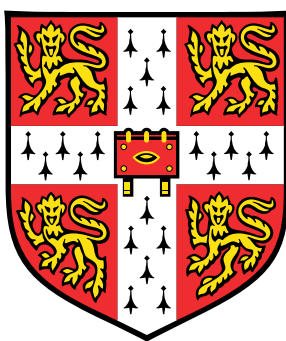


Identifying bioactivity events of small molecules from the scientific literature



Yan Ying

European Bioinformatics Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Gonville & Caius College

May 2015

This thesis is dedicated to my father 严世明, my mother 卢萍, and my husband Oliver.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations.

Yan Ying
May 2015

Acknowledgements

My first and sincere appreciation goes to my principal supervisor Dr Dietrich Rebholz-Schuhmann for all I have learned from him and for patient guidance, encouragement and advice he has provided throughout my time as his student. I would also like to thank him for being an open person to ideas. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly.

I would like to express my deep gratitude and respect to Dr Christoph Steinbeck, my co-supervisor, whose advice and insight was invaluable to me. My thanks go to him for all I learned from him, and for providing the opportunity of collaboration with the Cheminformatics and Metabolism group in EBI. I am most thankful for Christoph's generosity in providing me office space after Dietrich left EBI.

I would like to thank Dr Şenay Kafkas, a text miner in the Literature Services group at EBI, who has provided friendship and mentoring during my PhD and was a great help in implementing machine learning experiments. I was impressed by Şenay's expertise and knowledge in the text mining domain, and I was also continually amazed by her willingness to help me to solve research problems.

I would like to thank Dr Jee-Hyub Kim for his assistance and guidance in getting my PhD project started on the right foot and providing me with the foundation for becoming a text miner in the biomedical realm. Although Jee-Hyub was a mentor only for the first half of my research, in the long run, I believe that his help provided me with the unique opportunity to gain a wider breadth of experience.

I would like to thank the members of my thesis advisory committee, Dr Ann Copestake and Dr Carsten Schultz, for their input, valuable discussions and accessibility provided during my TAC meetings.

I would like to sincerely thank my thesis examiners, Dr Simone Teufel and Dr Goran Nenadic, for their patience and understanding, and the significant contribution they made to the final version of this thesis.

I would like to thank Dr Matt Conroy for his contribution to the work curating the DrugPro corpus. Matt has also been a very reliable and calm person to whom I could

always talk about my problems and anxieties. Thanks to him for questioning me about my ideas, helping me think rationally and for hearing my problems.

I express my gratitude to my husband Oliver for his support and encouragement, and the countless hours he spent proof-reading the thesis and improving my LaTeX code. Last but not least I recognise the patience of my mother, who experienced all of the ups and downs of my research.

Completing this work would have been all the more difficult were it not for the support and friendship provided by the other members of the European Bioinformatics Institute. I am indebted to them for their help. Those people, such as the postgraduate students at EBI and the research staff, who provided a much-needed form of escape from my studies, also deserve recognition for helping me keep things in perspective.

Finally, I would like to thank the European Bioinformatics Institute, not only for providing the funding which allowed me to undertake this research, but also for giving me the opportunity to attend conferences and meet so many interesting people.

Abstract

The PhD work focused on the automatic extraction of small molecule bioactivity information from scientific literature, a topic that ultimately aims to contribute to the small molecule ontology ChEBI. The ChEBI ontology plays an important role in assisting biomedical research, but requires further expansion in order to improve the ability to interrelate and link ChEBI to other primary biological resources. Small molecule bioactivity is a concept that presents small molecule characteristics in scenarios involving other biological objects. Bioactivity effects can have various targets including genes, proteins, organs, and complete organisms.

Two methods of extracting these events were examined. In the first, a supervised machine learning approach was proposed and developed based on a new corpus: DrugPro. The developed system comprised cascaded classifiers and a post-processing unit to extract the specific relationship between drugs and bacteria. The classification system delivered a good result when combining domain-specific features and a term weighting scheme, achieving precision of 92.56% and recall of 93.22% on the test corpus. A sentence classification system also achieved reasonable results, with precision of 58.9% and recall of 57.4% on the test corpus. A small set of language patterns was developed to extract the key items of information from the positive sentences.

The second method of extracting small molecule bioactivity events involved the development of an entirely rule-based extraction system. The performance of the system was evaluated on a random selection of 20 abstracts, which resulted in 94.7% precision and 94.7% recall on this (admittedly limited) data set. The results indicated the rule-based approach has the potential to precisely identify general small molecule bioactivity events in a much larger collection of biomedical text.

It is concluded that extension and enhancement of the ChEBI ontology could be based on the concept of small molecule bioactivity, and text mining provides a viable strategy by which this can be implemented. A further contribution of the work is the set of text mining resources created here, including the annotated corpus, and an investigation into the best way to undertake named-entity recognition for the multi-domain concept of bioactivity. It

is expected that these concepts can be adapted to similar biomedical text mining tasks.

Contents

Contents	xi
List of Figures	xv
List of Tables	xvii
Nomenclature	xx
Publications	xxi
Annotation scheme	xxiii
1 Introduction	1
1.1 Context	1
1.1.1 ChEBI	1
1.1.2 Limitations of the ChEBI ontology	2
1.1.3 Small molecule bioactivity	3
1.1.4 ChEMBL	4
1.1.5 Connection of ChEBI with other ontologies	4
1.1.6 The use of text mining in biomedical research	7
1.1.7 The presentation of bioactivity in biomedical text	9
1.1.8 Challenges	11
1.2 Research objectives	12
1.2.1 Use of explicit and non-explicit verbs in a specific case of small molecule bioactivity	12
1.2.2 General small molecule bioactivity on a wide range of possible targets	12
1.3 Biomedical text mining - tools and techniques	13
1.3.1 Named entity recognition	13

1.3.2	Corpora	17
1.3.3	Parsing	18
1.3.4	Relation extraction	18
1.3.5	Linguistic patterns	19
1.4	Related work	19
1.4.1	Extension of bioinformatics resources	19
1.4.2	Protein-protein interaction	21
1.4.3	BioNLP	22
1.4.4	Summary	23
1.5	Contributions and thesis organisation	24
1.6	Chapter summary	25
2	Drug-bacterium bioactivity events	27
2.1	Technical background	28
2.2	Classification using support vector machines	30
2.3	System overview	30
2.4	The DrugPro corpus	32
2.4.1	Selection of abstracts	33
2.4.2	Abstract annotation	33
2.5	Document-level classification	34
2.5.1	Document level pre-processing	35
2.5.2	Extracting features	38
2.5.3	Document SVM classification	41
2.6	Sentence-level classification	41
2.6.1	Sentence-level pre-processing	43
2.6.2	Sentence SVM classification	45
2.7	Relationship extraction	46
2.7.1	Keyword-based relationship identification	47
2.8	Chapter summary	50
3	General small-molecule bioactivity events	51
3.1	Technical background	52
3.1.1	Rule-based event extraction process	54
3.2	Rule development for event detection and characterisation	55
3.2.1	Rule classification	56
3.3	Named entity recognition	57

3.3.1	Dictionary selection and analysis	58
3.3.2	NER implementation	59
3.4	Noun phrase as event trigger	62
3.4.1	Patterns of biological role terms	62
3.4.2	Pre-processing	64
3.4.3	Identifying the bioactivity role term	66
3.4.4	Classifying bioactivity terms	67
3.5	Verb phrase as event trigger	67
3.5.1	Patterns of bioactivity verb phrases	69
3.5.2	Pre-processing	70
3.5.3	Special cases	72
3.6	Bioactivity triple extraction	74
3.7	Chapter summary	75
4	Results and discussion	77
4.1	Performance parameters	77
4.2	Results from the drug-bacterium study	78
4.2.1	Document classification results	78
4.2.2	Sentence classification results	81
4.2.3	Results of relationship extraction	82
4.2.4	Analysis of errors	83
4.2.5	Drug-bacteria bioactivity in the DrugPro corpus	85
4.3	Results from the general bioactivity study	86
4.3.1	Analysis of target entity recognition	86
4.3.2	Analysis of small molecule recognition	89
4.3.3	Manual assessment of the rule-based extraction pipeline	91
4.3.4	Extraction of small molecule bioactivity from MEDLINE	94
4.4	Qualitative comparison of the two approaches	96
4.5	Chapter summary	97
5	Conclusion	99
5.1	Summary of the research	99
5.2	Contributions	102
5.3	Future research	102
5.3.1	Implementing changes to ChEBI	103
5.3.2	Development of a general corpus for small molecule bioactivity	103

5.3.3	Solving the NER problem for small molecule bioactivity	104
5.3.4	Automatic pattern generation	104
5.4	Final remarks	105
References		107

List of Figures

1.1	Example of a ChEBI role classification	2
1.2	ChEBI bioactivity ontology model	7
1.3	PubMed Growth	8
1.4	Example bioactivity annotation	10
1.5	Whatizit web-based annotation tool	15
1.6	The interface of ABNER NER	16
2.1	A hyperplane of a support vector machine	31
2.2	System overview	32
2.3	A fragment from the annotated DrugPro corpus	35
2.4	The main components of the document-level SVM classifier	36
2.5	Pre-processing for document classification	37
2.6	Bag-of-words vector with concatenated domain-specific features	41
2.7	Overview of sentence-level classification	42
2.8	Dependency relations generated by Gdep	45
2.9	Shortest path dependency feature	45
2.10	Identification of relation type	49
3.1	The rule-based system of extracting bioactivity events used in this thesis	56
3.2	The dependency parse tree of an example bioactivity description	63
3.3	Overview of the rule-based extraction of bioactivity role terms	65
3.4	A dependency parser tree of an active sentence	68
3.5	A dependency parse tree of a passive sentence	69
3.6	Pre-processing work flow for bioactivity verb phrases	71
4.1	Trendline of SVM performance based on BOW vector size	80

List of Tables

1.1	Similar projects investigating links between chemical and biological entities	24
2.1	Contingency table for document frequency	38
2.2	Keywords found in the DrugPro corpus	40
2.3	Sentence classification feature set	46
2.4	Keywords found in the DrugPro corpus	47
2.5	Language patterns of rules used to extract drug-bacterium relationships . .	49
2.6	Negation cues	50
3.1	Two types of trigger word	57
3.2	NER dictionaries	58
3.3	Analysis of cross-tagging between different resources	59
3.4	NER performance against SCAI corpus	61
3.5	Patterns for bioactivity noun phrases	64
3.6	Patterns for bioactivity verb phrases	70
3.7	Patterns for general bioactivity event extraction	74
4.1	Classification performance using various BOW vector sizes	79
4.2	Classification performance using other features.	80
4.3	Classification performance using other concatenated features	81
4.4	Sentence classification performance using other concatenated features . . .	82
4.5	Performance of the keyword-based extraction on the filtered sentences . . .	83
4.6	Co-occurrence of bioactivity role term and NER tag, as found in MEDLINE abstracts	87
4.7	Analysis of triples from MEDLINE	90
4.8	Filtering result	91
4.9	Extraction of bioactivity events from all of MEDLINE	94
4.10	Bioactivity target type distribution	95

5.1	Review of language patterns used to extract drug-bacterium relationships .	100
5.2	Review of language patterns used for general bioactivity event extraction . .	101

Nomenclature

Acronyms / Abbreviations

ABNER A Biomedical Named Entity Recognizer

BNC British National Corpus

BOW Bag-of-words

ChEBI Chemical Entities of Biological Interest

DDI Drug-drug interaction

EBI European Bioinformatics Institute

FP False negative

FP False positive

GO Gene Ontology

IDF Inverse document frequency

KEGG Kyoto Encyclopedia of Genes and Genomes

NCBI National Centre for Biotechnology Information

NER Named entity recognition

NLP Natural language processing

NP Noun phrase

POS Part of speech

PPI Protein-protein interaction

RDP Relationship-depicting phrases

RelEx Relationship extractor

RNA Ribonucleic acid

SM Small molecule

SVM Support vector machine

TF Term frequency

TP True positive

UMLS Unified Medical Language System

VP Verb phrase

Publications

Parts of this work have been previously released in peer-reviewed publications, as follows:

- Rebholz-Schuhmann, Dietrich, Jee-Hyub Kim, Ying Yan, Abhishek Dixit, Caroline Friteyre, Robert Hoehndorf, Rolf Backofen, and Ian Lewin. "Evaluation and cross-comparison of lexical entities of biological interest (LexEBI)". PloS one 8, no. 10 (2013): e75185.
- Y Yan, J Hastings, J-H Kim, S Schulz, C Steinbeck, and D Rebholz-Schuhmann. Use of multiple ontologies to characterize the bioactivity of small molecules. In ICBO, 2011.
- Y Yan, S Kafkas, M Conroy, and D Rebholz-Schuhmann. Towards generating a corpus annotated for prokaryote-drug relations. In International Symposium on Semantic Mining in Biomedicine (SMBM), 2012.
- Y Yan, J-H Kim, S Croset, and D Rebholz-Schuhmann. Finding small molecule and protein pairs in scientific literature using a bootstrapping method. In Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, pages 172-175. Association for Computational Linguistics, 2012.

Where appropriate, these publications will be cited in the text in the usual way.

Annotation scheme

In this thesis a colouring scheme has been adopted to assist the reader in interpreting the numerous excerpts of biological text that will be presented. The link between colours and entities is as follows:

- Small molecule
- Protein (or enzyme)
- Organ
- Organism
- Biological process
- Gene or gene product
- **Trigger word**

Chapter 1

Introduction

This project has used text mining methods to extract small molecule bioactivity descriptions from the scientific literature.

In this introductory chapter, the context of the project will first be described. The research objectives will then be stated and there will be an overview of the work that has been previously carried out on this and related topics.

1.1 Context

Biological informatics resources have made a huge contribution to the processes of scientific research. One of the major categories of biological resources is ontologies. Ontologies define the basic terms and relationships in common use in a particular domain [1]. The usage of biological ontologies can be as a community reference, as a formal basis for interoperability between systems, and for search, integration and exchange of biological data [1, 2]. Numerous biological ontologies have been created or proposed, however the focus of this work is an ontology named ChEBI.

1.1.1 ChEBI

ChEBI [3] is an ontology of chemical entities of biological interest, created by the European Bioinformatics Institute (EBI). It describes chemical entities such as molecules and ions together with their structure and biologically relevant properties. ChEBI is primarily focused on chemical compounds which are small molecules. Within biochemistry, a small molecule is an organic compound with a low molecular weight and a size of around 1 nanometre. Most (but not all) drugs are small molecules.



Roles Classification 	
Biological Role(s):	<p>antibacterial drug A drug used to treat or prevent bacterial infections.</p> <p>antibiotic A substance that is biostatic or biocidal at low concentrations towards bacteria, yeasts, moulds, or other form of life, especially pathogenic or noxious organisms. Although the term was originally restricted to substances produced by microorganisms, its use was later expanded to include derivatives of such substances and it is now commonly used to include entirely synthetic compounds. (via peptide antibiotic)</p>
Application(s):	<p>antibacterial drug A drug used to treat or prevent bacterial infections.</p> <p>antibiotic A substance that is biostatic or biocidal at low concentrations towards bacteria, yeasts, moulds, or other form of life, especially pathogenic or noxious organisms. Although the term was originally restricted to substances produced by microorganisms, its use was later expanded to include derivatives of such substances and it is now commonly used to include entirely synthetic compounds. (via peptide antibiotic)</p>
View more via ChEBI Ontology	
ChEBI Ontology 	
Outgoing	<p>vancomycin (CHEBI:28001) has functional parent vancomycin aglycone (CHEBI:47724)</p> <p>vancomycin (CHEBI:28001) has role antibacterial drug (CHEBI:36047)</p> <p>vancomycin (CHEBI:28001) has role antibiotic (CHEBI:22582)</p> <p>vancomycin (CHEBI:28001) is a glycopeptide antibiotic (CHEBI:24395)</p> <p>vancomycin (CHEBI:28001) is conjugate base of vancomycin(1+) (CHEBI:76842)</p>
Incoming	<p>telavancin (CHEBI:71229) has functional parent vancomycin (CHEBI:28001)</p> <p>vancomycin(1+) (CHEBI:76842) is conjugate acid of vancomycin (CHEBI:28001)</p>

Fig. 1.1 ChEBI role classification for the small molecule Vancomycin

As of 2012, ChEBI consisted of around 25,000 entities, divided into a structure-based classification and a role-based classification. The structure-based classification pertains to physical characteristics, such as the entity's chemical formula and its mass. The role-based classification includes terms describing the biological activities of chemical entities, such as "cyclooxygenase inhibitor" and "immunomodulator". These terms describe small molecule *bioactivity*: the combined influence of a small molecular entity on the components of a living organism and on the organism as a whole. An extract from ChEBI showing the role classification for the small molecule Vancomycin is shown in Figure 1.1.

1.1.2 Limitations of the ChEBI ontology

The ChEBI role ontology allows the categorisation of chemical entities by their bioactivities. However, in its present form it suffers from two key limitations:

1. Role assertions are relatively sparse as compared to the full ontology of chemical entities. Fewer than 3,000 chemical entities are mapped to fewer than 500 biological roles, i. e. only ~10% of the full chemical entity ontology has a role mapping. The result is that the vast majority of chemical entities included in the ontology are not adequately described in terms of their biological context.
2. Role assertions currently only provide links to other ChEBI entries, and there is no

way to link to a primary reference source for a target of bioactivity (i. e. a biological entity). For example, the role term ‘cyclooxygenase inhibitor’ describes how a small molecule inhibits a cyclooxygenase enzyme, yet there is no link from the small molecule (in ChEBI) to an informatics reference for enzymes.

