/resources/biocreative-iv/chemdner-corpus/

4. General Architecture of Drug Name Recognition Systems

Many systems have been developed for DNR. Figure 1 shows the typical procedure of a DNR system. Generally, three steps are required to develop a DNR system.

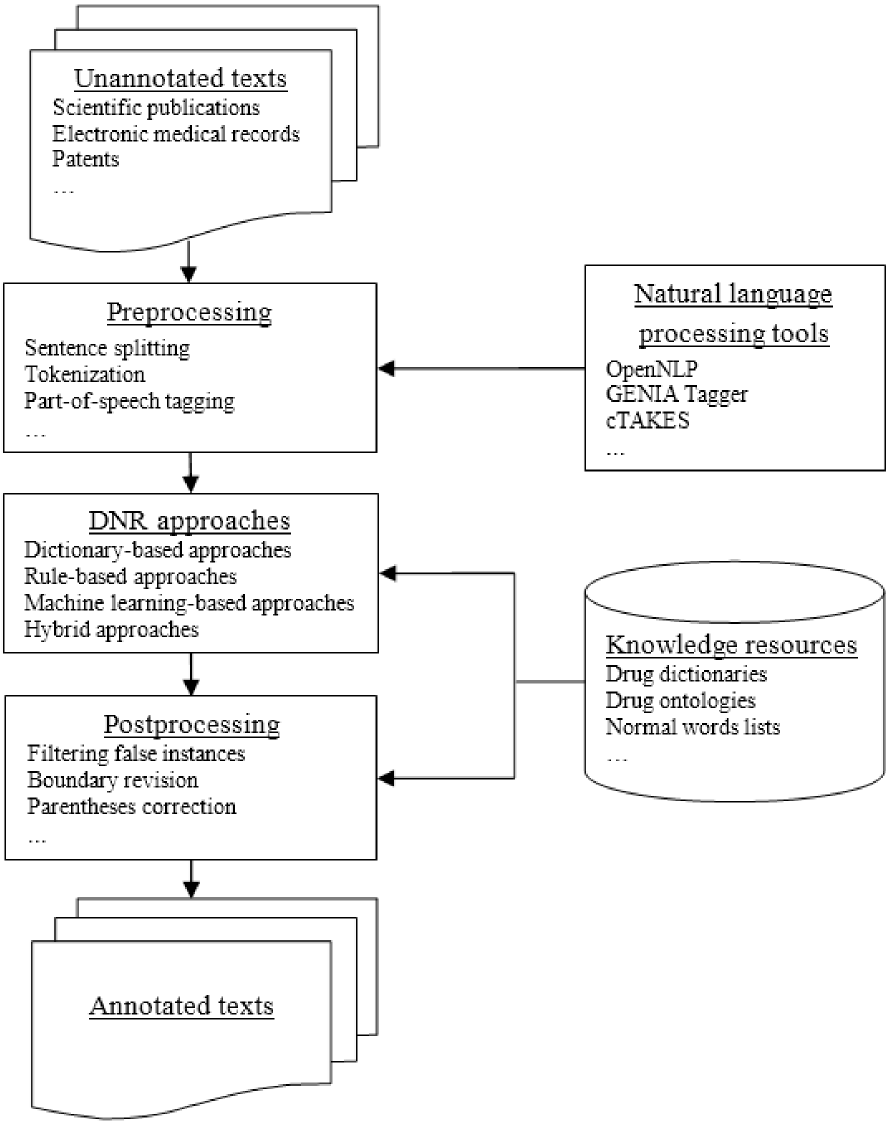

Figure 1. Typical procedure of a DNR system.

(1) Preprocessing: Preprocessing aims at transforming the original input texts into representations required by DNR approaches and enriching the original texts with lexical and syntactic information. Preprocessing includes sentence splitting, tokenization, part-of-speech (POS) tagging, text chunking, lemmatization, *etc.* The output information of preprocessing can be used to induce rules or generate features for DNR approaches. The selection of suitable strategies or methods for preprocessing has significant impact on the performances of DNR systems [16,17]. Dai *et al.* [16] investigated the effects of coarse-grained and fine-grained tokenization strategies on DNR. For the coarse-grained tokenization, Penn Treebank tokenization rules [18] were used. The fined-grained tokenization strategy applied some extra preprocessing steps on the generated tokens of coarse-grained tokenization, e.g., adding separations before and after symbols such as hyphens and dashes. It was demonstrated that fine-grained tokenization performed better than coarse-grained tokenization. Batista-Navarro *et al.* [17] focused on the effects of sentence splitting and tokenization on recognition of drugs and chemicals from chemical literature. They compared non-specialized implementations of sentence splitting and tokenization with specialized implementations tuned for chemical literature. Specialized implementations achieved better performance than non-specialized implementations.

There are many open source natural language processing (NLP) toolkits that can be used for preprocessing in DNR systems. Table 2 lists some commonly used NLP toolkits. For each preprocessing task, NLP tools based on different methods and tuned for different types of texts are available. It is important to select appropriate NLP tools to preprocess texts of different fields. The unstructured information management architecture (UIMA) [19,20] makes comparing and selecting NLP tools simple. Based on UIMA, it is easy to plug a NLP tool into existing text processing pipelines or combine NLP tools into text processing pipelines. For example, U-Compare [21,22] is an integrated NLP systems based on UIMA. It provides a large collection of NLP tools and allows sets of tools to be run in parallel on the same inputs. Moreover, it can automatically generate statistics for all possible combinations of these tools.

Table 2. Open source natural language processing (NLP) toolkits for preprocessing in DNR systems.

NLP Toolkit	Target Domain	URL
OpenNLP	General	http://opennlp.apache.org
LingPipe	General	http://alias-i.com/lingpipe
NLTK	General	http://www.nltk.org
ANNIE	General	https://gate.ac.uk/sale/tao/splitch6.html#chap:annie
NaCTeM	General	http://www.nactem.ac.uk/software.php
Stanford NLP Toolkit [23]	General	http://www-nlp.stanford.edu/software/
U-Compare [21,22]	General	http://u-compare.org
JULIE Lab [24]	General	http://www.julielab.de/Resources
GENIA Tagger [25]	Biomedical	http://www.nactem.ac.uk/GENIA/tagger/
GDep [26]	Biomedical	http://people.ict.usc.edu/~sagae/parser/gdep/
Neji [27]	Biomedical	http://bioinformatics.ua.pt/neji/
BioLemmatizer [28]	Biomedical	http://biolemmatizer.sourceforge.net/
cTAKES [29]	Clinical	http://ctakes.apache.org/

(2) Drug name recognition: This step recognizes drug names from unstructured texts and classifies them into predefined categories. Knowledge resources play important roles in DNR approaches. They can be used to match drug names, induce rules and generate features for DNR approaches. DNR approaches and knowledge resources will be introduced in detail in the following sections.