Linking ChEBI to other resources would greatly improve its functionality. The connection would serve as a bridge between small molecules and their biological targets, facilitating reasoning about the activities of chemical entities in a biological context. Moreover, chemical entities can be well characterised over more biological dimensions in relation with other biological types. A newer version of the ChEBI ontology can enhance the ability to perform semantic reasoning over different biological resources. The way in which ChEBI can be connected to other resources is through the bioactivity of the small molecules it describes.

1.1.3 Small molecule bioactivity

Guaadaoui *et al.*, after reviewing multiple sources, defined bioactivity as “a compound which has the capability and the ability to interact with one or more component(s) of the living tissue by presenting a wide range of probable effects” [4]. Therefore we can consider a bioactive small molecule as any compound which has an effect, positive or negative, on living tissue.

Bioactivity provides insights into the biological events associated with small molecules used for pharmaceutical and therapeutic purposes, not to mention understanding the pathway triggers and regulation. On a molecular level, small molecule bioactivity corresponds to the binding of the molecule to a macromolecular receptor, resulting in some observable physiological effect on the biological systems involving that macromolecule [5]. Bioactive molecules can have positive effects, such as repressing the development of disease, or they can have negative (toxic) effects, leading to illness or even death. The differentiation of bioactive molecules from non-bioactive molecules is one of the core requirements for *in silico* (i. e. computational) drug discovery approaches, as is delineating molecules which share similar activity profiles [5].

To properly formalise the description of activities of chemical entities in biological contexts requires reference to multiple terminological sources, some of which can be described as formal ontologies, whereas other ones are better characterised as thesauri, databases, or controlled vocabularies. For example, to formalise a description of enzymatic inhibition requires reference to the biological process of the enzyme and the description of the bioac-

tivity within that particular process. The bioactivity descriptions may involve reference to the exact location of the activity at a cellular level, or may instead refer to the effect on the organ or organism within which it takes place. This use of metonymy (i. e. where a target is described implicitly, through its association with another entity) is common in descriptions of bioactivity, but it means that there can be a certain amount of vagueness. This is a concept we will return to later.

1.1.4 ChEMBL

The concept of classifying small molecules by their bioactivity is not new. The ChEMBL database [6] stores bioactive drug-like small molecules together with their bioactivity information. The database contains two-dimensional structures, calculated properties, and binding and functional information for a large number of drug-like bioactive compounds. These data are manually abstracted from the primary published literature on a regular basis, then further curated and standardised to maximise their quality and utility across a wide range of chemical biology and drug-discovery research problems.

In contrast to ChEBI, chemical compounds present in ChEMBL are mostly drug-like molecules. The focus of ChEMBL is restricted to drug-like compounds and other compounds are ignored. On the other hand ChEBI is much wider in scope and features all small molecules with and without biological concern. The molecular entities in ChEBI cover both intentionally bioactive molecular entities (drugs), and accidentally bioactive entities, such as chemicals in the environment. Therefore it is preferable to use ChEBI as a small molecule collection in order to assess the bioactivity of small molecules with the broadest possible scope. However, this broadened scope makes the manual curation used for ChEMBL impractical.

1.1.5 Connection of ChEBI with other ontologies

Recall that the ultimate aim is to provide links from small molecules in ChEBI to entities in other bioinformatics resources. The first step of creating this connection is to define the nature of the relationship between the small molecule and the target entity. In order to proceed it is instructive to examine the fields already present within ChEBI records. Currently, role assertions of ChEBI entities are presented using a *has_role* attribute, which can be used to provide a link to other ChEBI entities. For example *metronidazole* **has role** *antibiotic*, where metronidazole and antibiotic both have dedicated pages within the ChEBI ontology.

Based on a heuristic analysis of bioactivity phrases in the literature, macromolecules

and biological processes have been identified as the most common types of targets for the bioactivity of small molecules. One option therefore is to introduce a *has_target* relationship to relate a bioactivity description to either a macromolecule or a biological process. Strictly speaking, the range of the *has_target* relationship should be restricted to those entities with which the chemical entity can physically interact – macromolecules. However in many cases the bioactivity will be described where the precise macromolecular target is unknown. In these cases the target will often be described as a biological process. In the same way, anatomical or subcellular locations may also be mentioned when the exact target is unknown. Therefore, the *has_target* relation can be further generalised to cover biological processes and other higher level entities such as organs and organisms.

The Manchester OWL syntax [7] provides a high level way of representing ontologies. In this case the target is a *macromolecule* and *participant_of* some *Process*. A set of examples will now be provided.

The sentence *m1 is a betaadrenergic receptor inhibitor* is a typical bioactivity description. It can be formalised as:

```
m1 subclassOf bearer_of some
(realized_by only
(Inhibition and
(has_target some BetaAdrenergicReceptor)))
```

In an example referring to a process, *m2 is a mitosis stimulator*:

```
m2 subclassOf bearer_of some
(realized_by only
(Stimulation and
(has_target some
(participant_of some Mitosis))))
```

Considering organs within the body, an example *m3 is a thyroid stimulator*:

```
m3 subclassOf bearer_of some
(realized_by only
(Stimulation and
(has_target some
(has_locus some ThyroidGland))))
```

Finally, an example which is further constrained in terms of the species is considered, *m4 is a mouse thyroid stimulator*:

```
m4 subclassOf bearer_of some
  (realized_by only
    (Stimulation and
      (has_target some
        (has_locus some (ThyroidGland and part_of someMouse))))))
```

The aim is to support the expansion of the current ChEBI ontology by identifying new types of relations to the current main ontology as well as completing the existing “role” ontology. Figure 1.2 presents how the expansion would be achieved, by attaching the new sub-ontology of bioactivity with connections to various target entities contained in external resources. The existing relation type of *has_role* enables the interlinking between small molecule entity and bioactivity; the relation type of *has_target* enables the ChEBI ontology to be linked with a target type in another domain. In that way, identifying and extracting the properties in a small molecule’s bioactivity event becomes an essential step toward constructing the sub-ontology and developing links to other resources.

Ontologies which represent suitable target resources include

- UniProt [8], a comprehensive database of protein sequence and functional information.
- Gene Ontology (GO) [9], an ontology of genes and gene products across all species.
- NCBI Taxonomy [10], an extensive database of organism names.

To manually create the links between ChEBI and these other ontologies would be an enormous undertaking, if not impossible. As we shall see, new biomedical research is published at a staggering rate and many of these projects will have resulted in the discovery of new bioactivity events. Therefore the focus is on automated approaches. This work takes the view that natural language processing of scientific publications, in other words text mining, is the most viable solution. From a text mining point of view, relation extraction is the task of finding semantic relations between relevant entities in natural language text [11]. Related work has shown that discovering or extracting relationships between biological objects from text documents is highly efficient [12]. To pursue a text mining form of relation extraction involves object identification, reference resolution and extracting object-object

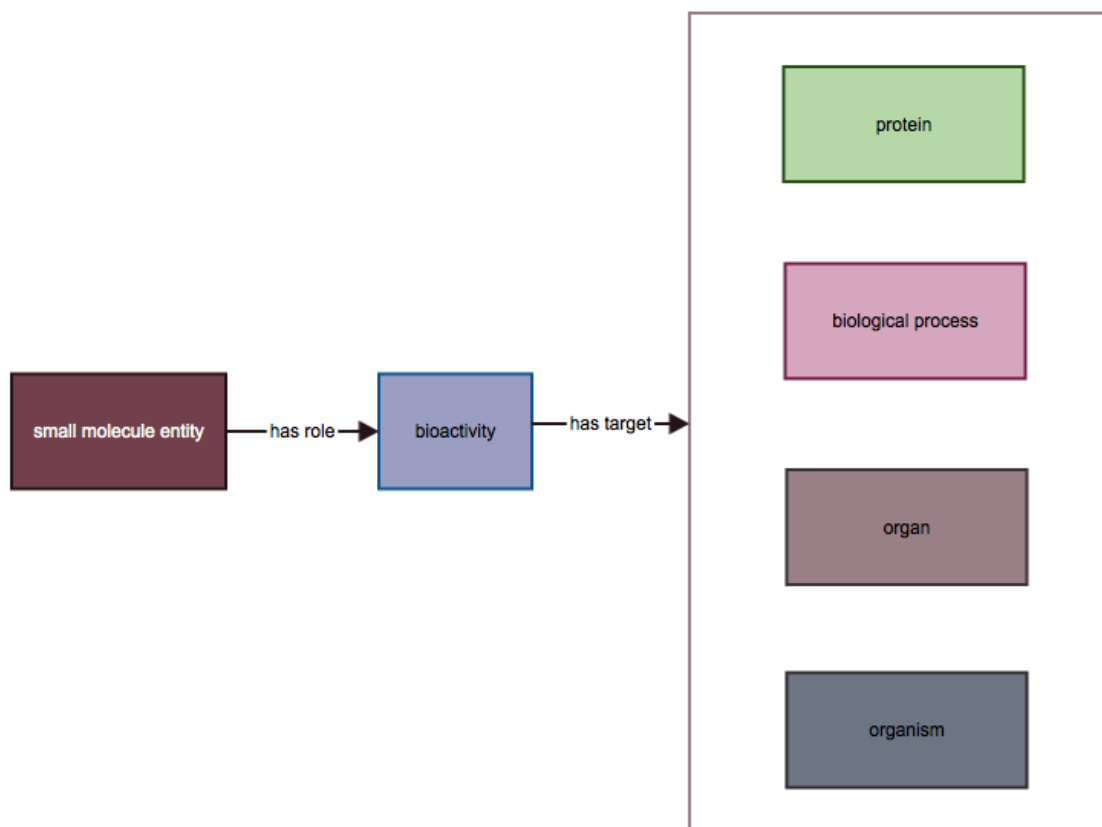


Fig. 1.2 ChEBI bioactivity ontology model. The small molecule entity is in ChEBI, and each of the four boxes on the right represents another ontology.

relationships [13]. The next section will give an introductory statement of the biomedical text mining method.

1.1.6 The use of text mining in biomedical research

Biomedical researchers nowadays require efficient means of handling and extracting information from literature and biomedical databases, because both show unbridled growth. Figure 1.3 shows annual growth of publications on PubMed [14], a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The number of articles in PubMed that were published before 1949 is 61,653, however that had grown to 23,937,995 articles published by the end of 2012. Researchers have realised that the exponentially increasing quantity of scientific literature has resulted in a problem of information retrieval and knowledge discovery. A major problem is that modern experimental methods are high-throughput, with single studies often leading to results

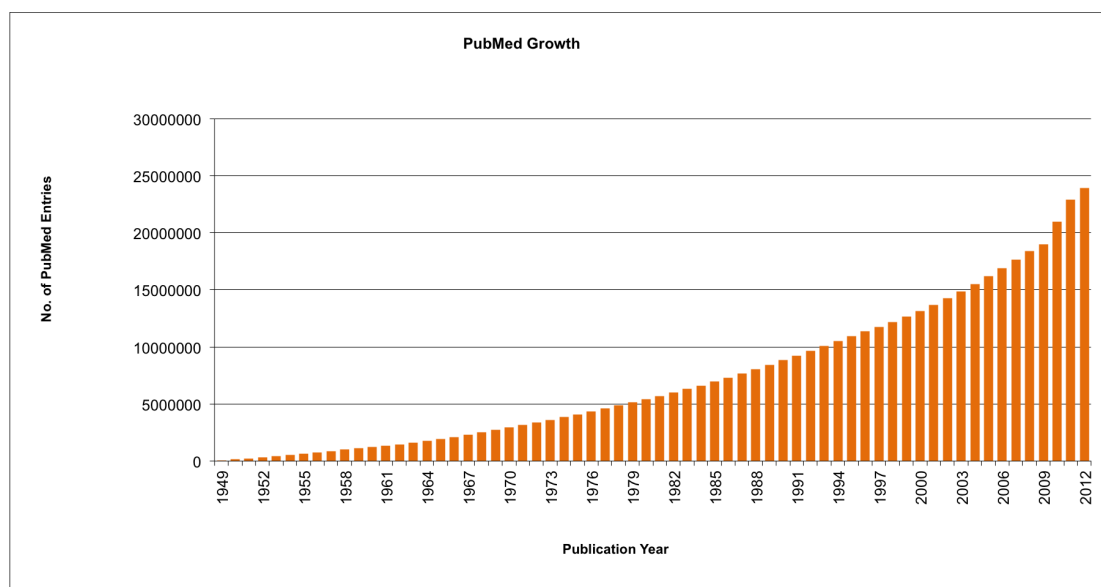


Fig. 1.3 PubMed Growth

for thousands of biological objects (e. g. genes or proteins) at the same time. These results are often only made available in free text (scientific publications). Results are not entered into structural databases and searching for the information later is incredibly difficult, which means that information can become 'hidden' [15].

In contrast to information retrieval, which relates only to the existence and whereabouts of information, the aim of text mining is the discovery of new, previously unknown information. As described by Manconi *et al.* [16], text mining systems analyse natural language textual data and detect lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information.

Biomedical text mining is a field of research which refers to the application of text mining methods to texts and literature of the biomedical and molecular biology domain in order to extract useful information. New approaches for knowledge discovery are highly sought after by the scientific community in order to guide research and efficiently provide relevant background information [17].

Previous research has demonstrated some successes in using computational text mining to automatically transform scientific publications from raw text into structured database content. The resulting content may be combined or integrated with other knowledge resources, potentially leading to the formation of novel hypotheses and new discoveries. For example, information extraction and textual data analysis could be particularly relevant and helpful in genetics research, in which up-to-date multi-domain information about complex

processes involving genes, proteins and phenotypes is crucial [18].

1.1.7 The presentation of bioactivity in biomedical text

With the bioactivity concept defined, investigation into the presentation of bioactivity events in scientific text can be carried out. The term “biomedical event extraction” is used to refer to the task of extracting descriptions of actions and relations among one or more entities from the biomedical literature [19]. As previously described, bioactivity is an event induced by a chemical entity exposing its activity towards a biological entity or structure. One chemical entity can induce different bioactivities at different scales, which may be linked to each other. For example, a chemical entity blocking the function of a protein in the coagulation cascade is expected to block the coagulation cascade overall.

According to the biomedical text annotation guidelines associated with the “gold standard” bioinformatics corpus GENIA [20] (to be described in more detail later), an event textual presentation includes a set of properties, such as *cause*, *target* and *theme*. This heuristic provides a framework with which to identify and annotate bioactivity events from text. In GENIA terminology, bioactivity is an event that is induced by an agent (a small molecule in this context), the *cause*, which is acting on a *target*. The target may be from a diverse set of biological types, for example proteins, cells, organisms, and biological processes. The target may also be another small molecule. The *theme* describes the nature of the interaction. Such bioactivity events can have quite complex representations and verbal expressions, but these annotation guidelines are mainly focused on the identification and annotation of specific trigger words (sometimes known as interaction words) which provide a clue to the type of bioactivity event. The trigger word provides the link between the small molecule agent and the biological target. For example, common trigger words are “inhibit”, and “stimulate”.

It is important to cater for morphology of the trigger word. The triggers can be in verb or noun forms, e. g. “inhibit” or “inhibitor”. Nominalisations of the verb forms, e. g. “activation” for “activate”, are relevant, together with negative prefixes e. g. “inactivate”.

The trigger words can be classified as terms of positive regulation (e. g. “activate”) and negative regulation (e. g. “block”), but neutral forms or regulatory trigger forms occur as well (e. g. “modulate”).

The combination of the small molecule, the biological target, and the trigger word forms the complete bioactivity description. We will denote this a *bioactivity triple*.

Once the annotation of bioactivity has been completed, additional contextual informa-

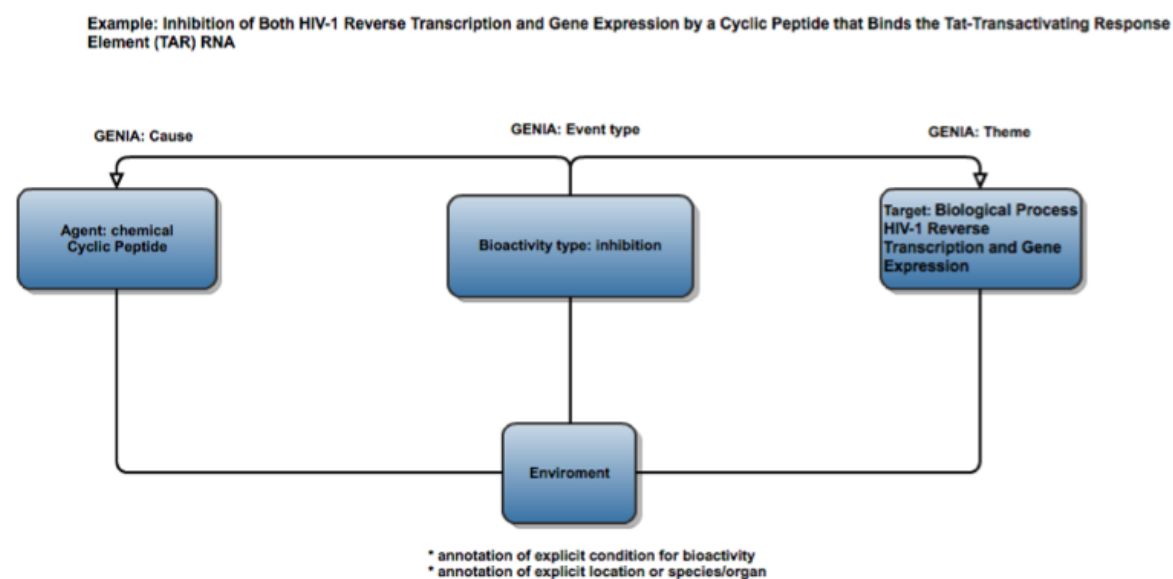


Fig. 1.4 Example bioactivity annotation

tion could be relevant for the full specification or description of the bioactivity event. Such additional information is not essential, but contributes additional information about the experimental context in which the bioactivity is measured or reported. For example, in a sentence containing the statement

“*Caffeine inhibits the antidiuretic hormone (ADH) in adult mice*”

the agent (small molecule) is *caffeine*, the bioactivity trigger word is “inhibits”, the target is the antidiuretic hormone *ADH* in the environment *adult mice*.

Figure 1.4 provides an example illustrating a full bioactivity annotation and its relationship to the GENIA annotation scheme from which the scheme was derived. The diagram shows the annotation result for the following article title:

“*Inhibition of both HIV-1 reverse transcription and gene expression by a cyclic peptide that binds the Tat-transactivating response element (TAR) RNA*”

In this case, *cyclic peptide* is the cause of the biological process and *HIV-1 reverse transcription and gene expression* is the bioactivity target.

A simple approach would be to search for the co-occurrence of relevant entities in a sentence. This would yield many results however there will inevitably be a large number of errors (noise). More sophisticated approaches based on natural language processing analyse the semantic relationships between terms in order to improve the signal-to-noise ratio. An overview of natural language processing techniques will be provided in Section 1.3.

1.1.8 Challenges

These are some of the issues that cause difficulty in using text mining to extract small molecule bioactivity.

Explicit vs non-explicit triggers

One way of looking for bioactivity is to use the verb as a trigger word. For instance, finding the verb “resistant” co-located with a small molecule entity is likely to indicate the presence of bioactivity. This can be thought of as an explicit bioactivity trigger. However, some other keywords are less reliable. The word “against” is often used to denote bioactivity, but this is also a relatively common word that could be used in many other situations. Extracting all instances of this term would lead to many incorrect results (i. e. low precision).

Use of metonyms

There is obviously an enormous amount of variation in natural language, even for tasks such as scientific writing which favour precision and conciseness. One of the ways in which this is expressed is in describing contextual information. In bioactivity descriptions this is often seen as metonymy: where the target can be expressed in a precise way (as a chemical binding site, for instance), or in an imprecise way, by talking about the effect on the organism as a whole.

For instance compare the following two sentences, which are both valid examples of small molecule bioactivity:

*“**Penicillin** did not **inhibit** the **recovery process**”*

*“The ability of some **N,N-dialkylaminosubstituted chromones and isoxazoles** to **inhibit** the protein **kinase C (PKC)** dependent signal transduction pathway was tested”*

This clearly places huge demands on detection of the entities within the sentence.

Negation

An added complexity is the issue of negation. For instance it will often be the case that a single word completely changes the nature of a bioactivity event. It is critical that the negation cue is identified. For example consider the following sentence:

“The cellulase component was not markedly inhibited by most metal ions tested.”

Detecting the word “not” in the vicinity of the bioactivity trigger “inhibited” is relatively straightforward. However there are clearly more subtle forms of negation. Consider the following:

“Without influence on WDS were: physotigmine, atropine, ganglionic- or adrenergic-blocking drugs, Dopa, MAO-inhibitors, serotonin- and histamin-antagonists and nonnarcotic analgesics.”

Here the word “without” means that there is no positive example of bioactivity in this sentence, despite it containing all the necessary components. It is necessary to develop patterns or rules to handle these cases.

1.2 Research objectives

Given the above context, the primary aim of this work was to develop methods of extracting bioactivity information from the scientific literature. Following preliminary research, this aim was decomposed into the following objectives:

1.2.1 Use of explicit and non-explicit verbs in a specific case of small molecule bioactivity

A list of verb trigger words can be used as an indication of bioactivity. Often these verbs will be used in a way that explicitly denotes a bioactivity event. However sometimes the verb will be used in a way that does not denote a bioactivity description. In order to handle this type of use, it is necessary to pre-filter the data. Hence techniques are required to perform intelligent pre-filtering of documents and sentences before attempting pattern-based extraction. The approach adopted here is the use of supervised machine learning algorithms (namely a support vector machine) to perform classification.

As a way of helpfully constraining this aspect of the work, there is a focus on a specific type of bioactivity: that between drugs and bacteria. In other words, this experiment studies vagueness in the bioactivity type, but not the target.

1.2.2 General small molecule bioactivity on a wide range of possible targets

In contrast to the above task, the second uses only explicit verbs but expands the scope to consider all types of small molecule bioactivity. This results in a high degree of vagueness,

particularly in the targets of the bioactivity. One way in which this is most often seen is through the use of metonymy: using one concept to refer implicitly to another one.

With such a broad scope it is not feasible to develop the training corpus required for a supervised machine learning solution. Hence the work follows a purely rule-based method of finding bioactivity descriptions in the literature.

1.3 Biomedical text mining - tools and techniques

Text mining has been widely used to support biological and medical research [21]. Biomedical text mining is a subfield of natural language processing, which deals with text from biology, medicine, and chemistry, where the text mining applications have to be tailored to specific types of text. Using computational methods to understand biomedical text could enable a vast number of biomedical publications to be processed with exponentially increasing speed. For instance, text relating to proteins has its own nomenclature. In order to machine-process this text it requires a standardised algorithm to be developed that recognises protein mentions from natural text. When extracting information relevant to the protein, such as protein interactions, protein disease association, and biological events that involve one or multiple proteins, the information extraction algorithm has to be designed appropriately. This may involve training to provide the ability to precisely localise extraction targets.

A lot of effort has been spent on developing specific tools for different biological text mining tasks, and the techniques generally fall into three main categories in terms of layers of text mining: named entity recognition (NER), relation extraction, and event extraction. The following is an overview of these approaches and their application to some typical biomedical text mining tasks.

1.3.1 Named entity recognition

Terms (names used in a specific domain) and terminology (a collection of terms) constitute important building blocks in biomedical text [22, 23]. In the biological and biomedical field, there has been an increasing amount of research on biomedical named entity recognition (NER). Biomedical NER has become the most basic problem in automatic text extraction [24]. A particular issue is the extremely large vocabulary, taking in specialised terms from fields as disparate as cell types, proteins, medical devices, diseases, gene mutations, chemical names, and protein domains [25], together with organs, organisms, biological processes, and many others. There are also multiple complications relating specifically to bio-

medical writing, for instance the fact that biological objects often have names consisting of multiple words which can often make it unclear where a name starts and ends, even for human readers [15]. New biological objects (e. g. genes and proteins) are discovered at such a rapid pace that naming conventions have not caught up, and so there is a lack of consistency in how objects are named, and extensive use of names containing mixtures of letters, numbers, and special characters. Use of abbreviations also causes problems, for instance the abbreviation ACE can refer to all of ‘angiotensin converting enzyme’, ‘affinity capillary electrophoresis’, ‘acetylcholinesterase’, and other things [15].

The following examples outline different NER techniques in terms of types of entity and extraction methods.

Whatizit

Whatizit [26] is a text mining tool containing NER modules developed in the European Bioinformatics Institute. It is primarily a web service tool which can be used to identify several different types of terms. Any additional vocabulary in the range of up to 500k terms can be easily integrated into a Whatizit pipeline. Previously integrated vocabularies are Swissprot, the Gene Ontology (GO), the NCBI taxonomy, and Medline Plus [26]. NER components in Whatizit are mainly dictionary-based or rule-based. The NER service from Whatizit satisfies the need for terminology-driven feature extraction from text, leading to some successful cases of downstream text mining tasks such as document classification and relation extraction [26]. Moreover, the integrated modules automatically link annotations to a database to offer additional support. Figure 1.5 shows a screenshot of the Whatizit web-interface when a passage of text is annotated. As well as the web interface, the back-end processing of Whatizit can also be used in a non-interactive manner.

ABNER

Another famous NER tool which has been highlighted in the field is ABNER [27]. ABNER is a standalone application which recognises five types of biological terms: DNA, RNA, cell line, cell type, and proteins. Similar to Whatizit, ABNER has optional built-in tokenisation and sentence segmentation algorithms. It is robust to wrapped lines and biomedical abbreviations. The core of the system is a machine learning algorithm based on conditional random fields. It is designed to maximise domain independence by not employing brittle semantic features or rule-based processing steps. Figure 1.6 gives an annotation example from the ABNER interactive user interface. The system takes a passage of input text and produces

Resulting tagged text

Diabetes represents one of the main chronic diseases worldwide . Diabetes and its associated complications may be detectable even at early stages in the urinary proteome . In this article we review the current literature on urinary proteomics applied to the study of diabetes and diabetic complications . Further, we present recent data that strongly indicate urinary proteome analysis may be a valuable tool in detecting diabetes-associated pathophysiological changes at an early stage, and also may enable assessment of disease progression and efficacy of therapy . Current data indicate that collagen-derived peptides represent one of the main peptidic components in urine, which are consistently found at reduced levels in diabetes . It is tempting to speculate that this decrease in urinary collagen-derived peptides is related to an increase in extracellular matrix deposition which is a major complication in diabetes . Therefore, urinary proteome analysis might enable noninvasive assessment of this process at an early stage via determination of specific collagen fragments . This may open an avenue towards targeted therapeutic intervention .



Fig. 1.5 Whatizit web-based annotation tool. Various biological concepts and entities have been highlighted by the tool. The colours used in the annotation depend on the dictionaries and options selected at run-time.

simultaneously recognised multiple named entities highlighted in different colours accordingly. ABNER includes routines for training on new corpora, together with a programming interface to incorporate ABNER into customised biomedical text applications [28].

NER of chemical entities

Named entity recognition of chemical entities has received significantly less research effort than NER for genes or proteins. It has been argued that chemical NER is even more challenging, because chemical names often use special characters, commas, spaces, hyphens, and parentheses.

Jochem [29], or the Joint Chemical Dictionary, is a dictionary-based NER tool to identify small molecules and drugs in free text. It combines information multiple sources, including ChEBI. In total it contains over 2.6 million terms. Performance was found to be acceptable on a test corpus of 100 MEDLINE abstracts, with precision of 67% and recall of 40%.

Some of the problems of dictionary-based systems (including Jochem and Whatizit) is that creation and maintenance of the dictionaries can be extremely labour intensive, and typically they can struggle to deal with natural spelling variations or misspellings in natural text. Solutions to this problem attempt to introduce ‘fuzziness’ to the tagging process. Machine learning techniques (which can be supervised or unsupervised) attempt to automatically detect relevant entities based on semantic or semantic features, or other contextual information. This means that machine learning NER approaches can also detect previously unknown entities.

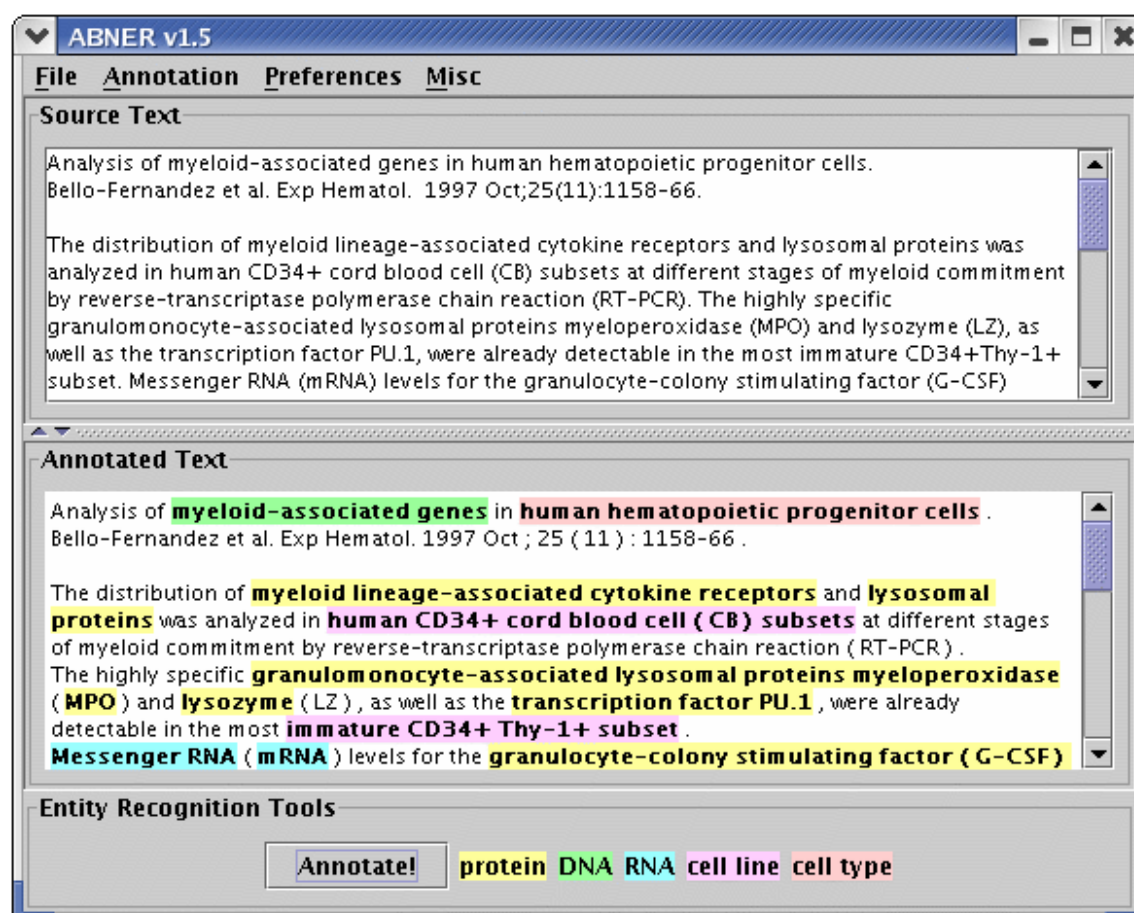


Fig. 1.6 The interface of ABNER NER. The source text in the upper window has been fully semantically annotated in the lower window. The colour coding can be seen near the bottom of the interface.

OSCAR [30] was also developed using information derived from ChEBI, but in contrast to Jochem, OSCAR is a hybrid approach that includes a machine learning element. The most recent version, OSCAR4, is claimed to be the state-of-the-art in the chemical text analysis. It features both a regular-expression pattern engine and a machine learning element implemented using a maximum entropy Markov model. Performance is generally impressive though it is recognised as being domain-dependent and varies with different styles of text.

1.3.2 Corpora

A corpus is a relatively small collection of texts that is designed to statistically represent a much larger collection. Such a corpus will usually be precisely annotated by hand, and provides reference material for algorithm development and statistical analysis. There have been several corpora published for use in bioinformatics applications. Three of the most well known will be described here.

GENIA

GENIA [20] is a widely used corpus in the field of biomedical text mining. GENIA contains over 9,000 annotated sentences taken from 1,000 paper abstracts. Various entities are annotated, including entities, terms, and sentence syntax.

Some useful tools have been created using GENIA. GDep [31] is a parser trained on the GENIA corpus. GDep will process text and perform dependency parsing, stemming, and part-of-speech tagging. These processes, which will be described in more detail later, are essentially used as a way of reducing and interpreting the huge amount of variability seen in natural language.

BioInfer

The BioInfer [32] corpus was developed to support information extraction in the biomedical text mining domain. It consists of sentences from 200 PubMed abstracts that contain interacting proteins as previously identified in the Database of Interacting Proteins. It contains 1,100 sentences with annotation of named entities, relationships, and syntactic dependencies. A further 30 negative abstracts were also added (i. e. with no PPI relation) to balance the corpus. It is primarily targeted at protein, gene, and RNA relationships.

AIMed

AIMed is a corpus created by Bunescu *et al.* [33] for PPI method comparison which has been widely used. It consists of 225 manually annotated abstracts from MEDLINE. Of these, 200 are known to describe interactions between human proteins, and the remaining 25 are known not to. As a further feature, of the 200 positive abstracts, 177 contain PPI within a single sentence. There are 4,084 protein references and around 1,000 tagged interactions in this data set.

1.3.3 Parsing

After NER, the subsequent processing step is to recognise the relationships between and among the identified entities. Parsing can take various forms. The concept is to identify semantic or grammatical features about the words in a sentence.

Perhaps the most common form of parsing is part-of-speech tagging. This performs a shallow parse of the sentence and will tag each word based on its part of speech and its context. Part of speech refers to common grammatical definitions such as verb, noun, adjective, preposition, et cetera. POS taggers can also work at a slightly higher level and reveal whole clauses, or noun phrases for example. In this work, POS tagging was performed by GDep, Enju, and a commercial tool from CIS [34].

Dependency parsing is a computationally expensive process that analyses the content of sentences and aims to reveal non-local dependencies. It does so through deep analysis of sentence syntax and semantics. This can be particularly useful in processing biomedical writing where sentences can be long and complicated, and frequently mention a number of relevant agent entities [35]. Dependency parsing allows processing to focus on only the terms connecting the entities of interest, and ignore other terms in the sentence that are irrelevant for the particular information extraction task being undertaken.

A useful comparative study of parsers for biomedical event extraction is available in [36].

1.3.4 Relation extraction

The relation extraction problem is to recognise a piece of text containing a semantic property of interest [37], and potentially extract the relevant items. This is a challenging task. The simplest method of relation extraction employs a co-occurrence based approach, identifying when two related entities commonly appear near each other in the text. Such an approach would be expected to find all relations but with a very low signal-to-noise ratio,

i. e. maximum recall but low precision.

The objective of machine learning and rule-based methods for relationship extraction is to use more sophisticated algorithms to improve on the results produced by co-occurrence techniques.

1.3.5 Linguistic patterns

Once a description of a bioactivity event has been found, the next task is to extract the relevant pieces of information and ignore the irrelevant words in the surrounding context. This can be done using pattern matching. The patterns are sometimes known as lexico-semantic patterns, or linguistic patterns [38]. The context around the trigger word has been called a concept node.

The patterns will depend on the verb which triggers the detection of bioactivity. For instance a passive verb will have a different sentence order to an active one, and will also influence whether a preposition is present or not. As highlighted above, special consideration must be given to negation. Speculation may also be a concern: this refers to writing that describes the possibility of an event occurring. An example sentence might be:

“In the future we will design an experiment to evaluate whether p9-RNA is also able to activate the RNA-dependent protein kinase (PKR) of human lymphocytes”

Ideally this sentence should be rejected by a bioactivity extraction system because it does not contain a valid bioactivity description. However, detecting this in a general way is particularly challenging.