(3) Postprocessing: In the postprocessing step, heuristic rules and knowledge resources are commonly used to refine the recognition results of DNR approaches [30–32]. For instance, Grego *et al.* [30] filtered the recognized drug names composed entirely by digits and removed characters such as "*", "–" and "." from recognized drug names if the characters appear at the end of recognized drug names. Leaman *et al.* [31] regarded a mention of drug or chemical with unbalanced parenthesis as an error. They balanced the parenthesis by adding or removing one character at the right or left of the mention. Grego *et al.* [32] calculate the semantic similarities between drugs identified by a DNR system in a given text window based on semantic relationships in a drug knowledge base. They assign a single validation score to each identified drug based on its similarities to other drugs and then filter falsely identified drugs using a given threshold to increase precision of the DNR system.

5. Approaches for Drug Name Recognition

Approaches for DNR can be classified into four categories: dictionary-based, rule-based, machine learning-based and hybrid approaches.

5.1. Dictionary-Based Approaches

Drug dictionaries refer to collections of drug names. They can be constructed manually or automatically from publicly available knowledge resources such as databases and ontologies containing synonyms or spelling variants of drug names. Different knowledge resources contain different terms. Some knowledge resources focus on drugs, while others focus on general chemicals. Therefore, drug dictionaries are usually constructed by merging several knowledge resources. Before reviewing the dictionary-based approaches, we introduce some freely available knowledge resources and describe how to construct drug dictionaries from them. The web-accessible URLs of the knowledge sources are listed in Table 3.

Table 3. Available knowledge resources for constructing drug dictionaries.

Knowledge Resource	URL
DrugBank	http://www.drugbank.ca/
KEGG DRUG	http://www.kegg.jp/kegg/drug/
PharmGKB	http://www.pharmgkb.org/
CTD	http://ctdbase.org/
RxNorm	http://www.nlm.nih.gov/research/umls/rxnorm/
RxTerms	http://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm
Drugs@FDA	http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm
TTD	http://bidd.nus.edu.sg/group/ttd/ttd.asp
ChEBI	http://www.ebi.ac.uk/chebi
MeSH	http://www.nlm.nih.gov/mesh/
PubChem	https://pubchem.ncbi.nlm.nih.gov/
UMLS Metathesaurus	http://www.nlm.nih.gov/research/umls/
Jochem	http://www.biosemantics.org/index.php/resources/jochem

DrugBank is an online database that contains chemical, pharmacological and pharmaceutical information about drugs and comprehensive drug target information [33]. Fields such as “name”, “synonyms” and “international-brands” in DrugBank can be extracted to build a drug name dictionary.

Kyoto Encyclopedia of Genes and Genomes (KEGG) DRUG is a drug information resource for approved drugs in Japan, USA and Europe [34]. The “Name” field in KEGG DRUG can be used for the creation of drug name dictionary.

Pharmacogenomics Knowledgebase (PharmGKB) is a comprehensive resource that curates knowledge about the impact of genetic variation on drug response [35]. It provides a drug list and the “name”, “generic names” and “trade names” fields in the drug list can be collected to construct a drug name dictionary.

Comparative Toxicogenomics Database (CTD) is a publicly available database that provides manually curated information about chemical-gene interactions, chemical-disease and gene-disease relationships [36]. The “ChemicalName” field can be extracted to build a drug name dictionary.

RxNorm is a standardized nomenclature for clinical drugs [37]. It is created by the United States National Library of Medicine (NLM) to let various systems using different drug nomenclatures share and exchange data efficiently. The “ingredient (IN)” and “brand name (BN)” fields can be extracted to build a drug name dictionary.

RxTerms is a drug interface terminology derived from RxNorm for prescription writing or medication history recording [38]. The “FULL_GENERIC_NAME”, “BRAND_NAME” and “DISPLAY_NAME” fields can be used to build a drug name dictionary.

Drugs@FDA is provided by the United States Food and Drug Administration (FDA). It contains information about FDA-approved drug names, generic prescription, over-the-counter human drugs, *etc.* Drug names can be extracted from the “drug name” and “activeingred” fields of Drugs@FDA.

Therapeutic Targets Database (TTD) is a database that provides information about therapeutic targets and corresponding drugs [39]. It contains many drugs including approved, clinical trial and experimental drugs. The “Name”, “Synonyms” and “Trade Name” fields in TTD can be collected to build a drug dictionary.

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities [40]. In addition, it incorporates an ontological classification, whereby the relationships between molecular entities, classes of entities and their parents, children and siblings are specified. The fields such as “ChEBI name”, “International Nonproprietary Name (INN)” and “Synonyms” can be extracted for dictionary creation. Moreover, the class information in ChEBI ontology can be used to classify drugs.

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus from NLM. It consists of sets of terms named descriptors [41]. MeSH descriptors are used for indexing, cataloging and searching for biomedical and health-related information and documents. MeSH descriptors are divided into 16 categories. Each category is further divided into some subcategories. Within each subcategory, descriptors are organized in a hierarchical structure. The category “D” covers drugs and chemicals. Terms belonging to category “D” can be extracted to build a drug dictionary.

PubChem is a public repository for biological properties of small molecules [42]. It consists of three interconnected databases: PubChem Substance, PubChem Compound and PubChem BioAssay. PubChem Substance contains entries of mixtures, extracts, complexes and uncharacterized substances and provides synonyms of the substances. PubChem Compound is a subset of PubChem Substance. It contains pure and characterized chemical compounds but no synonyms. In order to build a high quality dictionary consisting of as many synonyms as possible, names and synonyms of PubChem Substance entries that have links to PubChem Compound entries are usually collected.

Unified Medical Language System (UMLS) Metathesaurus is a large, multi-purpose, and multi-lingual thesaurus that contains millions of biomedical and health related concepts from over 100 vocabularies, their synonymous names and relationships among them [43]. Each concept in UMLS Metathesaurus is assigned to at least one semantic type. The concepts in the UMLS Metathesaurus with semantic types such as “Pharmacological Substance (PHSU)” and “Antibiotics (ANTB)” can be collected to build a drug dictionary.

The joint chemical dictionary (Jochem) is a dictionary developed for the identification of drugs and small molecules in texts. It combines information from UMLS, MeSH, ChEBI, and so on [44]. Concepts in Jochem can be extracted to build a drug dictionary.