1.4 Related work

The following section will review related projects and discuss where there are similarities and differences. Related projects will also be treated in later chapters to highlight specific connections to the work being presented here.

1.4.1 Extension of bioinformatics resources

Several previous studies have attempted to unify bioinformatics resources. With a similar motivation of extending ChEBI, STITCH [39] is a resource to explore known and predicted interactions of chemicals and proteins. The linkage evidence of chemicals and proteins is

derived from databases and the scientific literature. Initially, STITCH integrates information about interactions from metabolic pathways, crystal structures, binding experiments and drug-target relationships. Inferred information from phenotypic effects, text mining and chemical structure similarity is used to predict relations between chemicals [39]. In the later developments of STITCH, the number of chemicals and proteins was increased by integrating information from other databases, such as BindingDB [40], PharmGKB [41] and the Comparative Toxicogenomics Database [42]. The expanded resource enabled by the interaction network can be explored interactively or used as the basis for large-scale analyses. So far, STITCH contains interactions of over 300,000 chemicals and 2.6 million proteins from 1,133 organisms [43].

Another similar solution for exploring small molecule and target relations, SuperTarget is a resource which reports drug-related information regarding medical indication areas, adverse drug effects, drug metabolism, pathways and Gene Ontology terms of the target proteins [44]. The resource provides an easy-to-use query interface allowing the user to retrieve drug-related information, including aspects of finding drugs that target a certain pathway, drug screening and target similarity inclusion. SuperTarget currently contains over 6,000 target proteins, which are annotated with over 330,000 relations to 196,000 compounds [45]. SuperTarget has the ability to provide a profile of a drug by combining information pertaining to the drug and its target. However the resource is only focused on drugs, a subset of small molecules.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [46] is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information. The genomic information is stored in the GENES database [47], which is a collection of gene catalogues for all completely sequenced genomes and some partial genomes with up-to-date annotation of gene functions. A third database in KEGG is LIGAND, which provides information about chemical compounds, enzyme molecules and enzymatic reactions. In accordance with new chemical genomics initiatives, the scope of KEGG LIGAND has been significantly expanded to cover both endogenous and exogenous molecules. Specifically, RPAIR contains curated chemical structure transformation patterns extracted from known enzymatic reactions, which would enable analysis of genome-environment interactions, such as the prediction of new reactions and new enzyme genes that would degrade new environmental compounds. Additionally, drug information is now stored separately and linked to new KEGG DRUG structure maps [48].

1.4.2 Protein-protein interaction

Protein-protein interaction has been described as the most extensively studied information extraction problem in the biomedical domain [49]. The aim of PPI text mining is to find new or updated information on interaction between proteins or protein-coding genes. A PPI extraction system would be expected to return instances featuring two protein names and an interaction type. This is therefore similar in concept to small molecule bioactivity, but significantly narrower in scope due to the agent and target being of the same semantic type. However many of the techniques are relevant and instructive for this work.

Lee *et al.* [49] developed a rule-based system for PPI. They argue that for this application, precision is significantly more important than recall. Their system was designed with this in mind, rejecting machine learning approaches because they are viewed as having a negative impact on precision, and produced very specialised patterns for PPI extraction. The resulting system did indeed demonstrate high precision (sometimes >95%) but recall was considerably lower than baseline techniques. It is debateable whether this is an acceptable trade-off.

He *et al.* [50] developed a system to mine protein-protein interactions from MEDLINE abstracts. They used a combination of co-occurrence and NLP algorithms, which they call a frame-based approach. It uses the co-occurrence technique to find (protein-coding) genes related to a gene of interest, and then computational linguistics to extract the semantic descriptions of interactions. It also checks PPI information against databases of known interactions, in order to distinguish between novel and previously-known events.

A proliferation of basic research on proteins and their biological processes has resulted in a high demand for text mining systems and methods. The authors of [51] developed a text mining framework for extracting gene and protein interactions from the literature. The framework is designed to be flexible in its uses, the authors' own demonstration of its capabilities was PRISE (protein interaction search engine), a tool used to extract protein-protein interaction (PPI). PRISE is an end-to-end system which comprises a querying site and a way of displaying results. It incorporates several other tools including NER, machine learning, and dependency parsing. In this sense it is conceptually similar to the approach that will be presented in Chapter 2.

Another important relation extractor developed for protein-protein interaction is RelEx [35]. RelEx comprises a rule-based core and detects the candidate proteins which potentially interact with each other in text. RelEx was developed for the identification of interactions between genes and proteins in terms of metabolic and signalling pathways. The system has the capability to extract relations from free text based on natural language processing,

producing dependent parse trees and applying a small number of rules to these produced trees. RelEx collects potential relations by extracting paths connecting pairs of proteins from dependency trees. Three rules then are applied to the trees. These rules are meant to reflect the constructs that are most frequently used in the English language for describing relations, namely:

1. effector-relation-effectee ('A activates B')
2. relation-of-effectee-by-effector ('Activation of A by B')
3. relation-between-effector-and-effectee ('Interaction between A and B')

The performance of RelEx was evaluated against the LLL-challenge data set. It found 85% of the results returned by a co-occurrence approach but with significantly higher precision. The explanation given was that these numbers correspond to inter-annotator agreement for the recognition of gene names and biomedical annotations. Error analysis of the results also concluded that the performance of RelEx heavily relied on publicly available pre-processing tools.

The RelEx program greatly surpassed previously obtained results, especially when applied to more stringent requirements such as finding the direction and type of the interaction rather than just the interaction itself.

A related field to PPI, though less well-researched, is drug-drug interaction (DDI). Munger & Teredesai [52], taking advantage of PPI extraction systems and a well formatted training data set from a contest to evaluate the converted system, developed a DDI detection engine using support vector machines (SVM). Performance was found not to out-perform other state-of-the-art examples, but DDI is considered a particularly difficult problem.

1.4.3 BioNLP

BioNLP [53, 54] is a community initiative to collaborate on biomedical natural language processing tools and techniques, developed by the University of Colorado. In order to motivate researchers to contribute to biomedical text processing, BioNLP announced several series of shared tasks: research problems published in the community and designed to be tackled simultaneously by different groups all using a common data set.

There have been several notable results from past BioNLP activities. For instance, in 2009, a BioNLP shared task relied on GENIA [20], a corpus containing PubMed abstracts focused on transcription factors in human blood cells. Two sets of event extraction tasks

were proposed, Epigenetics & Post-translational Modifications (EPI) and Infectious Diseases (ID).

FAUST [55] is an event extraction tool developed in 2011 for the BioNLP shared task. It used a simple model combination strategy on two competitive systems, producing a new system which delivers better and more accurate results. The stacking model, which is the backbone of the system, was originated for another event extractor, UMASS [56]. Among the competitors in BioNLP 2011, the FAUST system ranked in first place in three out of four tasks: first place in GENIA Task 1 (56.0% F-score) and Task 2 (53.9%), first place in the Infectious Diseases track (55.6%), and second place in the Epigenetics & Post-translational Modifications track (35.0%) [55].

The Stanford event extraction system [57] is a well-known event extraction tool which was released in a BioNLP shared task. The tool shows competitive performance, obtaining third place in the Infectious Diseases task (50.6% F-score), fifth place in the Epigenetics & Post-translational Modifications task (31.2%), and seventh place in GENIA (50.0%). The event parser includes three main parts, as is typical of event extraction routines. They are: an anchor detection to identify and label event anchors; event parsing, to structure events by linking entities and event anchors; and event re-ranking to select the best candidate event structure.

1.4.4 Summary

To conclude, the different resources outline different semantic relations between small molecules and other types. Table 1.1 summarises the characteristics of some of the resources presented above. As can be seen, the projects facilitate small molecule profiling by introducing interacting partner proteins. The discriminators between the projects are the biological concerns. Each project has focused on a specific type of interaction (e. g. interactions between small molecules and proteins), rather than looking at the more general concept of small molecule bioactivity. In addition the above projects have resulted in new databases, rather than providing interconnectivity between existing ones. At a time of “information overload” for biomedical researchers, the value of creating entirely new databases can be questioned. It is concluded that the proposal to examine bioactivity is both novel and offers value to the scientific community.

Having investigated the various approaches for biomedical text mining and selected examples of biomedical text mining applications, we see that biomedical text mining falls into a general problem of extracting information from a vast quantity of textual data. We have

Table 1.1 Similar projects investigating links between chemical and biological entities

Project	Chemical entity	No of chemical entities	Biological entity
STITCH	Small molecule	300,000	Protein
SuperTarget	Drug	196,000	Protein
KEGG	Small molecule	17,091	Protein
ChEMBL	Drug-like bioactive compound	1,324,941	Drug target protein

seen that the general topic of text mining has drawn significant research attention, but often with a focus on other applications. The reason for this can be explained by considering where there is most demand for text mining services. Development of text mining systems heavily relies on the availability of an annotated corpus used as training data. When this has been made available to the community, the results can be more robust and give higher accuracy. This is because the annotated corpus is the reference for extraction work, and the modelling of events can be simulated when the event morphology is analysed. Secondly, the availability of a corpus enables easy assessment of the system.

At present, since a great deal of scientific research and funding is directed into events involving genes and proteins, naturally this has resulted in the production of corpora and other text mining tools which has in turn led the direction taken by much of the biological text mining community. Hence most event extraction systems seen to date were built to achieve the target of extracting gene/protein related events. Consequently it has been found that certain areas have not been well explored, such as the chemical domain and the phenotypic domain.

1.5 Contributions and thesis organisation

In the course of this research, the following contributions have been made and will be described in this thesis:

- A concept of small molecule bioactivity will be presented which integrates four types of biological target, enabling the creation of linkage between the chemical domain and other biological domains.
- A supervised machine learning method will be described for classification of documents and sentences according to the detected presence of bioactivity between drugs and bacteria.

- A rule-based approach will be described for extraction of small molecule-target relations and small molecule bioactivity events.
- An annotated corpus developed for training supervised machine learning classifiers will be presented. The corpus, known as the DrugPro corpus, focuses on bioactivity relationships between drugs and prokaryotes (the class of organisms which includes bacteria).
- An experimental semi-supervised machine learning system was developed to find interacting pairs of proteins and small molecules. The work used a bootstrapping technique to self-generate language patterns based on a small number of high quality seeds. As such, it did not require an extensive training corpus. The work will not be described in detail in this thesis but an overview has been published elsewhere [58].

These points will be described in the following chapters. Following this introduction is Chapter 2, which presents a hybrid approach for extraction of drug-bacterium relationships from the literature. The approach uses both supervised machine learning to classify documents and sentences, and a collection of relatively simple language patterns for event extraction. The chapter will provide information on the algorithms used to extract the relevant information, and details the development of the DrugPro corpus. Chapter 3 describes a rule-based approach to extract general small molecule bioactivity events from the literature. Chapters 2 and 3 are primarily concerned with methodology, and both chapters provide a brief summary of related work as a means of setting out the technical background and providing justification for the chosen approach.

Chapter 4 presents and compares the results returned from both the drug-bacterium study and the general bioactivity study. Chapter 5 is a summary of this thesis, including the conclusions that have been drawn from the research.

1.6 Chapter summary

The current limitations of the ChEBI ontology, together with the difficulties that can be faced when using it, motivate the idea of extracting bioactivity events from text. Extracting small molecule bioactivity would make a contribution to the future ChEBI ontology in the direction of interrelating chemical entities and biological objects and processes. The rationale of using text mining techniques to extract small molecule bioactivity events from the scientific literature has been discussed. The research has selected two directions, one a

focused study of drug-bacterium bioactivity, and the other a much broader study into small molecule bioactivity in general. There was an overview of the tools and techniques available in developing text mining applications. The drug-prokaryote study will be presented in the next chapter.

Chapter 2

Drug-bacterium bioactivity events

Relation extraction is the process of finding small molecule bioactivity in the literature and extracting the key components: the bioactivity triple. In this chapter, a machine learning based technique for relation extraction will be presented. The work presented in this chapter fits within the wider context of small molecule bioactivity, but with a focus on a specific class of small molecule: drugs, and their effects on bacteria. This includes the field of antibiotics, a topic which is very widely researched and which we can expect to see highly represented in the literature.

By limiting this task to drug-bacterium relationships, the target of the bioactivity is constrained to be a specific entity type (bacterium). This allows us to focus on vagueness in nature of the bioactivity: i. e. the verb we take to be an indicator of bioactivity may or may not actually be an expression of bioactivity.

In order to handle this vagueness, a machine learning technique is adopted. The aim is to perform intelligent filtering of the documents in order to locate sentences containing bioactivity. The relationship extraction method uses cascaded classifiers to first select the documents likely to contain drug-bacterium relationships, then secondly to find the sentences containing the description of the relationship. In both cases, the classification is via supervised machine learning implemented using a support vector machine (SVM). Once a set of likely-positive sentences has been extracted, a set of keyword-based patterns are used to extract the bioactivity triple. The overall system is therefore a hybrid of a machine learning filtering step and a rule-based relationship extraction step.

In order to train the classifiers, a new corpus was created. The corpus, known as Drug-Pro, was developed in collaboration with a team in EBI, specifically for relationship identification between drugs and prokaryotes. (Prokaryotes are single celled organisms; bacteria are a type of prokaryote.) The DrugPro corpus is made up of nearly 400 abstracts and has

been fully annotated for training classifiers for this and similar tasks. It is expected the corpus will be useful for the text mining community in development of tools for biomedical NER and relation extraction.

In this chapter, a technical survey will first summarise the state of the art in this specific field of relation extraction. Following that, the cascaded classification approach will be presented. There will also be a description of the development of the DrugPro corpus.

2.1 Technical background

In the presence of a training corpus, approaches to relation extraction based on supervised machine learning techniques become available. According to many related studies, supervised machine learning approaches are often modelled as classification [59]. Various algorithms are used for relation extraction, such as Bayesian network and maximum entropy methods, not to mention support vector machine (SVM) methods.

Machine learning approaches for relation extraction can be divided into two groups: feature-based methods and kernel-based methods [60]. Feature-based methods are based on standard kernels, in which each data item is represented as a vector derived from a set of features generated from the training data set [61]. Feature selection can be used to identify the most discriminative features which are then chosen to train the system. The features can be of semantic or syntactic type, and serve as a cue for the system to decide whether the entities in a sentence are related or not. There are many types of syntactic features which can be computed, including:

- The semantic entities in the sentence.
- A word sequence between two related entities.
- The number of words between two related entities.
- The path in the parse tree containing the two related entities.
- Part-of-speech (POS) tags of the entities.
- N-grams (i. e. contiguous sequences).

The most prevalent representation of text for machine classification is the bag-of-words vector (BOW) [62], also known as the vector space model. In the standard BOW representation a document corresponds to a vector of values, where each value represents a weight

associated with a feature of the document. Representing documents in this way facilitates mathematical analysis; for instance the similarity of two documents can be quantified by calculating the angle between their vectors. This representation is widely used in methods of document classification, where the frequency of each word is used as a training feature for a classifier [63].

Feature-based approaches for relation extraction have become popular in the biomedical text mining domain [64]. Jiang & Zhai [12] presented work using a combination of features of different levels of complexity and from different sentence representations. Together with task-oriented feature pruning, the system achieved very promising performance. McDonald *et al.* [65] present a simple two-stage method for extracting complex relations between named entities in biomedical text. They create a graph from pairs of entities that are likely to be related at the first stage, and the second stage scores maximal cliques in that graph as potential complex relation instances. A strength of the work was its feasibility to adapt in any binary relation classification, potentially achieving high accuracy results.

In the biomedical text mining realm, the SVM has been applied to solve relation extraction problems. Kafkas *et al.* [66] developed a protein-protein interaction (PPI) pipeline to identify the variability of protein isoform interaction by extracting PPI pairs from MEDLINE. One example is the work by Kafkas for protein splice variants mentioned above, which aimed to extract and identify PPI pairs from MEDLINE. The core of the system was a set of SVM classifiers. The system performs two levels of classification: document classification to select related document abstracts that contain PPI information, and sentence classification to select sentences containing interacting protein pairs. The performance of the Kafkas system showed a good result using an SVM classifier and outperformed other state-of-the-art approaches, obtaining F1-measure of 54.20% cross-validation performance [66].

A second study of PPI extraction with SVMs is the system developed by Mitsumori *et al.* [67]. The authors did not attempt any syntactic or semantic analysis of the text, simply using the words around the protein names to create bag-of-word features for the SVM. An F-score of 48% was achieved on the AImed corpus.

Another study which aimed to extract protein-protein interaction from biological text using SVM is [68]. The system has the ability to combine features from two parsers with very different models in order to extract good features for the tree kernels in SVM^{light}, a method which automates the discovery of rules for PPI extraction. As result, the group achieved an F-score of 69%.

Taking inspiration from the related work presented above, it was decided that a system to extract drug-bacterium relations from biomedical text would be a useful and novel con-

tribution to the field. A period of time was spent developing such a system as part of the overall research into extracting small molecule bioactivity. The developed system has a core involving cascaded SVM classifiers, and a full description of its operation will be provided in the following section.

2.2 Classification using support vector machines

In machine learning, a support vector machine refers to a supervised learning model with associated learning algorithms [69]. The SVM model is commonly used for binary classification and regression analysis. In these systems a set of training data is analysed and the SVM algorithm builds a model that is then able to assign future data points into mutually exclusive categories. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New data points are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. Figure 2.1 illustrates the basic principle of how a linear SVM constructs a hyperplane in feature space and separates data into classes. In this simple example the hyperplane is a straight line which is separating points in a two-dimensional space, however SVM kernels also exist that derive linear and non-linear hyperplanes for classification of data of higher dimensionality.

The training of an SVM requires suitable training data; in text mining applications this is generally a pre-annotated corpus.

The bioactivity extraction task is formalised as a binary classification task in which only true and false classes are identified, i. e. we are using the SVM to classify text as positive (likely to contain bioactivity) or negative (unlikely to contain bioactivity).

2.3 System overview

It was decided that the system would comprise three distinct stages, as shown in Figure 2.2. Initially an SVM classifier is used to perform classification at a document level, that is, to extract the documents likely to contain a drug-bacterium mention. Secondly, an additional SVM classifier uses the likely-positive documents to perform classification at a sentence level; that is, to extract the sentences likely to contain a drug-bacterium mention. Finally a simple algorithm uses keyword patterns to extract the drug-bacterium relationship from the surrounding context.



Fig. 2.1 A hyperplane of a support vector machine

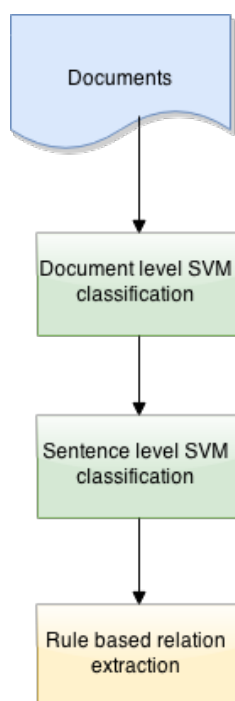


Fig. 2.2 System overview

This system design mirrors that used by the related works discussed in Section 2.1. Performing hierarchical classification (i.e. at a document, then at a sentence level) using cascaded classifiers significantly reduces the complexity of the task and is expected to lead to higher quality results. One might wonder why the classifiers are required at all, when the patterns alone could be applied to the whole data set. The reason is that the classifiers improve the quality of the final results: without the classifiers the patterns would be expected to produce a lot of output but with many false positives (low precision). The classifiers therefore perform intelligent pre-filtering as a way of removing noise. In order to achieve this, the SVMs must first be trained using suitable data. It was found that existing corpora (as detailed in Section 1.3.2) were unsuitable for this task, lacking the cross-domain coverage required for drug-bacterium applications. The decision was taken to develop a new corpus.

2.4 The DrugPro corpus

Before any system based on machine learning can be developed it is necessary to identify a suitable training corpus. In this system the training corpus is used to train both the document-level and the sentence-level classifier. Since a suitable corpus for small molecule

bioactivity did not previously exist, a new one was created for this work. The procedure followed to generate the corpus is summarised in the following section. The process includes selection of abstracts and abstract annotation steps. The corpus has been named DrugPro, because of its focus on drugs and prokaryotes. The development of the corpus is described in more detail in [70]. This is thought to be the first corpus in the drug-prokaryote domain and may find use in text mining applications relating to the field of antibiotic development, amongst others.

2.4.1 Selection of abstracts

We selected 400 abstracts published after 1999 as a basis for the DrugPro corpus. The corpus consists of two main sets of data, a likely-positive set and a random set. As a pre-processing step for selecting the likely-positive set as well as helping in the curation process, named entity recognition was applied to the entire MEDLINE corpus to annotate drug and bacteria names. These were taken from two dictionaries: DrugBank [71] and the DSMZ Bacteria Catalogue [72]. The DrugBank database contains 6,711 drug entries including 1,447 FDA-approved small molecule drugs, 131 FDA-approved biotech (protein/peptide) drugs, 85 nutraceuticals and 5,080 experimental drugs. DSMZ is one of the largest biological resource centres worldwide which provides a total number of 11,367 validly published bacterial species names, of which about 1,723 are synonyms belonging to 2,135 different named genera [72]. The tool used to apply NER was the Whatizit back-end, as described in Section 1.3.1.

Abstracts were initially selected using a co-occurrence approach and dictionary-based identification. If an abstract contains one or more sentences annotated with drug name(s), bacterium name(s) and one of the keywords indicating a type of bioactivity relationship described in Section 2.7, then this abstract is considered as a likely-positive abstract. A random selection of abstracts was carried out after the selection of the likely-positive set to form the random set.

2.4.2 Abstract annotation

The next step in preparation of the corpus was manual annotation of the abstracts. The format proposed for the BioInfer corpus [32] was followed to deliver a standard format for the corpus. A domain expert carried out the annotation. The domain expert corrected the errors of named entity recognition produced during the dictionary-based identification. The relation between drug and bacterium was also identified by the curator, as was the trigger

word (e. g. “inhibits”) which syntactically triggers the relation type.

Annotation followed some agreed principles. The annotator would not annotate the sentence unless all three main components of bioactivity occur in the sentence, namely trigger word, small molecule and biological target. Therefore the annotation of small molecule bioactivity in the scientific literature requires three components, which we term a bioactivity triple:

1. Annotation of small molecules
2. Annotation of targets (of different kinds)
3. Annotation of the bioactivities as the relation between, or event involving, the small molecule and the target.

The following general rules apply to the annotation task:

1. A single annotation spans a continuous stretch of text, i. e. discontinued annotations are currently not considered.
2. Nestedness of annotations has to be avoided. The annotation should cover the longest span that describes the entities.
3. For the annotation of the event, only the event trigger is annotated. Complex event annotations, including the entities involved in the event, are covered separately from the event triggers. In principle, selected resources could be provided for the normalisation of the chemical entities and the targets, but the curators were not limited to the semantic resources for the annotation of the involved entities.

A fragment from the annotated corpus is depicted in Figure 2.3. Once the DrugPro corpus was complete, it could be used in machine learning applications, beginning with binary classification at the document level. It is thought the corpus can have wider use in developing NERs for drug and bacteria names as well as identification of relations between them.

2.5 Document-level classification

A crucial prior step for relationship extraction is detecting which documents contain a relationship between the entities. This is a coarse initial binary classification. A system

```

<document pmid="11822771">
<sentence id="s1" text="Clinical strains of Staphylococcus aureus UCN7 and UCN8 were inducibly
resistant to erythromycin, clindamycin, lincomycin, and quinupristin.">
<entity id="1" charOff="21-41" type="bacteria">
<entity id="2" charOff="85-96" type="drug">
<entity id="3" charOff="99-109" type="drug">
<entity id="4" charOff="112-121" type="drug">
<entity id="5" charOff="128-139" type="drug">
<relation id="r1" e1="1" e2="2" type="resistant" keyWord="resistant">
<relation id="r2" e1="1" e2="3" type="resistant" keyWord="resistant">
<relation id="r3" e1="1" e2="4" type="resistant" keyWord="resistant">
<relation id="r4" e1="1" e2="5" type="resistant" keyWord="resistant">
<docType class="P" resistance="yes ">
</document>

```

Fig. 2.3 A fragment from the annotated DrugPro corpus

was required to classify documents into those likely to contain a drug-bacterium relation (a “positive document”) and those unlikely to contain one (a “negative document”). The main components of the document-level classification system are illustrated in Figure 2.4. The system consists of three stages: a pre-processing unit, a feature extractor, and a trained SVM classifier. The pre-processing unit identifies tokens in a document and rejects noisy tokens. The feature extractor extracts sensible features and combines them. The set of combined features is passed to the SVM classifier to output classified document sets. These three stages will now be described in more detail.

2.5.1 Document level pre-processing

The pre-processing step is essential to provide a structured representation for the subsequent processing stages. In reality, most human readable text is delivered and stored in rather complicated structure, so that in the most common cases of text mining a pre-processing step is usually required to clean the document and add in structure. Pre-processing of documents before classification involves tokenisation and sentence splitting, followed by normalisation. In this particular case, the work-flow is shown in Figure 2.5.

Firstly, every document in the corpus is scanned and annotated for NER purposes. The dictionary was the same dictionary used for creation of the DrugPro corpus, a collection of drug and bacterium names taken from the DrugBank database and the DSMZ Bacteria Catalogue respectively. Once located in the text, drug names and bacterium names were replaced with the tags “DRUG” and “BACTERIUM” respectively. Using generic terms like this avoids the huge amount of variation that would otherwise exist within noun phrases and

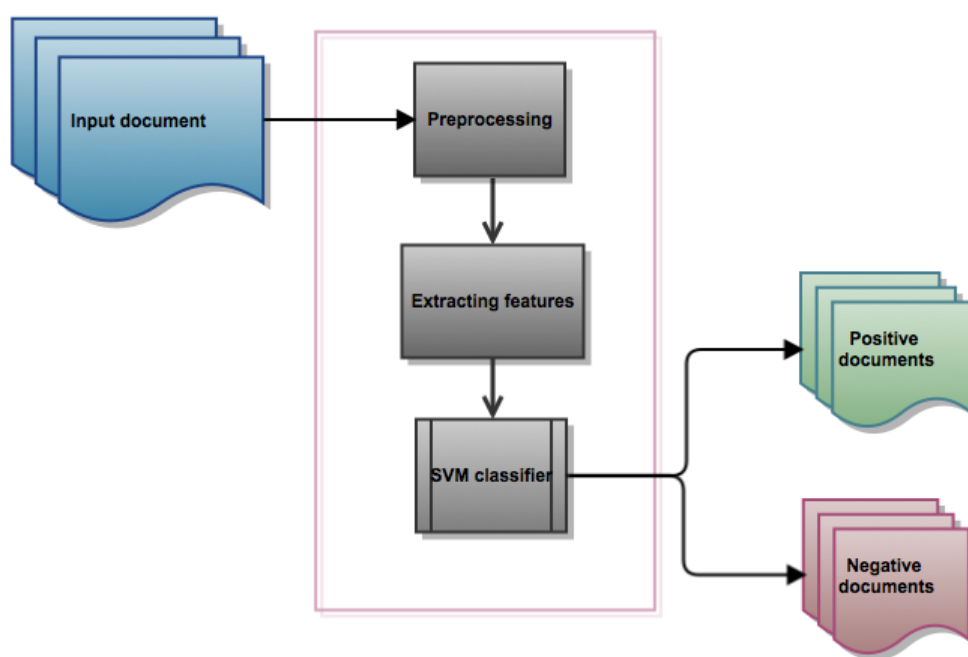


Fig. 2.4 The main components of the document-level SVM classifier

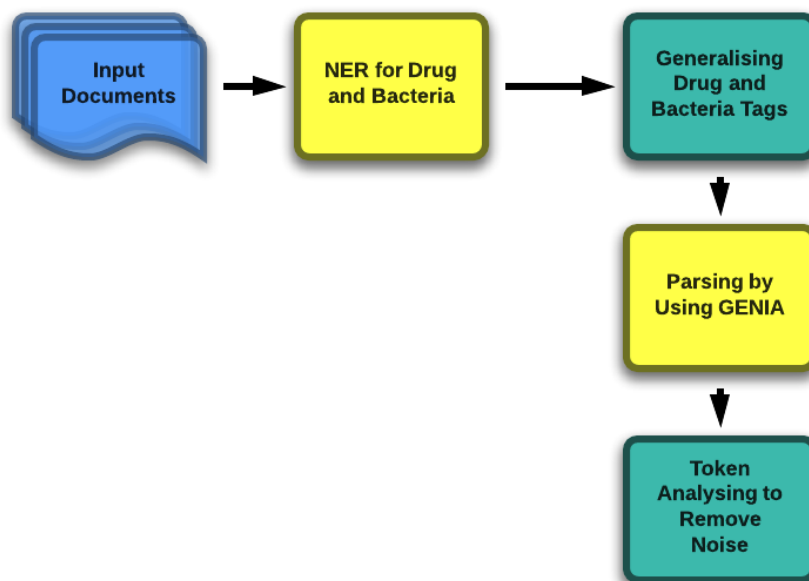


Fig. 2.5 Pre-processing for document classification

therefore greatly reduces the complexity of the classification task. Secondly, each document was split into sentences and then parsed using the GDep parser [20]. As described in Section 1.3.3, dependency parsing assigns part-of-speech tags to each token, a process which assists the subsequent normalisation task.

Normalisation uses a newly developed token analyser to scan through all the tokens. The token analyser software was written in Java. It was designed to perform a number of tasks. Capitalisation is removed, and all tokens comprising fewer than three letters were also eliminated. Digits and measurement units were eliminated since they do not positively contribute to the results and would introduce noise into the subsequent analysis stage: digits are commonly used as modifiers of a noun phrase and do not play useful roles in the relation extraction task which is only interested in the nouns themselves. Finally, the token analyser removed a list of stop words (frequently used English words considered not useful for analysis purposes) from the document collection. Removal of stop words effectively increases the signal-to-noise ratio of the input data, and for this reason is a common task in natural language processing, e. g. see [73]. The input list of stop words was obtained from online sources.

Table 2.1 Contingency table for frequency of term t_i in different document classes

Document class	t_i	\bar{t}_i
Positive	a	b
Negative	c	d

2.5.2 Extracting features

From a machine learning point of view, a feature is an individual measurable heuristic property of a phenomenon being observed [74]. Document classification is a feature driven matter [75]. Therefore, choosing discriminating and independent features is key to the success of document classification. We are interested in features that are indicative of a document containing drug-bacterium bioactivity. The required input to the SVM is a set of weighted features. The weights can be thought of as a measure of how relevant each feature is to the classification task.

In the study of feature selection, experimental work was carried out to test different feature sets and discover which methods were most successful. In the following subsections, each feature will be introduced and its relationship to the classification task will be explained.

Term weighting schemes

A set of features for a document can be the set of unique terms it contains (after pre-processing), also known as the vocabulary of the document. This concept can be extended to a whole corpus. The terms must be weighted to be used as SVM input data. There are various approaches to generating weights, but the simplest is the term frequency (TF), calculated for term t_i as follows:

$$TF = \frac{\text{occurrences of term } t_i \text{ in document}}{\text{total terms in document}} \quad (2.1)$$

In order to make the term frequency more useful, it can be combined with the inverse document frequency (IDF):

$$TF * IDF = TF \times \log_2 \frac{N}{a + c} \quad (2.2)$$

where N represents the total number of documents in the corpus, and a and c are the contingency values defined in Table 2.1. The TF*IDF score is a statistical weighting score reflecting the importance of a word to a document within a text collection [76, 77]. The

function of the scaling factor IDF is to reveal the information content of a particular term. It enables discrimination of the useful terms that may be highly prevalent in a few documents from common words that appear in all documents. TF*IDF has been used in information retrieval and text mining [78], in particular for the task of document classification [79, 80]. The TF*IDF value is the most popular term weighting scheme among various schemes used with BOW methods [81].

The third component to the term weighting scheme is the statistical χ^2 measure, which uses the full set of contingency values from Table 2.1.

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \quad (2.3)$$

χ^2 is a statistical independence test. We are using it to determine whether there is a correlation between positive documents and a particular term.

The set of features of a given data instance is grouped into a feature vector. As previously described, the reason for doing this is that the vector can be treated mathematically. In classification applications, many algorithms compute a score to classify an instance into a particular category by linearly combining a feature vector with a vector of weights, using a linear predictor function [82].

In this experiment the classifier was trained using sets of TF*IDF weighted BOW features, with χ^2 used to rank the features by usefulness [83] (i. e. by how correlated each term is with a document being positive). Additionally, experiments were performed where the BOW features were combined with domain-specific features, as follows:

Number of drug and bacterium occurrences

As noted in Section 2.4.1, the abstract selection of the DrugPro corpus was performed by manually checking the occurrences of drug and bacterium names in the text. During the annotation period, both annotators commented that the prevalence of drug and bacterium names made a marked difference in whether the document would be classified as a positive or negative. Therefore, it is believed that the probability of a randomly selected document from the training set being a positive drug-bacterium interaction document is strongly correlated with the frequency of drug and bacterium mentions in the text.

Table 2.2 Keywords found in the DrugPro corpus

Keywords	Frequency
resistant	111
resistance	63
susceptible	22
against	14
susceptibility	9
tolerant	8
sensitive	4
inhibit	3
sensitivity	1
inhibitory	1
inhibition	1
clear	1
activate	1

Co-occurrence of drug-bacterium pairs

Another item of feedback returned by the annotators was that the co-occurrence of drug-bacterium pairs is strongly associated with a relationship between drugs and bacteria. When the whole corpus was screened, it was found that one hundred of the positive abstracts had at least one occurrence of a drug and bacterium pair.

The co-occurrence level was used in two different ways as an SVM feature, one using co-occurrence at the level of the document abstract, the other at a sentence level. The result of using the two features will be compared in the results chapter. The co-occurrence provides a simple binary weight which is used in the training vector.

Keywords

This feature was proposed because the presence of certain words can indicate interaction between entities. Analysis showed that when the sentence contains a keyword such as *resistant* or *inhibit*, there is a higher probability that there is a relationship present than those cases when such a keyword is absent from sentence. When annotating the text, the annotators highlighted keywords which indicated and provided evidence of a relationship. The full list of keywords identified by the annotators, together with their frequency in the corpus, is shown in Table 2.2. TF*IDF scores were found for each of these terms and used as a feature with the SVM.

A simplified representation of training vectors comprising BOW features and domain-

TF*IDF scores for Terms $T_1..T_n$							Domain specific features $D_1..D_n$			
0.802	0.916	0.185	0.638	0.292	0.079	0.353	1	1	22	0.878
0.819	0.582	0.155	0.815	0.215	0.745	0.612	0	1	10	0.244
0.641	0.898	0.529	0.646	0.084	0.437	0.568	1	0	5	0.945
0.012	0.935	0.946	0.433	0.067	0.347	0.805	0	1	49	0.905

Fig. 2.6 Simplified representation of a bag-of-words vector with concatenated domain specific features. Each row is input to the SVM for document level classification.

specific features is shown in Figure 2.6. For clarity, this example only shows seven BOW features, but in reality there are several thousand. These are concatenated with the handful of domain specific features described above.