Dictionary-based approaches identify drug names by matching drug dictionaries against given texts. Exact matching approaches [45,46] usually achieve high precision, but suffer from low recall. This is because there are spelling mistakes or variants of drug names not covered by drug dictionaries. Therefore, approximate matching is used to improve the recall of dictionary-based approaches. Lexical similarity measures and approximate string matching methods such as edit distance [47], SOUNDEX [48] and Metaphone [49] can be used for approximate matching. For example, Levin *et al.* [50] used Metaphone to match generic and trade names of drugs in RxNorm [37] against anesthesia electronic health records. Moreover, there are approaches utilizing existing systems to map textual terms to drug dictionaries [51,52]. For example, Rindflesch *et al.* [51] utilized the UMLS MetaMap program [53] to map biomedical texts to UMLS Metathesaurus concepts. Phrases that were mapped to concepts with the semantic type "Pharmacological Substance" were considered to be drug names.

Dictionary-based approaches also may yield low precisions because of low quality of drug dictionaries. Sirohi *et al.* [54] investigated the effects of using varying drug dictionaries to extract drugs from electronic medical records and concluded that the precision and recall could be considerably enhanced by refining the dictionaries. Many methods have been used to improve the quality of drug dictionaries [44,55,56]. Hettne *et al.* [44] proposed several filtering rules to filter terms in a dictionary developed for DNR. For example, the short token filtering rule removed a term if the term was a singular character or an Arabic number after tokenization and removal of stop words. Moreover, they manually checked highly frequent terms in a set of randomly selected MEDLINE abstracts. If a term corresponded to a normal English word, it was added to a list of unwanted terms. Xu *et al.* [55] compared the drug dictionary with the SCOWL list [57], which is a list of normal English words. They manually reviewed ambiguous words and removed unlikely drug terms from the dictionary. At the same time, they expanded the dictionary by adding drug names annotated in a training dataset.

Due to the rapid development of pharmaceutical research, new drugs are constantly developed and enter the market. However, drug dictionaries cannot be updated regularly. It is impossible for a drug dictionary to cover all existing drugs. Therefore, approaches that do not rely too much on drug dictionaries are necessary for DNR.

5.2. Rule-Based Approaches

Rule-based approaches use rules that describe the composition patterns or context of drug names. Composition pattern-based rules are usually used to identify drug names that are generated following specific rules (e.g., systematic names and international nonproprietary names). For example, Lowe D. [58] *et al.* encoded the nomenclature rules as formal grammars to identify systematic names of drugs and chemicals. Segura-Bedmar *et al.* [3] built a regular expression for each international nonproprietary name stem recommended by World Health Organization to identify and classify drugs. However, composition pattern-based rules are ineffective for drug names generated without nomenclature rules. Context-based rules identify drug names by the context of drug names in free texts [59,60]. For example, Gold *et al.* [59] and Hamon *et al.* [60] used contextual clues to extract misspelled drug names and drug names not in drug dictionary from discharge summaries. Phrases that were surrounded by enough information such as dosages, frequencies and durations were considered as drug names and were extracted accordingly.

In addition to hand-crafted rules, rules that are automatically learned also have been used for DNR [61,62]. For example, Xu *et al.* [61] developed an iterative pattern leaning approach to extract drugs and other medical treatment concepts from randomized clinical trial abstracts. The approach started with a seed pattern such as “treated with NP (noun phrase)” or some seed instances (*i.e.*, drug names). Then it looped over a procedure consisting of two steps: pattern discovery and instance extraction. The discovered patterns and extracted instances were scored. Only top ranked patterns were used to extract instances and top ranked instances were considered as reliable instances.

Although rule-based approaches perform well when expert rules are available, the generation of rules is time-consuming. Moreover, rules developed for a specific class of drug names are not applicable for other classes of drug names, and too specific rules usually achieve high precision but low recall.

5.3. Machine Learning-Based Approaches

Machine learning-based approaches usually formalize DNR as a classification problem or a sequence labeling problem. Each token is presented as a set of features and then is labeled by machine learning algorithms with a class label. The class label denotes whether a token is part of a drug name and its position in a drug name. **BIO** is the most popular tagging scheme used for DNR. Tags in the **BIO** tagging scheme respectively represent that a token is at the beginning (**B**) of a drug name, inside (**I**) of a drug name and outside of a drug name (**O**). Figure 2 shows an example of **BIO** tagging results of a sentence from the DDIExtraction 2013 dataset, where four types of drugs (**drug**, **brand**, **group**, **non-human**) are defined. Moreover, there are some more expressive tagging schemes such as **BEIO**, **BESIO** and **B$_{12}$EIO** [63]. The tagging schemes are derived from **BIO**. Tag **E** represents that a token is at the end of a drug name. Tag **S** represents a single token drug name. Tags **B1** and **B2** in **B$_{12}$EIO** stand for the first token in a drug name and the second but not the last token in a drug name, respectively. Dai *et al.* [16] compared the effects of above four tagging schemes on DNR. It was demonstrated that **BESIO** outperformed other tagging schemes under the same experimental settings.

Sentence:PXR-ligands include a wide variety of pharmaceutical agents, such as antiepileptic drugs, taxol, rifampicin, and human immunodeficiency virus protease inhibitors such as ritonavir and saquinavir.

BIO:PXR-ligands\O include\O a\O wide\O variety\O of\O pharmaceutical\O agents\O ,\O such\O as\O antiepileptic\B-group drugs\I-group ,\O taxol\B-brand ,\O rifampicin\B-drug ,\O and\O human\B-group immunodeficiency\I-group virus\I-group protease\I-group inhibitors\I-group such\O as\O ritonavir\B-drug and\O saquinavir\B-drug .\O

Figure 2. An example of BIO tagging results of a sentence from the DDIExtraction 2013 dataset.

The selection of machine learning models is very important for machine learning-based approaches. Classification models commonly used for DNR include Maximum Entropy (ME) [64] and Support Vector Machine (SVM) [65]. They only consider individual tokens or phrases and do not take the order of tokens into account. Different from classification models, sequence tagging models such as Hidden Markov Model (HMM) [66] and Conditional Random Fields (CRF) [31,67–69] consider the complete sequence of tokens in a sentence. They aim at predicting the most probable sequence of tags for a given sentence. CRF is widely used and demonstrated to be superior to other machine learning models used for DNR. For example, CRF-based systems achieved the best performances in the DNR tasks of i2b2 medication extraction [67], CHEMDNER [31] and DDIExtraction 2013 [68] challenges. In most cases, only one machine learning model is used in a machine learning-based DNR approach. However, there are approaches using multiple models [31,70–73]. For example, Leaman *et al.* [31] employed two independent CRF models with different tokenization strategies and feature sets. Results of the two models were combined with heuristic rules. Lu *et al.* [70] used a character-level CRF and a token-level CRF to learn the internal structure and context of drugs, respectively. Results of the two CRF models were also merged in a heuristic method. Lamurias A. *et al.* [72] train multiple CRF models on different training datasets and combine the confidence scores returned by the models to rank and filter the identified drug names. Sikdar *et al.* [73] combine one SVM model and six CRF models that use different features to recognize drug names based on an ensemble framework. Table 4 lists some open source toolkits that can be used as implementations of commonly used machine learning models.