2.5.3 Document SVM classification

Classification was performed using the selected features as a basis for the SVM method. The method described is a machine learning approach that has been used for many classification issues including text and document classification, such as classifying protein interaction articles during the BioCreAtIvE-II task [84].

The software used for the training procedure was an implementation of SVM called SVM^{light} [85]. SVM^{light} has been successfully used for tackling the problem of pattern recognition, regression and learning ranking function. In this experiment, the three typical kernel functions are used with an SVM learner to integrate two sources of information from syntactic parse trees. The three kernels are: linear, polynomial and radial basis [85, 86].

The output of the document classification is a set of likely-positive documents, i.e. those documents that are likely to contain bioactivity descriptions. These are then used as input to the sentence level classifier, as will now be described.

2.6 Sentence-level classification

The design of the sentence-level classification system is shown in Figure 2.7. It comprises three main parts: a pre-processing unit, an SVM classifier and a keyword-based relation extraction unit. In this particular task, the SVM is used to classify whether a sentence contains a small molecule and one or more target entities in an instance of a bioactivity relationship. The following example sentence illustrates such a relationship and a pair of interacting partners:

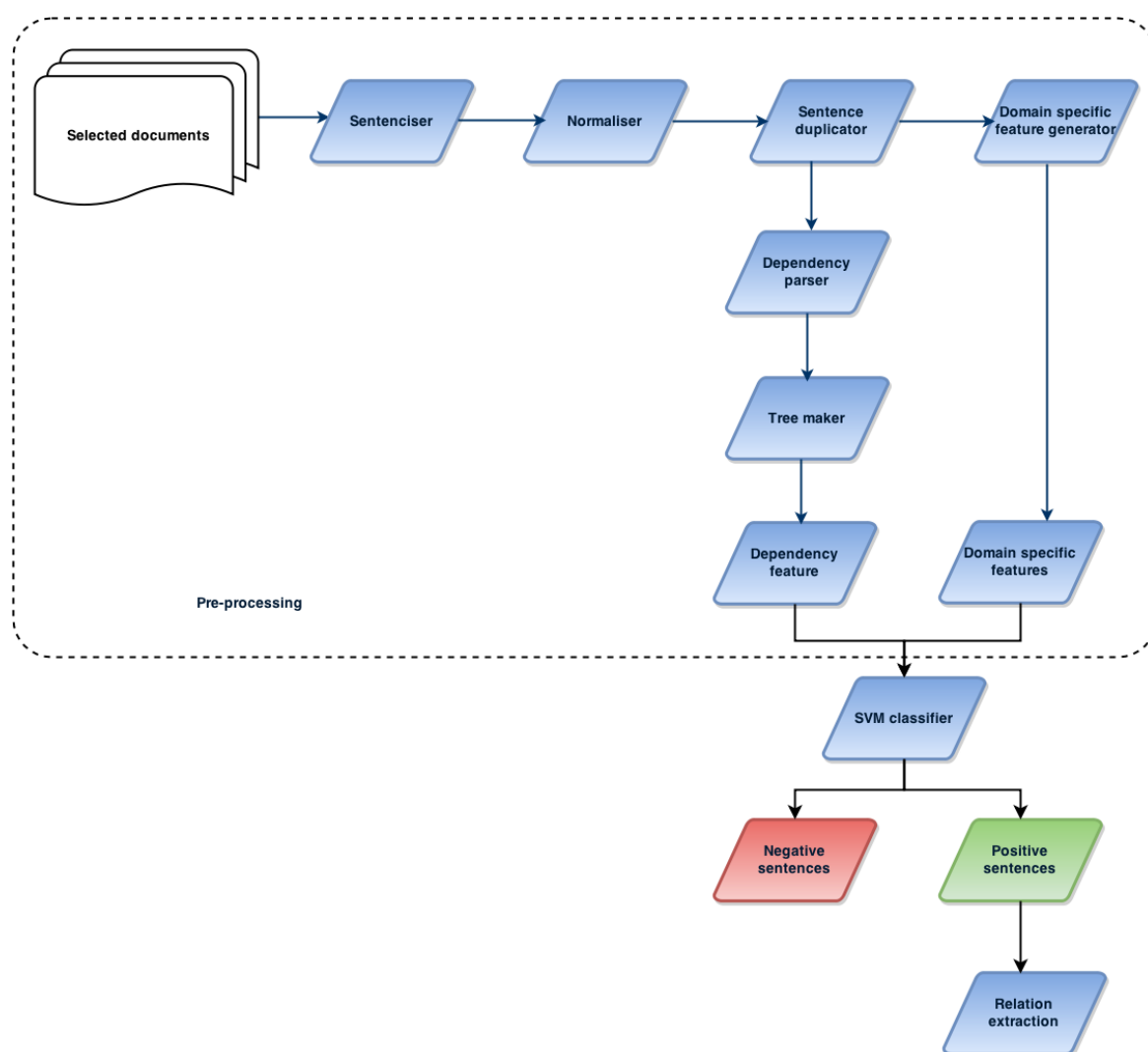


Fig. 2.7 Overview of sentence-level classification

*“In conclusion, our findings support the in vitro efficacy of **imipenem**, **meropenem** and **trimethoprim-sulfamethoxazole** against **B. pseudomallei**.”*

In this sentence the four targets of the extraction system are *imipenem*, *meropenem*, *trimethoprim-sulfamethoxazole* and *B. pseudomallei*. The terms *imipenem*, *meropenem* and *trimethoprim-sulfamethoxazole* are drugs, and *B. pseudomallei* is a bacterium. The preposition “against” connects the two types of entity and illustrates the relationship between them. These are the prospective outputs of the system.

The three processing stages involved in this task will now be described.

2.6.1 Sentence-level pre-processing

The pre-processing unit is constructed to eliminate noise in natural language text and construct a data input format for the sentence-level SVM. The pre-processing unit has multiple components including a sentenciser, a normaliser, a dependency parser, a tree maker and a domain-specific features generator. Some of these components are as described for the document level pre-processing task in Section 2.5.1. The other components are presented below.

Normaliser

After sentencising the likely-positive set of documents (separating each document into collections of individual sentences) a normaliser is used. The normaliser used for sentence-level classification plays the role of collapsing small molecule names and bioactivity target names into generic representations. For instance, in the sentence

“Corynebacterium striatum is resistant to the antibiotics chloramphenicol, erythromycin, kanamycin and tetracycline”

the drug names “chloramphenicol”, “erythromycin”, “kanamycin” and “tetracycline” are replaced by a generic tag “DrugTag”. The bacterium name “Corynebacterium striatum” is replaced by “BacteriaTag”. The sentence is therefore normalised into

BacteriaTag1 is resistant to the antibiotics DrugTag1, DrugTag2, DrugTag3 and DrugTag4

Sentence duplicator

The aim of the study is to identify relationships between entity pairs (of different semantic types), however some natural language sentences contain multiple possible combinations of the two entity types in a coordinate structure. For example, the above sentence contains four small molecule names as conjuncts in a relationship with a single target bacterium name. In this case, four different interacting pairs can be observed in one sentence.

The solution adopted to handle coordination structures like this is to duplicate the sentence for each conjunct. An algorithm detects the presence of multiple entities of the same semantic type and causes the duplication to take place.

Again referring to the above example, four copies of the sentence with different interacting pairs would be produced. In each case, a different conjunct is exposed for subsequent processing.

BacteriaTag1 is resistant to the antibiotics DrugTag1, DrugTag, DrugTag and DrugTag
BacteriaTag1 is resistant to the antibiotics DrugTag, DrugTag2, DrugTag and DrugTag
BacteriaTag1 is resistant to the antibiotics DrugTag, DrugTag, DrugTag3 and DrugTag
BacteriaTag1 is resistant to the antibiotics DrugTag, DrugTag, DrugTag and DrugTag4

Dependency parser

As discussed in Section 1.3.3, dependency parsing is a technique for analysing the grammatical structure of a natural language sentence, to attempt to understand its meaning. The benefit of dependency parsing is that it reveals non-local relationships between terms in the sentence [35].

A dependency parser takes sentences as input and produces a graph for each sentence where the nodes are the words and the arcs are dependency links between words. Dependency links are most commonly syntactic or semantic, though other dependencies can also be used. As before, the dependency parser used is the Gdep parser [31], which is a version of the KSDep dependency parser trained on the GENIA Treebank [87] for parsing biomedical text.

Gdep uses the LR algorithm to generate a parse tree which presents the relations within a sentence. The LR algorithm refers to a type of method that efficiently parses deterministic context-free languages in linear time [88]. The output presents binary relations between the head and dependent nodes. For example, Figure 2.8 shows a relational representation of the normalised sentence

“BacteriaTag1 is resistant to the antibiotics DrugTag1, DrugTag, DrugTag and DrugTag”

The labels shown in the figure indicate the type of dependency. In this particular example, the parser has identified that the verb “is” is the root of the sentence. The parser has further identified that “BacteriaTag1” is the subject of the sentence and “resistant” is the root node of the predicate; all other nodes in the predicate depend directly or indirectly on “resistant”. The tags *AMOD* and *NMOD* denote adjective modifier and noun modifier respectively. The *P* tag denotes punctuation.

It is believed that the information required to assert a relationship between two named entities in the same sentence is typically captured by the shortest path between the two entities in the dependency graph [89, 90], because it provides the essential components of the sentence without the surrounding context. A new software routine was developed to

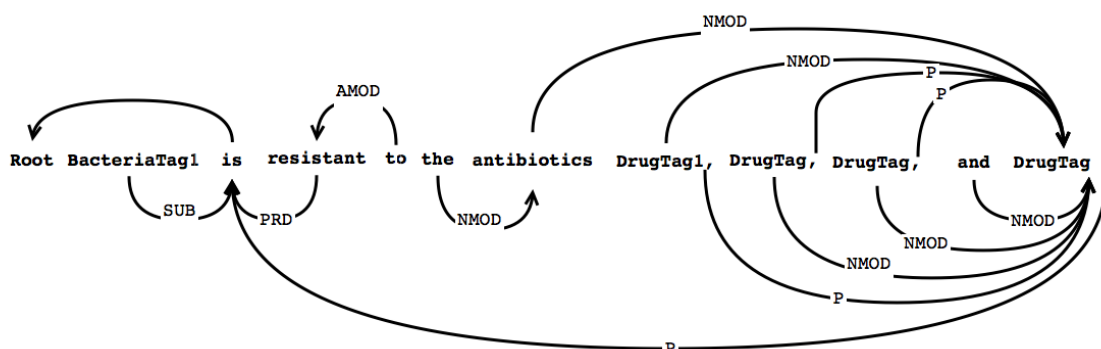


Fig. 2.8 Dependency relations generated by Gdep

```
(GDEP(r (DrugTag1 DrugTag ))(r (DrugTag to ))(r (to resistant ))(r (resistant be ))(r (be BacteriaTag1 )))
```

Fig. 2.9 Shortest path dependency feature

extract this shortest path from the GDep output. Figure 2.9 shows the resulting shortest dependency path from “DrugTag1” to “BacteriaTag1” for the example shown in Figure 2.8.

As a result the shortest path between a drug entity and bacterium entity can be used as a possible feature to input into the sentence-level SVM classifier. We call this a dependency feature.

Domain specific feature generator

Domain specific features are generated to be combined with the dependency features. In particular, three main types of domain-specific features are generated, namely, the binary presence of drug name, target name, or keyword; the number of drug names in the sentence; the number of target names in the sentence.

2.6.2 Sentence SVM classification

Once extracted, the above features are used as input into the sentence SVM classifier. The sentence classifier was trained on the DrugPro corpus. The 400 manually annotated abstracts comprise 3,800 sentences, of which only 760 are qualified as positive sentences in which

Table 2.3 Sentence classification feature set. The dependency feature used most often is shortest path between drug and bacterium name.

Feature set
Dependency feature
Dependency feature + number of drug names
Dependency feature + number of bacterium names
Dependency feature + binary occurrence of drugs, bacteria and keywords

one or more relations exist. In order to balance the training data set, 800 negative sentences were randomly selected and added to the training set. An experiment to compare the effects of balancing the training set found that it is crucial and majorly affects the performance of the classifier.

After training, the model can be used to classify unseen sentences. Four experiments were carried out and 5-fold cross-validation was applied to each experiment to assess the performance. Each experiment was designed by combining different features. The SVM classifier takes the syntactic features identified by the dependency parser (as described in Section 2.6.1), optionally concatenated with a domain-specific feature. This is used to train the classifier and ultimately find the feature set that produces the best results. Table 2.3 shows the combination of feature sets.

The SVM model separates the collections of sentences harvested from the DrugPro corpus into positive and negative classes. The positive class of sentences contain a pair of entities with one kind of relation and therefore presents a potential bioactivity scenario. The negative class of collection is a collection with no recognised relations.

2.7 Relationship extraction

We then analysed the entire set of correctly classified relationship-bearing sentences in the corpus. This analysis showed that most of the relationships are signalled by verb-based forms. This led to development of patterns to perform the event extraction. Once candidate sentences had been extracted, a simple rule-based system was used to identify the relationship connecting the drug entity and the bacterium entity. An extraction algorithm was implemented to identify and extract relationships at the sentence level using verb-based keywords.

Table 2.4 Keywords found in the DrugPro corpus

Keywords	Frequency
resistant	111
resistance	63
susceptible	22
against	14
susceptibility	9
tolerant	8
sensitive	4
inhibit	3
sensitivity	1
inhibitory	1
inhibition	1
clear	1
activate	1

2.7.1 Keyword-based relationship identification

A list of keywords was provided by domain experts when annotating the DrugPro corpus. In total, they examined 249 relationship-bearing sentences in 400 annotated abstracts. The list of keywords and their frequency in the corpus is shown in Table 2.4. The presence of one of these keywords in a sentence with a drug and bacterium is taken to signal the presence of a relevant bioactivity.

Some examples of true positive cases were studied and it was possible to identify several key criteria for a relationship-bearing sentence:

- Adjective keywords: Several of the keywords are adjectives, for example the high frequency keyword “resistant” reflects the major relationship type between drug and bacterium. The following example shows when a true relation is detected in a relationship bearing sentence:

“The organisms tested included four erythromycin-resistant isolates of B. pertussis from a single patient.”

In this sentence, “resistant” was identified as the relationship. It is directly adjacent to the drug *erythromycin*.

Alternatively, a common pattern of the relationship signalled by “resistant” includes the preposition “to”, as in:

*“106 strains of **Escherichia coli** were examined, 68% of these were **resistant** to **tetracycline**, and 57% were **resistant** to **ampicillin** and **cotrimoxazole** respectively.”*

So two language patterns of relations involving the word “resistant” were defined, which are <DRUG resistant BACTERIA> and <BACTERIA_NP <be> resistant to DRUG>. In this case the token <be> represents the verb “be” and its inflectional morphological variants (is, am, were etc). Similar patterns were created for other adjective keywords, including the antonym “susceptible”, and the pair “sensitive” and “tolerant”. For example in the following sentences:

*“Whatever the serotype, 76% of **P. aeruginosa** strains with bla PSE-1 are **susceptible** to **fosfomycin**”*

*“**Cefepime**, **cefuroxime**, **ceftazidime**, and **erythromycin** all demonstrated excellent efficacy **against** fully **penicillin**-susceptible isolates of **S. pneumoniae**”*

- Noun keywords: The nouns ‘resistance’ and ‘susceptibility’ often appear in the drug-bacteria relationship-bearing sentence when bacteria sensitivity to certain antibiotic drugs is tested. The following examples show possible forms of linguistic realisation in this category. The general pattern of such relational presentation can be seen from the sentences, where mentions of the bacteria are often connected with a preposition word such as “in”, or “of”. In addition, the word “resistance/susceptibility” is often the head of a noun phrase, and a bacterium name is usually in the modifier position. For example:

*“Previous antibiotic prescriptions were strongly associated with **resistance** to **ampicillin** in **E. coli**”*

*“This study aimed to test the **susceptibility** of Malaysian isolates of **B. pseudomallei** against **imipenem**.”*

- Preposition keyword: When the keyword “against” is used to present a type of relationship between drug and bacterium, it forms a prepositional phrase with a noun phrase representing the bacterium. For example:

*“**Telithromycin** showed good activity **against** **S. pyogenes** isolates.”*

A high frequency of **S. aureus** (30%) was **resistant** to **ciprofloxacin**.

Fig. 2.10 Identification of relation type

Table 2.5 Language patterns of rules used to extract drug-bacterium relationships

Keyword	Example patterns
Resistant/susceptible/ sensitive/tolerant	<NP1><KEYWORD><NP2> <NP1><be><KEYWORD><preposition><NP2>
Resistance/susceptibility/ sensitivity/tolerance	<KEYWORD><preposition><NP1><preposition><NP2>
Against	<NP1><KEYWORD><NP2>
Other	<NP1><verb><NP2>

- Other cases: A NP1-Verb-NP2 relationship construction rule [91] is applied on other types of relationship identification when the keywords presented in the above cases were not found. The “NP1-Verb-NP2” rule was proposed to identify the relation between two noun phrases, NP1 and NP2 respectively [92]. A generic pattern of this form is able to catch many cases that have been missed by the preceding rules. The pattern is used to find the occurrence of two noun phrases separated by a verb (and possibly a preposition). The phrase is extracted as a relationship if the verb matches one of the keywords.

These four categories explain how to approach the matter of relationship extraction in a relationship-bearing sentence. The rules are generally pattern-based which also employ natural language processing (NLP). The language patterns were collected as in Table 2.5, and used to extract the relationship. An example of a sentence after relationship identification is shown in Figure 2.10.

When the patterns are applied the system also searches for possible negation cues in the immediate context. It was found simple lexical cues can be used as a relatively crude way of detecting negation. The presence of one of the negation words of Table 2.6 indicated that the relationship was a negative relation. These are still extracted, but the negative nature of the relationship is noted as part of the bioactivity triple (e. g. the bioactivity type is recorded as “not inhibit”).