Table 4. Open source implementations of machine learning models.

Toolkit	Machine Learning Models	URL
MALLET	Naïve Bayes (NB), Decision Trees (DT), ME, HMM, CRF	http://mallet.cs.umass.edu/
WEKA	NB, DT, SVM	http://www.cs.waikato.ac.nz/ml/weka/
CRFsuite	CRF	http://www.chokkan.org/software/crfsuite/
CRF++	CRF	http://taku910.github.io/crfpp/
LIBSVM	SVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
SVM^{light}	SVM	http://www.cs.cornell.edu/People/tj/svm_light/

Performances of machine learning-based approaches highly depend on the features they used. Various types of features have been explored for DNR. Table 5 lists some features that are commonly used in machine learning-based DNR systems. Features based on the linguistic, orthographic and contextual information of tokens are widely used and the effectiveness of them is extensively studied. For example, Campos *et al.* [71] investigated the effects of features including lemma, POS, text chunking, dependency parsing, *etc.* Lemma, POS and text chunking features produced significant positive impacts on the performance of the CRF-based approach, while dependency parsing brought negative effects on the performance. Halgrim [64] examined the effects of POS, affix and orthographic feature in a ME-based approach and all the features provided positive outcomes.

Table 5. Features used in machine learning-based DNR systems.

Feature	Description	Reference
Character feature	*N*-grams of characters in a word.	[17,31,68,70,71,74]
Word feature	*N*-grams of words in a context window.	[17,31,64,68,70,71,74,75]
Lemma	*N*-grams of lemmas of words.	[17,31,68,71,74]
Stem	*N*-grams of stems of words.	[31,74]
POS	*N*-grams of POS tags.	[17,31,64,68,71,74,75]
Text chunking	*N*-grams of text chunking tags.	[17,71,75]
Dependency parsing	Dependency parsing results of words in a sentence.	[71]
Affix	Suffixes and prefixes of a word.	[17,31,64,68,71,74,75]
Orthographic feature	Starting with a uppercase letter, containing only alphanumeric characters, containing a hyphen, digits and capitalized characters counting, *etc.*	[17,31,64,68,71,74,75]
Word shape	Uppercase letters, lowercase letters, digits, and other characters in a word are converted to "A", "a", "0" and "O", respectively. For example, "Phenytoin" is mapped to "Aaaaaaaaa".	[17,31,68,71,74,75]
Dictionary feature	Whether an n-gram matches with part of a drug name in drug dictionaries.	[17,31,64,68,71,74,75]
Outputs of NER tools	Features derived from the output of existing chemical NER tools.	[31,68,74]
Word representation	Word representation features based on Brown clustering, word2vec, *etc.*	[70,75]
Conjunction feature	Conjunctions of different types of features, e.g., conjunction of lemma and POS features.	[17,71,75]

Domain-specific features such as dictionary features and features derived from outputs of existing chemical NER are also widely used. For example, Batista-Navarro *et al.* [17] compiled dictionaries from domain-specific knowledge resources including ChEBI, DrugBank, Jochem, *etc.* Each token was tagged by the dictionaries and the tagging results were used as features by a CRF-based approach. Rocktäschel *et al.* [68] generated domain-specific features from ChEBI, Jochem and the outputs of ChemSpot [76], which is a chemical NER tool. In general, domain-specific features can significantly improve the performances of machine learning-based approaches.

Recently, word representation features are exploited and demonstrated to be effective for DNR. Word representation features are generated by unsupervised machine learning algorithms on unstructured texts. They contain rich syntactic and semantic information of words. Many unsupervised machine learning algorithms have been proposed to learn word representation features and Brown Clustering algorithm [77] and word2vec [78] are most commonly used. For example, Lu *et al.* [70] employed Brown Clustering algorithm and word2vec to learn word representation features on MEDLINE documents. Then the word representation features were used to improve the performances of CRF-based DNR systems.

Moreover, conjunction features that combine different types of features are also used for DNR. Conjunction features can capture multiple linguistic characteristic of a word. For example, Batista-Navarro *et al.* [17] used conjunction features that combined lemmas and POS tags. Liu *et al.* [75]

selected 8 types of features including word feature, POS, text chunking, *etc.*, and combine them into conjunction features in two ways in their CRF-based DNR system.

Noisy features can significantly affect the performances and efficiencies of machine learning-based approaches. Therefore, the selection of informative and discriminative features is very important. However, determining the optimal subset of features by testing different combinations of features is time-consuming. Moreover, it is very likely that the optimal feature subset on a dataset will not perform well on another dataset. Therefore, automatic feature selection is necessary. In [75], Liu *et al.* employed three automatic feature selection methods, Chi-square [79], mutual information [80] and information gain [81], to eliminate noisy features for a CRF-based DNR system. Experimental results showed that each feature selection method could improve the performance of the CRF-based system.

Although machine learning-based approaches can achieve promising results, they require a sufficiently large and high quality annotated dataset for training. However, the creation of an annotated dataset is costly and time-consuming. Moreover, domain experts are required in the process of creating an annotated dataset.

5.4. Hybrid Approaches

Hybrid approaches combine multiple types of approaches to exploit the advantages and avoid the limitations of each type of approaches. In general, a post-processing step is needed to deal with the conflicting results of multiple approaches. Hybrid approaches usually produce better results than each component. Akhondi *et al.* [82] proposed a hybrid approach combining a dictionary-based approach and a rule-based approach based on the observation that different classes of drug names have different naming characteristics. The dictionary-based component is used to extract non-systematic names such as brand and generic drug names, and the rule-based component is used to extract systematic names, which are generated following standard nomenclature rules. Finally, the outputs of the dictionary-based and rule-based components are merged and the shorter one of two overlapping terms is removed. He *et al.* [83] constructed a drug name dictionary from DrugBank and MEDLINE abstracts. Then dictionary look-up was combined with a CRF-based approach to recognize drug names. For the overlapping terms, the results of dictionary look-up were kept. Due to the small size of training set, Tikk *et al.* [84] firstly developed a rule-based approach to label drug names in a large document set. Then a CRF-based approach was trained on the union of a small training set and the output of the rule-based approach. The CRF-based approach achieved better performance than that trained only on the small training set. Korkontzelos *et al.* [85] develop a voting system to combine a maximum entropy model, a perceptron classifier and a dictionary-based method to enhance the performance for DNR. Usié *et al.* [86] employ a CRF-based method, a dictionary and some regular expressions to recognize different types of drug names and then integrate the recognition results of the methods.

The performances of representative DNR systems on datasets of different types of texts are listed in Table 6. In the third column, “Dict”, “Rule”, “ML” and “Hybrid” denote dictionary-based, rule-based, machine learning-based and hybrid approaches, respectively. The fifth column lists the F-scores of DNR systems that only recognize drug names from texts (drug detection), while the sixth column lists the F-scores of DNR systems that not only recognize drugs from texts but also classify the recognized drugs into predefined classes (drug classification).

The fourth column of Table 6 lists the datasets on which the systems have been run on. We can see that machine learning-based systems or hybrid systems containing a machine learning component outperform other systems on the same dataset. Therefore, machine learning-based approaches or hybrid approaches that contain a machine learning component are the best choices if annotated datasets are available. Performance differences between machine learning-based approaches are mainly because of the selection of different machine learning models and different features.

Table 6. Performances of representative DNR systems on different types of texts.

Type of Texts	Reference	Category of Approach	Dataset	F-Score Detection	F-Score Classification
Discharge summaries	[64]	ML	i2b2 Medication Extraction	89.80%	*
	[67]	ML	i2b2 Medication Extraction	88.35%	*
	[84]	Hybrid (ML + Rule)	i2b2 Medication Extraction	87.06%	*
	[87]	Dict	i2b2 Medication Extraction	85.89%	*
	[60]	Rule	i2b2 Medication Extraction	80.00%	*
	[59]	Rule	26 discharge summaries	87.92%	*
Clinic office visit notes	[54]	Dict	52 clinic office visit notes	73.80%	*
Biomedical literature abstracts	[31]	ML	CHEMDNER	87.39%	*
	[70]	ML	CHEMDNER	87.11%	*
	[58]	Rule	CHEMDNER	86.86%	*
	[88]	Dict	CHEMDNER	77.91%	*
	[82]	Hybrid (Dic + Rule)	CHEMDNER	77.84%	*
	[44]	Dict	100 abstracts	50.00%	*
DrugBank documents	[83]	Hybrid (ML + Dic)	DDIExtraction 2011	92.54%	*
Mix of DrugBank documents and biomedical literature abstracts	[75]	ML	DDIExtraction 2013	83.85%	79.36%
	[68]	ML	DDIExtraction 2013	83.30%	71.50%
	[52]	Dict	DDIExtraction 2013	66.70%	*
	[89]	Dict	DDIExtraction 2013	60.90%	52.90%

By comparing the fifth and the sixth column, we can see that drug classification is more difficult than drug detection. The performances for drug classification are relatively poor and more efforts should be devoted to drug classification.

6. Concluding Remarks and Future Perspectives

Many approaches have been proposed for DNR, ranging from simple dictionary-based approaches to complex hybrid approaches. These approaches differ in the degree of manual intervention, portability, and applicable situation. Each type of the approach has advantages over other types. Dictionary-based approaches are effective when comprehensive and up-to-date drug dictionaries are available. Moreover, dictionary-based approaches can normalize drug names in texts by mapping them to unique identifiers in drug dictionaries. In contrast, machine learning-based approaches can only identify drug names from texts. However, the creation and maintenance of comprehensive drug dictionaries are costly and time-consuming. Rule-based approaches are suitable when drug names are generated regularly. Rule-based approaches can be easily optimized by modifying existing rules or adding new rules. However, there is an unavoidable trade-off between precision and recall for rule-based approaches. Rules that are too specific achieve high precision but low recall. On the other hand, rules that are too general lead to high recall but low precision. Furthermore, the portability of rule-based approaches is poor. Rules defined for a class of drugs cannot be adapted to other classes. In contrast, machine learning-based approaches for a class of drugs can be easily retrained for other classes on corresponding training datasets. Machine learning-based approaches often outperform dictionary-based and rule-based approaches when sufficiently large and high quality annotated training datasets are available. However, it is costly to annotate datasets manually. Given the above, hybrid approaches that combine different approaches have been increasingly used.

At the present time, the state-of-the-art approaches for DNR are mainly based on traditional machine learning models such as CRF and SVM. Performance improvements of the state-of-the-art approaches depend heavily on exploring and using new effective features. However, performance improvements from new features are limited. It is necessary to explore new machine learning models for DNR. In recent years, deep neural networks (DNNs) [90] have been used in many machine learning tasks such as speech recognition [91] and visual object recognition [92] and achieved unprecedented success. It is worth exploring the use of DNNs for DNR.

The lack of sufficiently large and high quality training datasets is a major barrier to future work on DNR. Semi-supervised learning is a machine learning technique, which requires a small amount of annotated data and a large amount of unannotated data for training. Typical semi-supervised learning methods such as bootstrapping [93] and active learning [94] have demonstrated their effectiveness for improving the performances of systems when annotated data is scarce. Therefore, semi-supervised learning is a promising solution to lack of training datasets for DNR.

Another barrier to further development of DNR is the imbalance of training datasets. For example, drugs of the "**drug**" class account for 63% of all drugs in the DDIExtraction 2013 dataset, while drugs of the "**no-human**" class account for only 4%. As a result of the imbalance of training datasets, the top ranked system of the DDIExtraction 2013 challenge achieved an F-score of 79.0% for "**drug**", but only 14.1% for "**no-human**". Automatic text generation techniques based on formal grammar are likely to solve the imbalance problem of training datasets. Formal grammar can capture the morphology, syntax and semantic information of a language. It has been demonstrated that automatic text generation techniques based on formal grammar can automatically build realistic chemical-related training documents for chemical name extraction [95]. Moreover, automatic text generation techniques can control the density of different classes of training examples, the variety and the complexity of contexts, as well as the size of the training sets. For future work, it is worth trying to improve the performance of DNR systems by automatically generating training datasets without data imbalance.

Although there are a few approaches proposed to recognize both continuous and non-continuous named entities such as disorders [96–98], they still perform poorly for non-continuous named entities. For example, in the CHEMDNER task of the BioCreative IV challenge, non-continuous drugs (*i.e.*, multiple drugs) account for less than one percent and no participating system specially deals with them. All participating systems achieve poor performance for the multiple drugs. Therefore, effective solutions to non-continuous drug name recognition are needed for future work.