Results and evaluation of the techniques presented in this chapter will be provided in

Table 2.6 Negation cues

Negation cues
not
cannot unable, not able
fail

Chapter 4.

2.8 Chapter summary

In this chapter, a hybrid method of extracting drug-bacterium relations has been presented. The system comprises a core of two cascaded support vector machine classifiers. One is for document classification in order to detect the abstracts which include one or more drug-prokaryote relations. The system is carefully designed to select features by combining different domain-specific features with bag-of-word features. The second classifier performs sentence classification, where the system intends to select the drug-bacterium relationship at the sentence level. For the second step, an approach similar to document classification was applied, where sentences are classified combination of semantic and syntactic features.

Chapter 3

General small-molecule bioactivity events

In this chapter, a rule-based extraction system able to extract small molecule bioactivity events will be presented. The system identifies the event participants and their dependencies by using a pattern matching approach. The system consists of a core that contains two different sets of rules, selected according to the syntactic type of trigger words associated with the bioactivity event.

In contrast to the work presented in Chapter 2, the work presented in this chapter is concerned with all types of small molecule bioactivity, not only drug-bacterium relationships. This gives rise to a high degree of variation in the targets of bioactivity encountered in the text. For example, the following sentence is a valid example of small molecule bioactivity:

*“**Penicillin** did not **inhibit** the **recovery process**”*

In this example the target of the bioactivity is *recovery process* which is clearly a general, high-level concept. On the other hand, the effects of small molecule bioactivity can be very precise, for instance

*“**Ceramide** did not directly **inhibit** **CDK2** in vitro but caused **activation** of **p21**”*

In order to handle this extremely broad scope, it is necessary to be able to identify a wide range of entity types. This places very demanding requirements on named entity recognition. It also means that creation of a fully annotated corpus for supervised machine learning is very challenging and time consuming, and ultimately beyond the scope of this project.

In the absence of a corpus, a purely rule-based approach was used to identify small molecule bioactivity. This chapter will describe its operation.

3.1 Technical background

A common theme in applications of rule-based systems for natural language processing is the choice between two approaches: the syntactic approach and the lexical approach.

In general terms, a syntactic approach refers to a formal, hierarchical definition of text structure, invariably some form of grammar [93]. In a typical case, syntactic systems parse the text into a tree of elements, and then search the tree for an input text sequence.

The lexical system is less formal and rarely hierarchical. Lexical systems simply use regular expressions or a similar pattern language to describe the semantic structure of a certain relation or event. Omitting the step of syntactic parsing, a lexical system considers the input text as a sequence of flat segments, for instance, characters, tokens, or lines.

In the biomedical text mining realm, rule-based extraction approaches have been applied to relation extraction and biological event extraction. Several rule-based extraction systems have been developed for biomedical purposes, especially when domain-specific event extraction competitions were held. The systems developed have overcome some of the difficulties of complex sentence structure, notably in [94].

Rule-based approaches typically work by detecting predefined event triggers in the input text, applying syntactic tools such as dependency parsing, and then applying lexical patterns on the results to find the key items of information. There are several published works in the field that are rule-based and utilise a combination of both syntactic approach and a lexical approach, such as RelEx [35], and MeTAE [95].

RelEx was initially described in Chapter 1. It is a widely cited PPI text mining project. It is a rule-based approach which uses dependency parse trees to identify semantic pathways between relevant entities in a sentence. On a particular data set, RelEx achieves 85% precision and 79% precision. This is stated as being significantly higher precision than a co-occurrence search. Although the results are promising, this is on the specialised task of protein-protein interaction rather than the very broad scope that is of interest to us here. It also ignores negation, whereas we are interested in extracting negative as well as positive instances of bioactivity.

MeTAE [95] is used for automatic extraction of semantic relations between medical entities. The work provided a platform to extract and annotate medical entities and relationships from medical texts and to explore the produced annotations. The authors utilised linguistic patterns and domain knowledge to recognise medical entities and identify the correct semantic relation between each pair of entities. The result of extracting relationships between a treatment type and a problem was evaluated and 75.72% precision and 60.46% recall was

obtained.

The highlight of the MeTAE system is yielding linguistic patterns semi-automatically from a corpus selected according to semantic criteria. However a limitation with the approach is the fact that the knowledge-based starting point would not always be transferable when applying the system to different types of relations.

In addition, the corpus for generating linguistic patterns requires a specific qualifier for the target relation to obtain a more focused corpus for pattern construction. If this is not the case, a decrease in the relevance of the obtained abstracts/texts may be expected. Despite this, good precision is usually obtained by a ruled-based approach. The author also anticipated the disadvantage of pattern-based methods which is the expensive cost needed to obtain good recall.

A rule-based methodology was developed by Kilicoglu & Bergler [96] addressing the BioNLP'09 shared tasks. The approach uses dependency parsing as the underlying principle for extracting and characterising events. The approach was formed of three steps. They are:

1. Determining whether a word is an event trigger.
2. Identifying potential participant(s) of an event trigger.
3. Validating the event and sub-events.

Dependency-based rules were used to determine whether any entity or trigger in the sentence qualifies as an argument of the event. Additionally, grammar rules are applied in order of simplicity; rules that involve a direct dependency between the clue and any word of the entity are considered first. The performance of the system was evaluated at an overall recall of 32.68% and precision of 60.83% (the F_1 -score is 42.52%). The favouring of precision over recall is typical of rule-based approaches. The system includes speculation recognition, i. e. detecting whether a bioactivity is confirmed to have taken place or whether the text being processed is simply discussing the possibility of it (often indicated with verbs such as evaluate, investigate, etc.).

A rule-based system noted for its speed was presented in [97]. The system incorporates a learning phase and an extracting phase. Firstly, in the learning phase, a dictionary and patterns are generated automatically from annotated events. In the second phase the patterns are applied to extract events from input text. The system is particularly fast, and it produces reasonable results when evaluated against the GENIA event extraction task of the BioNLP 2013 shared task, with F_1 -score of 50.68%. Once again, the requirement for an annotated training corpus limits its scope and applicability.

During event extraction, patterns obtained from annotated events in the input text were used. The input sentences were converted into a structured representation by hand before NLP processing. Then the tokens of each sentence were matched against a dictionary generated from annotated events to detect the event trigger. Once an event trigger is detected, the retrieved patterns are applied to the input text.

Bui & Sloot [98] extracted biomedical events from the GENIA event extraction task in BioNLP'09. The approach uses manually created syntactic patterns derived from a parse tree and consists of three main components: a dictionary to detect triggers, text pre-processing, and event extraction. There are two rules which are applied based on the POS tag of the trigger word. Rule 1 is for extracting events from a noun phrase, and Rule 2 is for extracting events from a verb phrase. Recall and precision are relatively low, at 38% and 52% respectively.

The simplicity of this last approach is appealing, because it avoids the need for a training corpus. Splitting the rules depending on the POS tags is also thought to be a sensible approach. Although it cannot be directly applied because we have broader scope in terms of target types, it was taken as inspiration for much of the work to be presented in this chapter.

3.1.1 Rule-based event extraction process

Hobbs [99] recommends that event extraction should be divided into five sequential processes:

1. *Complex words*: Refers to recognition of multi-word terms and proper names. In the case of finding biological role terms, one needs to look for the trigger word and a biological mention in a particular place.
2. *Basic phrases*: The sentence is segmented into different phrasal groups, such as noun phrase, verb phrase, and so on.
3. *Complex phrases*: Complex noun groups and complex verb groups are identified.
4. *Domain patterns*: Phrases are scanned for patterns of interest.
5. *Merging structures*: Semantic structures from different parts of the text are merged if it is found that they provide information about the same entity or event.

Each successive stage operates on larger segments of text. The reasoning behind this approach is that the earlier stages recognise smaller linguistic objects and can operate in

a largely domain independent manner. The later stages take the linguistic objects as input and perform higher level, domain-specific processing. Hobbs calls this approach cascaded finite-state transducers, because each stage is implemented using finite state automata.

It is worth noting that this is a *bottom-up* process, beginning with individual words and then working upwards to phrases and sentences. This is in contrast to the *top-down* machine learning approach of Chapter 2, where a classifier was used to first find relevant documents and then secondly find relevant sentences, before using patterns to find the words of interest. We will return to this comparison in Chapter 4.

In summary, implementation of rule-based approaches requires extensive domain knowledge to find event trigger words and generic patterns. Rule-based extraction systems usually produce a high precision result on account of applying these highly restricted and specific rules, however this comes at the expense of recall.

3.2 Rule development for event detection and characterisation

Using the above background knowledge, a rule-based extraction system was developed in order to find general bioactivity events within the biomedical text. The ultimate aim is extraction of a bioactivity triple comprising three items of information:

- a small molecule term;
- a term for bioactivity type, possibly negated, which presents the relation type; and
- a term representing the target of the small molecule.

The five step process suggested by Hobbs [99] (described in Section 3.1.1) was followed and implemented, with the aim of scanning MEDLINE abstracts for bioactivity events.

The essential step is to determine good trigger words which can be used to select the documents potentially containing bioactivity event information. The input data (the MEDLINE abstract collection) were used to construct a dictionary of event triggers, drawing from MEDLINE annotations of triggers and making further refinements. The event triggers proposed are often found as the predicate in a sentence [100]. Trigger words are therefore restricted to verbs and nouns. In general other POS tags are tentatively annotated as event triggers and in fact require more context to fully qualify as triggers.

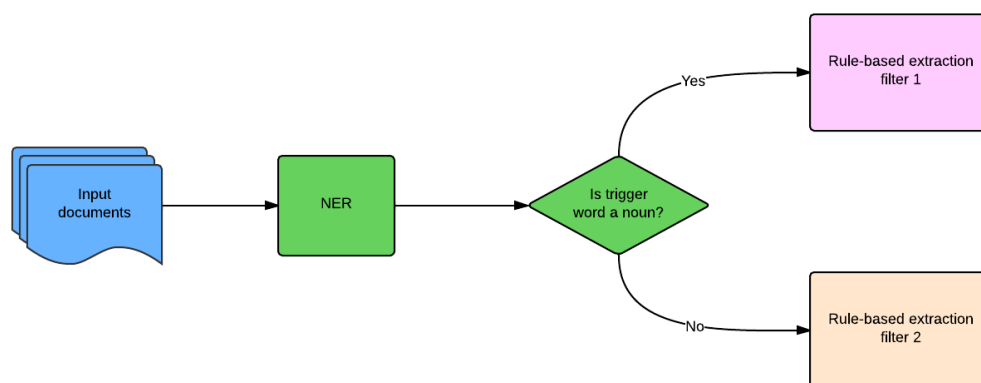


Fig. 3.1 The rule-based system of extracting bioactivity events used in this thesis

Before the extraction units are applied, the trigger words are firstly identified and marked up. The list of trigger words was proposed by domain experts, and they act as *roles* in bioactivity event indicators. These trigger words not only signal the presence of a bioactivity event but also characterise the event type.

3.2.1 Rule classification

The extraction rules are divided according to the part-of-speech (POS) of the event trigger words. The type of the trigger word usually infers its own characteristic of event presentation, which can be represented by using a particular language pattern. When developing the system, bioactivity formations were classified based on the POS of the trigger word, then language patterns for each were generated and tested.

Figure 3.1 shows an overview of the operation of the resulting rule-based extraction system. The classification of trigger words into verb and noun forms strongly influences and simplifies the subsequent processing. Noun and verb forms of trigger words can be handled separately leading to definition of two sets of rules for the different bioactivity presentation forms, leading to two different extraction units.

Trigger word in noun form

In these types of sentences the trigger word appears in the predicate noun phrase of a sentence. A predicate noun is a single noun or noun phrase that provides information about the subject of the sentence. The subject and the predicate are linked via a copula: usually a form of the verb “to be” or occasionally another verb (see below). Sentences of this form

Table 3.1 Two types of trigger word

Noun trigger words	Verb trigger words
stimulator	stimulate
adaptor	adapt
modulator	modulate
messenger	
agent	
blocker	block
toxin	toxicate
suppressor	suppress
regulator	regulate
factor	
agonist	agonise
activator	activate
antagonist	antagonise
inhibitor	inhibit

are often used to denote the presence (or absence) of equivalence between the subject and the predicate noun phrase. For example, *Aspirin is a COX-1 inhibitor*, where the subject is *Aspirin* and the predicate noun phrase is *COX-1 inhibitor*. In this case the trigger word is the head noun *inhibitor*.

Trigger word in verb form

In these sentences the trigger word appears as the main verb of a sentence, characterised (as all verbs are) by a mood, voice and tense. As with the noun forms described above, these sentences are also based around a predicate noun phrase but in this case the trigger word is the copula. In the simplest case a bioactivity event could be a complete declarative sentence in the form of <subject><predicate>, for instance *Aspirin inhibits COX-1*. This is a form often seen in biomedical text.

Table 3.1 shows the full list of trigger words after normalisation. Two set of rules to extract bioactivity were investigated accordingly.

3.3 Named entity recognition

As a general first stage of biological event extraction, pre-processing included named entity recognition, tokenisation and syntactic parsing. Keeping with the main principle of cascade finite-state transducers [99], the system first attempts to identify mentions of relevant entities

Table 3.2 NER dictionaries

Entity	Dictionary
Biological process	GO-BP [9]
Protein	UniProt [8]
Small molecule	ChEBI
Organ	Unified Medical Language System (UMLS) [102]
Organism	NCBI Taxonomy [10]

and their semantic types. The associated entity sets, small molecule names, trigger words and the target mentions of bioactivity are involved in this task. Therefore, in this particular case, the first step is to recognise the domain-relevant mentions and further to identify the semantically related phrases. In the following subsections, individual NER of different types of entity will be presented and recognition of two types of related phrases will be introduced. The method to extract them will also be discussed.

NER is a crucial step to provide high accuracy event extraction. As many types of entity must be identified in this particular case, the complexity of NER depends on the level of boundary overlap and how complex the relationship is between different entity types [101].

The identification work was mainly applied to lexicon resources and lexicon patterns. For the purpose of disambiguation, the NER system must first identify the entity that corresponds to the string with the largest extension.

Since bioactivity potential covers a wide range of entity types, the implementation of NER is not straightforward. There is no single dictionary resource available to perform NER across the relevant target entities of small molecules, proteins, biological processes, organs, and species.

3.3.1 Dictionary selection and analysis

A dictionary-based approach was used for the annotation of entities. Since we are searching for bioactivity against a range of targets, it is necessary to apply multiple NER passes. For the bioactivity extraction task, the following NER tagging priorities were identified: biological process, protein, small molecule, organ and organism. The selection of dictionary for each of these types is shown in Table 3.2.

Applying these NER passes in an uncontrolled way will lead to sub-optimal results. When tagging terms, we want to ensure that it is the head of a noun phrase that receives the tag. For example, in the fragment *human liver*, the term *human* is not the target entity, it is merely a modifier for the true target *liver*. This means that we should apply the Organ

Table 3.3 Analysis of cross-tagging between different resources

Dict Corp	ChEBI		UniProt		GO		Organ		Species	
ChEBI			13093	13.0%	15	0.01%	1412	1.4%	3684	3.6%
UniProt	27337	12.1%			366	0.2%	4157	2.0%	3858	1.7%
GO	4719	44.1%	1372	12.8%			645	6.0%	307	2.9%
Organ	4879	3.8%	5295	4.1%	48	0.4%			6901	5.4%
Species	20456	7.7%	13396	5.1%	82	0.03%	2915	1.1%		

dictionary before the Organism one.

In order to generalise this it was therefore necessary to cross-evaluate the different NER dictionaries to understand their overlap and complementarity. The approach taken was to use the dictionaries to tag the other dictionaries. The result of this cross-tagging is shown in Table 3.3. These results show, for example, that the dictionary derived from ChEBI was able to tag 12% of the terms in the UniProt dictionary, and 44% of the terms in the GO dictionary. Of course these results are not reciprocal. The GO dictionary was able to tag very few of the terms in the other dictionaries, which illustrates how specialised the GO dictionary is.

This experiment informed the application of dictionaries in the NER task. Using these data a suitable cascade could be designed which achieved the most effective tagging of all entity types. This is particularly important for this application because our target entities cover such a broad range of semantic types. To the author's knowledge this is a novel technique which has not been reported elsewhere.

3.3.2 NER implementation

Once again the Whatizit back-end was used for NER. This required some initial work to prepare the dictionaries in the appropriate file format.

After dictionary-based identification, a set of rule-based filters were applied in order to reduce noise. The filters were selected for each type of entity. Dictionary preparation and filtering for each entity type will now be described.

Biological process terms

Biological process (BP) terms have the largest boundary among the five types of entities and potentially extend across other types of mention, so the identification starts with applying a BP dictionary on the input text. First, using the BP subset of GO terms, biological pro-

cess names were retrieved from the GO online resource and converted into the appropriate format. Then the dictionary was applied to the input text.

Protein terms

A protein dictionary was next used to tag the MEDLINE abstracts after GO-BP terms had been identified. Protein name identification is based on the UniProt dictionary which is generated from the online UniProt database and converted into the dictionary format.

Two filters were applied to clean up the results of protein NER:

1. The first rule was designed to remove a potential conflict between protein names and chemical names. Many protein names include the name of a chemical. This rule removes chemical tags if one or more are seen within the boundary of a protein name string. The restriction was based on the aforementioned study (Table 3.3) revealing the morphological association between chemical names and protein names.
2. The second rule filters out false positives caused by tagging of common English terms. For example the English term *Be* could be confused with the chemical term Beryllium's short form. In order to produce a list of common English words the British National Corpus (BNC) was used to derive a filter. Words were selected based on their frequency in the BNC. Through experimentation it was found that a frequency of 250 was a good threshold which produced a useful list of common words.