Acknowledgments

This paper is supported in part by grants: National 863 Program of China (2015AA015405), NSFC (National Natural Science Foundation of China) (61402128, 61473101, 61173075, and 61272383) and Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20140508161040764, JCYJ20140417172417105 and JCYJ20140627163809422).

Author Contributions

The work presented here was a collaboration of all the authors. All authors contributed to designing the methods and experiments. Shengyu Liu performed the experiments. Shengyu Liu, Buzhou Tang, Qingcai Chen and Xiaolong Wang analyzed the data and interpreted the results. Shengyu Liu wrote the paper. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Segura-Bedmar, I.; Martínez, P.; Pablo-Sánchez, C. Using a shallow linguistic kernel for drug-drug interaction extraction. *J. Biomed. Inform.* **2011**, *44*, 789–804.
2. Warrer, P.; Hansen, W.; Juhl-Jensen, L.; Aagaard, L. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *Br. J. Clin. Pharmacol.* **2012**, *73*, 674–684.
3. Segura-Bedmar, I.; Martínez, 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*, 816–823.
4. Uzuner, Ö.; Solti, I.; Cadag, E. Extracting medication information from clinical text. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 514–518.
5. Krallinger, M.; Leitner, F.; Rabal, O.; Vazquez, M.; Oyarzabal, J.; Valencia, A. CHEMDNER: The drugs and chemical names extraction challenge. *J. Cheminform.* **2015**, *7(S1)*, S1.
6. Segura-Bedmar, I.; Martínez, P.; Herrero-Zazo, M. SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 341–350.
7. Vazquez, M.; Krallinger, M.; Leitner, F.; Valencia, A. Text mining for drugs and chemical compounds: Methods, tools and applications. *Mol. Inf.* **2011**, *30*, 506–519.
8. Gurulingappa, H.; Mudi, A.; Toldo, L. Challenges in mining the literature for chemical information. *RSC Adv.* **2013**. *3*, 16194–16211.
9. Eltyeb, S.; Salim, N. Chemical named entities recognition: A review on approaches and applications. *J Cheminform.* **2014**, *6*, 17, doi:10.1186/1758-2946-6-17.
10. Gurulingappa, H.; Rajput, A.; Roberts, A.; Fluck, J.; Hofmann-Apitius, M.; Toldo, L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed. Inform.* **2012**, *45*, 885–892.

11. Wu, H.; Karnik, S.; Subhadarshini, A.; Wang, Z.; Philips, S.; Han, X.; Chiang, C.; Liu, L.; Boustani, M.; Rocha, L.M.; *et al.* An integrated pharmacokinetics ontology and corpus for text mining. *BMC Bioinform.* **2013**, *14*, 35, doi:10.1186/1471-2105-14-35.
12. Boyce, R.; Gardner, G.; Harkema, H. Using natural language processing to extract drug-drug interaction information from package inserts. In Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Montreal, QC, Canada, 3–8 June 2012; pp. 206–213.
13. Mulligen, E.; Fourrier-Reglat, A.; Gurwitz, D.; Molokhia, M.; Nieto, A.; Trifiro, G.; Kors, J.A.; Furlong, L.I. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *J. Biomed. Inform.* **2012**, *45*, 879–884.
14. Segura-Bedmar, I.; Martínez, P.; Sánchez-Cisneros, D. The 1st DDIExtraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts. In Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction, Huelva, Spain, 5 September 2011; pp. 1–9.
15. Herrero-Zazo, M.; Segura-Bedmar, I.; Martínez, P.; Declerck, T. The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *J. Biomed. Inform.* **2013**, *46*, 914–920.
16. Dai, H.; Lai, P.; Chang, Y.; Tsai, R.T. Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *J. Cheminform.* **2015**, *7(S1)*, S14.
17. Batista-Navarro, R.; Rak, R.; Ananiadou, S. Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. *J. Cheminform.* **2015**, *7(S1)*, S6.
18. Treebank tokenization. Available online: http://www.cis.upenn.edu/~treebank/tokenization.html (accessed on 24 November 2015).
19. Ferrucci, D.; Lally, A. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.* **2004**, *10*, 327–348.
20. Apache UIMA. Available online: http://uima.apache.org/ (accessed on 24 November 2015).
21. Kano, Y.; Baumgartner, W.A., Jr.; McCrohon, L.; Ananiadou, S.; Cohen, K.B.; Hunter, L.; Tsujii, J. U-Compare: Share and compare text mining tools with UIMA. *Bioinformatics* **2009**, *25*, 1997–1998.
22. Kano, Y.; Miwa, M.; Cohen, K.B.; Hunter, L.E. U-Compare: A modular NLP workflow construction and evaluation system. *IBM J. Res. Dev.* **2011**, *55*, 1–11.
23. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.J.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 22–27 June 2014; pp. 55–60.
24. Hahn, U.; Buyko, E.; Landefeld, R.; Mühlhausen, M.; Poprat, M.; Tomanek, K.; Wermter, J. 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", Marrakech, Morocco, 26–27 May 2008; pp. 1–7.
25. Tsuruoka, Y.; Tsujii, J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, BC, Canada, 6–8 October 2005; pp. 467–474.

26. Miyao, Y.; Saetre, R.; Sagae, K.; Matsuzaki, T; Tsujii, J. Task-oriented evaluation of syntactic parsers and their representations. In Proceedings of the 45th Meeting of the Association for Computational Linguistics, Columbus, OH, USA, 19–20 June 2008; pp. 46–54.
27. Campos, D.; Matos, S.; Oliveira, J. A modular framework for biomedical concept recognition. *BMC Bioinform.* **2013**, *14*, 281.
28. Liu, H.; Christiansen, T.; Baumgartner, W.A., Jr.; Verspoor, K. BioLemmatizer: A lemmatization tool for morphological processing of biomedical text. *J. Biomed. Semant.* **2012**, *3*, 3.
29. Savova, G.; Masanz, J.; Ogren, P.; Zheng, J.; Sohn, S.; Kipper-Schuler, K.C.; Chute, C.G. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 507–513.
30. Grego, T.; Pinto, F.; Couto, F.M. LASIGE: Using conditional random fields and ChEBI ontology. In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 660–666.
31. Leaman, R.; Wei, C.; Lu, Z. tmChem: A high performance approach for chemical named entity recognition and normalization. *J. Cheminform.* **2015**, *7(S1)*, S3.
32. Grego, T.; Couto, F.M. Enhancement of chemical entity identification in text using semantic similarity validation. *PLoS ONE* **2013**, *8*, e62984.
33. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A.C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; *et al.* DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **2013**, *42*, D1091–D1097.
34. Kanehisa, M.; Goto, S.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.* **2014**, *42*, D199–D205.
35. Hernandez-Boussard, T.; Whirl-Carrillo, M.; Hebert, J.; Gong, L.; Owen, R.; Gong, M.; Gor, W.; Liu, F.; Truong, C.; Whaley, R.; *et al.* The pharmacogenetics and pharmacogenomics knowledge base: Accentuating the knowledge. *Nucleic Acids Res.* **2008**, *36*, D913–D918.
36. Davis, A.P.; Grondin, C.J.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; King, B.L.; Wiegers, T.C.; Mattingly, C.J. The Comparative Toxicogenomics Database's 10th year anniversary: Update 2015. *Nucleic Acids Res.* **2015**, *43*, D914–D920.
37. Liu, S.; Ma, W.; Moore, R.; Ganesan, V.; Nelson, S. RxNorm: Prescription for electronic drug information exchange. *IT Prof.* **2005**, *7*, 17–23.
38. Fung, K.; McDonald, C.; Bray, B. RxTerms—A drug interface terminology derived from RxNorm. In Proceedings of the AMIA 2008 Annual Symposium, Washington, DC, USA, 8–12 November 2008; pp. 227–231.
39. Qin, C.; Zhang, C.; Zhu, F.; Xu, F.; Chen, S.Y.; Zhang, P.; Li, Y.H.; Yang, S.Y.; Wei, Y.Q.; Tao, L.; *et al.* Therapeutic target database update 2014: A resource for targeted therapeutics. *Nucleic Acids Res.* **2014**, *42*, D1118–D1123.
40. Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Res.* **2013**, *41*, D456–D463.
41. Lipscomb, C. Medical Subject Headings (MeSH). *Bull. Med. Libr. Assoc.* **2000**, *88*, 265–266.
42. Li, Q.; Cheng, T.; Wang, Y.; Bryant, S.H. PubChem as a public resource for drug discovery. *Drug Discov. Today* **2010**, *15*, 1052–1057.

43. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270.
44. Hettne, K.; Stierum, R.; Schuemie, M.; Hendriksen, P.J.; Schijvenaars, B.J.; Mulligen, E.M.; Kleinjans, J.; Kors, J.A. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **2009**, *25*, 2983–2991.
45. Kolářik, C.; Hofmann-Apitius, M.; Zimmermann, M.; Fluck, J. Identification of new drug classification terms in textual resources. *Bioinformatics* **2007**, *23*, i264–i272.
46. Chhieng, D.; Day, T.; Gordon, G.; Hicks, J. Use of natural language programming to extract medication from unstructured electronic medical records. In Proceedings of the AMIA 2007 Annual Symposium, Chicago, IL, USA, 10–14 November 2007; p. 908.
47. Wanger, R.; Fischer, M. The string-to-string correction problem. *J. ACM* **1974**, *21*, 168–173.
48. Hall, P.; Dowling, G. Approximate string matching. *Comput. Surv.* **1980**, *12*, 381–402.
49. Philips, L. Hanging on the Metaphone. *Comput. Lang.* **1990**, *7*, 12.
50. Levin, M.; Krol, M.; Doshi, A.; Reich, D. Extraction and mapping of drug names from free text to a standardized nomenclature. In Proceedings of the AMIA 2007 Annual Symposium, Chicago, IL, USA, 10–14 November 2007; pp. 438–442.
51. Rindflesch, T.; Tanabe, L.; Weinstein, J.; Hunter, L. EDGAR: Extraction of drugs, genes and relations from the biomedical literature. In Proceedings of the Pacific Symposium on Biocomputing 2000 (PSB 2000), Honolulu, HI, USA, 5–9 January 2000; pp. 517–528.
52. Sanchez-Cisneros, D.; Martínez, P.; Segura-Bedmar, I. Combining dictionaries and ontologies for drug name recognition in biomedical texts. In Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics, Miami, FL, USA, 4–7 December 2013; pp. 27–30.
53. Aronson, A. Effective mapping of biomedical text to the UMLS Metathesaurus: The metamap program. In Proceedings of the AMIA 2001 Annual Symposium, Washington, DC, USA, 3–7 November 2001; pp. 17–21.
54. Sirohi, E.; Peissig, P. Study of effect of drug lexicons on medication extraction from electronic medical records. In Proceedings of the Pacific Symposium on Biocomputing 2005, Big Island of Hawaii, HI, USA, 4–8 January 2005; pp. 308–318.
55. Xu, H.; Stenner, S.; Doan, S.; Johnson, K.B.; Waitman, L.R.; Denny, J.C. MedEx: A medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 19–24.
56. Ata, C.; Can, T. DBCHEM: A database query based solution for the chemical compound and drug name recognition task. In Proceedings of the 4th BioCreative Challenge Evaluation Workshop, Bethesda, MD, USA, 7–9 October 2013; pp. 42–46.
57. SCOWL (And Friends). Available online: http://wordlist.aspell.net/ (accessed on 24 November 2015).
58. Lowe, D.; Sayle, R. LeadMine: A grammar and dictionary driven approach to entity recognition. *J. Cheminform.* **2015**, *7(S1)*, S5.
59. Gold, S.; Elhadad, N.; Zhu, X.; Cinimo, J.J.; Hripcsak, G. Extracting structured medication event information from discharge summaries. In Proceedings of the AMIA 2008 Annual Symposium, Washington, DC, USA, 8–12 November 2008; pp. 237–241.
60. Hamon, T.; Grabar, N. Linguistic approach for identification of medication names and related information in clinical narratives. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 549–554.

61. Xu, R.; Morgan, A.; Das, A.; Garber, A. Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Stroudsburg, PA, USA, 4–5 June 2009; pp. 63–70.
62. Coden, A.; Gruhl, D.; Lewis, N.; Tanenblatt, M.; Terdiman, J. SPOT the drug! An unsupervised pattern matching method to extract drug names from very large clinical corpora. In Proceedings of the IEEE 2nd International Conference on Healthcare Informatics, Imaging and Systems Biology, San Diego, CA, USA, 27–28 September 2012; pp. 33–39.
63. Zhao, H.; Huang, C.; Li, M.; Lu, B. A unified character-based tagging framework for Chinese word segmentation. *ACM Trans. Asian Lang. Inf. Process.* **2010**, *9*, 1–32.
64. Halgrim, S.; Xia, F.; Solti, I.; Cadag, E.; Uzuner, O. A cascade of classifiers for extracting medication information from discharge summaries. *J. Biomed. Semant.* **2011**, *2(S3)*, S2.
65. Björne, J.; Kaewphan, S.; Salakoski, T. UTurku: Drug named entity detection and drug-drug interaction extraction using SVM classification and domain knowledge. In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 651–659.
66. Malyszko, J.; Filipowska, A. Lexicon-free and context-free drug names identification methods using hidden markov models and pointwise mutual information. In Proceedings of the 6th International Workshop on Data and Text Mining in Biomedical Informatics, Maui, HI, USA, 29 October–2 November 2012; pp. 9–12.
67. Patrick, J.; Li, M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 524–527.
68. Rocktäschel, T.; Huber, T.; Weidlich, M.; Leser, U. WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 356–363.
69. Abacha, A.B.; Chowdhury, M.F.M.; Karanasiou, A.; Mrabet, Y.; Lavelli, A.; Zweigenbaum, P. Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification. *J. Biomed. Inform.* **2015**, *58*, 122–132.
70. Lu, Y.; Ji, D.; Yao, X.; Wei, X.; Liang, X. CHEMDNER system with mixed conditional random fields and multi-scale word clustering. *J. Cheminform.* **2015**, *7(S1)*, S4.
71. Campos, D.; Matos, S.; Oliveira, J. A document processing pipeline for annotating chemical entities in scientific documents. *J. Cheminform.* **2015**, *7(S1)*, S7.
72. Lamurias, A.; Grego, T.; Couto, F.M. Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI. In Proceedings of the 4th BioCreative Challenge Evaluation Workshop, Bethesda, MD, USA, 7–9 October 2013; pp. 75–81.
73. Sikdar, U.K.; Ekbal, A.; Saha, S. Domain-independent model for chemical compound and drug name recognition. In Proceedings of the 4th BioCreative Challenge Evaluation Workshop, Bethesda, MD, USA, 7–9 October 2013; pp. 158–161.
74. Huber, T.; Rocktäschel, T.; Weidlich, M.; Thomas, P.; Leser U. Extended feature set for chemical named entity recognition and indexing. In Proceedings of the 4th BioCreative Challenge Evaluation Workshop, Bethesda, MD, USA, 7–9 October 2013; pp. 88–91.

75. Liu, S.; Tang, B.; Chen, Q.; Wang, X.; Fan, X. Feature engineering for drug name recognition in biomedical texts: Feature conjunction and feature selection. *Comput. Math. Method Med.* **2015**, doi:10.1155/2015/913489.
76. Rocktäschel, T.; Weidlich, M.; Leser, U. ChemSpot: A hybrid system for chemical named entity recognition. *Bioinformatics* **2012**, *28*, 1633–1640.
77. Brown, P.F.; de Souza, P.V.; Mercer, R.L.; Pietra, V.; Lai, J. Class-based *N*-gram models of natural language. *Comput. Linguist.* **1992**, *18*, 467–479.
78. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, CA, USA, 5–10 December 2013; pp. 3111–3119.
79. Forman, G. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* **2003**, *3*, 1289–1305.
80. Yang, Y.; Pedersen, J. A comparative study on feature selection in text categorization. In Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, USA, 8–12 July 1997; pp. 412–420.
81. Zheng, Z.; Wu, X.; Srihari, R. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explor. Newslett.* **2004**, *6*, 80–89.
82. Akhondi, S.; Hettne, K.; Horst, E.; van Mulligen, E.M.; Kors, J.A. Recognition of chemical entities: Combining dictionary-based and grammar-based approaches. *J Cheminform.* **2015**, *7(S1)*, S10.
83. He, L.; Yang, Z.; Lin, H.; Li, Y. Drug name recognition in biomedical texts: A machine-learning-based method. *Drug Discov. Today* **2014**, *19*, 610–617.
84. Tikk, D.; Solt, I. Improving textual medication extraction using combined conditional random fields and rule-based systems. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 540–544.
85. Korkontzelos, I.; Piliouras, D.; Dowsey, A.W.; Ananiadou, S. Boosting drug named entity recognition using an aggregate classifier. *Artif. Intell. Med.* **2015**, *65*, 145–153.
86. Usié, A.; Cruz, J.; Comas, J.; Solsona, F.; Alves, R. A tool for the identification of chemical entities (CheNER-BioC). In Proceedings of the 4th BioCreative Challenge Evaluation Workshop, Bethesda, MD, USA, 7–9 October 2013; pp. 66–69.
87. Yang, H. Automatic extraction of medication information from medical discharge summaries. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 545–548.
88. Irmer, M.; Bobach, C.; Böhme, T.; Laube, U.; Püschel, A.; Weber L. Chemical named entity recognition with OCMiner. In Proceedings of the 4th BioCreative Challenge Evaluation Workshop, Bethesda, MD, USA, 7–9 October 2013; pp. 92–96.
89. Sanchez-Cisneros, D.; Gali, F.A. UEM-UC3M: An ontology-based named entity recognition system for biomedical texts. In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 622–627.
90. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural. Netw.* **2015**, *61*, 85–117.
91. Hinton, G.; Deng, L.; Dahl, E.; Fadiga, L.; Metta, G. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal. Process. Mag.* **2012**, *29*, 82–97.
92. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems 25 (NIPS 2012), Lake Tahoe, NV, USA, 3–6 December 2012.

93. Liu, X.; Zhang, S.; Wei, F.; Zhou, M. Recognizing named entity in tweets. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011; pp. 359–367.
94. Majumder, M.; Barman, U.; Prasad, R.; Saurabh, K.; Saha, S. A novel technique for name identification from homeopathy diagnosis discussion forum. *Proc. Technol.* **2012**, *6*, 379–386.
95. Yan, S.; Spangler, W.; Chen, Y. Chemical name extraction based on automatic training data generation and rich feature set. *IEEE-ACM Trans. Comput. Biol. Bioinform.* **2013**, *10*, 1218–1233.
96. Tang, B.; Wu, Y.; Jiang, M.; Denny, J.; Xu, H. Recognizing and encoding disorder concepts in clinical text using machine learning and vector space model. In Proceedings of the Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, Valencia, Spain, 23–26 September 2013.
97. Cogley, J.; Stokes, N.; Carthy, J. Medical disorder recognition with structural support vector machines. In Proceedings of the Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, Valencia, Spain, 23–26 September 2013.
98. Leal, A.; Martins, B.; Couto, F.M. ULisboa: Recognition and normalization of medical concepts. In Proceedings of the 9th International Workshop on Semantic Evaluation, Denver, CO, USA, 4–5 June 2015; pp. 406–411.