Small molecule terms

The implementation of small molecule NER was a hybrid system that employs a dictionary-based and rule-based approach. To identify the small molecules, results from two different chemical taggers were compared, namely the (at the time) newest versions of OSCAR (OSCAR3 [30]) and Jochem [29].

Jochem, being purely dictionary-based, has the advantage that all chemical entities it recognises are known entities, whereas OSCAR has a machine learning element, enabling recognition of unknown strings that resemble syntactic structures denoting chemical entities (thus giving higher recall). A corpus-based assessment of the two approaches was carried out using the SCAI corpus [103], which was mainly developed for NER of chemical entities. The purpose of the study was to assess the performance of OSCAR and Jochem when identifying chemical names in literature, and to find the associations between different chemical terminological resources and ChEBI.

Table 3.4 NER performance against SCAI corpus

	ChEBI	OSCAR3	Jochem
Recall	29.8%	72.9%	47.8%
Precision	34.0%	39.5%	40.2%
F-Measure	31.8%	51.3%	43.7%

Table 3.4 compares the performance of the three NER methods: ChEBI, OSCAR3 and Jochem. OSCAR3 outperforms the other two, out of which ChEBI has the poorest performance. From these results it was found necessary to improve the performance using a small ChEBI dictionary. The ChEBI dictionary was built by using the primary data available from the online ChEBI database. A program was written to convert the ChEBI data, obtained in XML format, into our working dictionary format (as used by the Whatizit backend). After that, several major pre-processing steps were carried out including suppression of some terms to reduce noise, so that the performance of the dictionary can be improved. The strategy included manual removal of terms potentially confused with other semantic terminology types. This avoids ambiguity, for example that between proteins and genes.

There are four filters applied to the dictionary to improve its quality:

1. Short token filter rule: a term is removed if it is a single letter (Arabic or single character) or Roman numeral. Greek alphabet characters, usually used to discriminate protein sub-units, are also removed from the dictionary. Terms consisting of letters with a number as a suffix are also removed from the dictionary because they are often the short form of a protein complex, for example *p45*. This filter has the effect of removing ambiguous terms.
2. Role term filter rule: removes terms which are a role term according to the ChEBI ontology (e. g. *antibiotic*)
3. Suffix filter rule: filters out terms with the suffix “-ase”, since these terms are normally enzyme names.
4. BNC filter rule: the BNC filter used with proteins was also applied for small molecules

After applying these filters to the raw ChEBI data, a dictionary of small molecules was available for NER.

Organ terms

Identification of organ entities was based on the the Unified Medical Language System (UMLS) [102] terminological resource “organ and body part” subset. As before, the subset was extracted and converted into appropriate format, and applied to the input MEDLINE text.

Species terms

The method of performing NER of species entities is also a hybrid method combining dictionary-based and rule-based approaches. The list of species names was retrieved from the NCBI taxonomy [10] subset, then it was converted into the appropriate format and used to identify the species names from the input MEDLINE text. Once again the BNC filter was applied to reduce the number of false positives caused by common English terms.

3.4 Noun phrase as event trigger

Situations where the event trigger is a noun are classified as “biological role terms”, because the sentence is providing a description of the biological role being played by the small molecule. The biological role term acts as the predicate noun of the subject small molecule; that is, the biological role term is modifying the small molecule.

In order to produce patterns for this type of sentence, it is useful to undertake dependency analysis. Figure 3.2 shows a dependency parsed tree result returned from the dependency parser Gdep [104] when applied to the following text:

“Aspirin is a COX-1 inhibitor”

This sentence contains a subject ‘Aspirin’ which is described as a ‘COX-1 inhibitor’. Here *COX-1 inhibitor* is a predicate noun that assigns a role to the drug. This format is often seen in biomedical text to present the biological role term and its equivalent chemicals.

Similar language patterns were defined for bioactivity terminology based on the examination of relevant portions of the Metathesaurus of UMLS [102] and the ChEBI biological roles.

3.4.1 Patterns of biological role terms

Any of the trigger words can occur as a head noun in a phrase structure where they are pre-modified by the target, as follows:

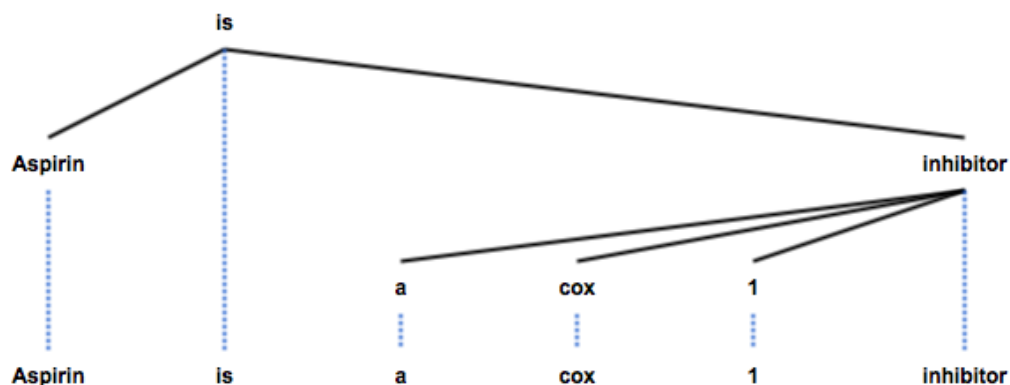


Fig. 3.2 The dependency parse tree of an example bioactivity description

<modifier><head>

Ideally, the phrase composing (*<modifier>*) is constituted by one or more tokens which denote the *target* of the bioactivity, whereas the head word specifies the *nature* of the interaction between the small molecule and the target. For example,

“beta-adrenergic receptor inhibitor”

has as modifier ‘beta-adrenergic receptor’ (the target) and as head word ‘inhibitor’ (the nature of the interaction).

The basic language pattern was further extended to include alternative, compatible language patterns such as ‘inhibitor of X’, where ‘X’ corresponds to the modifier and ‘inhibitor’ is the head word. Three variants for bioactivity role terms were identified, namely:

- Noun phrase or adjective/adverb compositions as modifier. This is the most commonly seen structure of the noun phrase (also called “basic noun phrase”), and a considerable number of bioactivity terms are presented in this way. For example:

“Kinase suppressor ”

“HIV transcriptase inhibitor”

- Prepositional phrase as modifier. Prepositional phrases are generally formed by the trigger word linked via a preposition to a subordinated noun phrase. It was found that

this structure often describes bioactivity events. The preposition will commonly be 'of', as in the following example extracted from MEDLINE:

“Suppressor of fused protein Oct-1 CoActivator”

- Relative clauses as modifier. Relative clauses are defined as subordinate clauses that begin with a relative pronoun. This type of modifier is also used in bioactivity events. For example:

“Factor that binds to inducer of short transcripts protein 1”

These variants were used to define two abstract patterns for extraction of bioactivity events from role term descriptions. They are presented in Table 3.5. The token <SM> refers to the small molecule. Also note that in rule NP1, the token <be> represents the verb *be* and its inflectional morphological variants (*is*, *am*, *were* etc).

Table 3.5 Patterns for bioactivity noun phrases

NP1	<SM><be><TARGET><TRIGGER>
NP2	<SM>,<TARGET><TRIGGER>

The implementation of the rule-based extraction system for biological role terms is shown in Figure 3.3. The extraction system consists of three main parts: a pre-processing stage, the pattern filters, and triple extraction (where triple refers to the three concepts of interest: the role term represented as one of the trigger words in noun form, the small molecule name, and the target). These processes will be explained in the following subsections.

3.4.2 Pre-processing

Pre-processing is required to remove noise introduced by natural language processing in biomedical text. The pre-processing solution developed for biological role term extraction is based on a POS tagger licensed from CIS [34].

Following the principle of cascade finite-state transducers, as described in Section 3.1.1, the “Complex words” were first identified. In this specific case, “complex words” refers to small molecule mentions and bioactivity role term mentions. A dictionary-based approach was used to tag all the possible small molecule mentions, and bioactivity trigger words were also identified at this step. Sentences containing both a small molecule name and a trigger word were taken forward to the next stage.

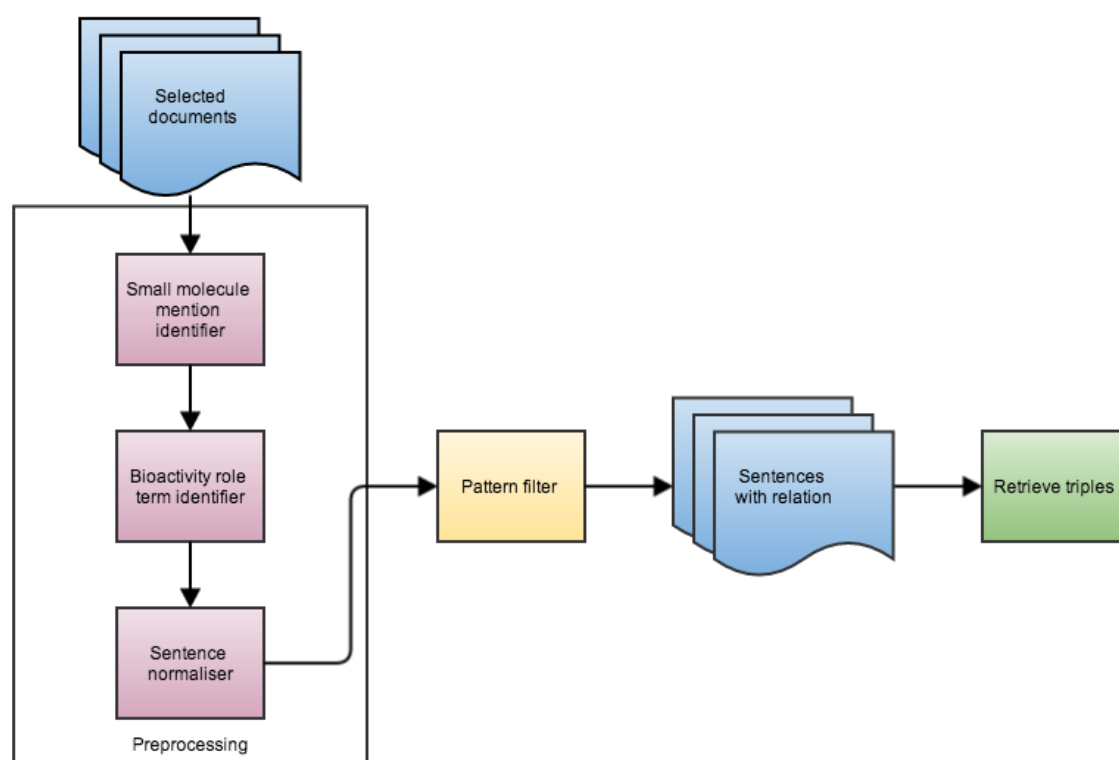


Fig. 3.3 Overview of the rule-based extraction of bioactivity role terms

Next, the sentences were further segmented into basic phrasal groups. To do this, the CIS tagger was used for POS annotation. The output of the CIS tagger shows POS labels for each word in a sentence. Based on the POS annotation, the next step is to identify complex noun phrases by combining the basic phrases as head noun with other modifiers. Once the bioactivity role mention problem is tackled, small molecule identification is performed.

3.4.3 Identifying the bioactivity role term

After the language patterns of bioactivity role terms were defined, potential bioactivity descriptions were extracted from MEDLINE abstracts. The method identifies noun phrase structures by matching syntactic language patterns using regular expressions. These hand-crafted language patterns form an alternative to syntactic parsing, which requires significant compute resources and is still error prone [105].

After candidate bioactivity role terms were identified using the previously described patterns, outliers which contained the trigger word in the wrong contexts were pruned. For example,

“Mononuclear cell growth inhibitor assay”

is not considered to represent a valid bioactivity phrase because the trigger word *inhibitor* is not the head noun (which in this case is *assay*). The approach taken to correctly filter out correct noun phrases is based on hierarchically organised language patterns. This was previously used in the extraction of protein noun phrases in a protein-protein interaction pipeline [49].

Domain-independent role terms were eliminated after the previous step, which resulted in the trigger word contexts most likely to contain a bioactivity description. Referring back to Section 3.4.1, each bioactivity role term consists of two parts, namely the head noun and the modifier. A dictionary-based rule was once again applied to the phrase. The rule says that if the context of a bioactivity role term contains a biological entity within the modifier part of the sentence, the full noun phrase is likely to be a bioactivity description; otherwise it is ruled out.

An analysis of the target types was carried out by tagging the modifier using four taggers for named entity and concept label identification: UniProtKB [8], Organ [106], Organisms [10], and GO [106]. These were applied to the modifiers of candidate noun phrases extracted from MEDLINE.

3.4.4 Classifying bioactivity terms

After extracting bioactivity descriptions from MEDLINE, the aim was to find an efficient method of classifying all the candidates with a high rate of recall, in order to allow for variability in natural language descriptions of bioactivity events.

In cases where the entire modifier is annotated by a tagger, its semantic type can normally be identified unambiguously. For example,

“CaM kinase I activator”

is easily classified as a protein activator since the whole modifier *CaM kinase I* would be annotated as a protein name (by reference to UniProtKB). However, in the majority of cases, it was found that the result is a nested case in which the semantic tagger annotates just part of the modifier, i. e. the tagged result resides within the boundaries of the whole phrase for the modifier. For instance,

“Agkistrodon blomhoffi ussuriensis protein C activator”

In this case, *ussuriensis protein C* is the real target of the activation, though *Agkistrodon blomhoffi* (a species of snake) is identified as belonging to the target. A simple method to rule out un-associated tagging is used. The tag which is in the position within the modifier nearest the role term is retained, ignoring other tags. In this example, the tag nearest the role term is the protein type *ussuriensis protein C*, not the preceding tag of species type (*Agkistrodon blomhoffi*).

An analysis carried out earlier provides a way of understanding the overlap between entity types. The analysis (described in Section 3.3.1) applied a method called cross tagging, meaning using a dictionary-based approach to identify the relation between two types of entity. The result of this process has been previously shown in Table 3.3 (page 59). The result implies that protein mentions are normally nested inside biological process mentions, and that organ and species mentions are usually nested inside protein mentions. According to this finding, identifying the largest extent of entities can find the true target (the head of the modifier of the bioactivity role term) by applying four dictionaries in the following order: Biological Process, UniProt, Organ, Organism.

3.5 Verb phrase as event trigger

As shown in Figure 3.1, the first stage of the process selects documents according to the phrasal type of the trigger word. The second rule-based method used to extract bioactivity

event presentations uses the verb forms of trigger words, defined in Table 3.1. The verb trigger words indicate the bioactivity type. The rules applied to extract bioactivity events from documents in the verb category will now be described.

Figure 3.4 uses a dependency parser tree to illustrates the following bioactivity event presented in this form:

“Vancomycin inhibits cell wall synthesis”

In this sentence, the subject is the chemical *vancomycin*, and the predicate verb phrase (VP) is *inhibits cell wall synthesis*. This example presents an active verb, but often passive verbs are used in writing, and especially in scientific and technical writing where the passive form is common. A passive verb always requires a form of the auxiliary verb *be*, and when a passive sentence contains an agent it is presented in a prepositional phrase headed by *by*. For example, Figure 3.5 shows a dependency tree of a passive presentation of the example from Figure 3.4.

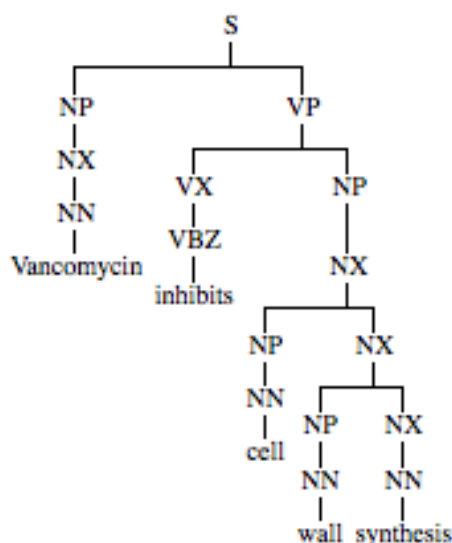


Fig. 3.4 A dependency parser tree of an active sentence

Consequently, another set of rules were formulated for passive sentences, and language patterns were defined for use in the extraction system. These language patterns will now be described.

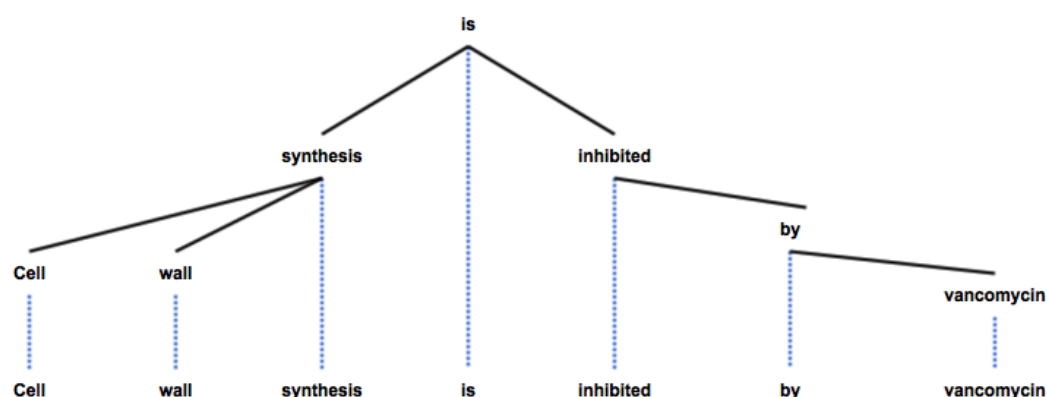


Fig. 3.5 A dependency parse tree of a passive sentence

3.5.1 Patterns of bioactivity verb phrases

The second class of language patterns was generated according to the predicate verb structure explained previously. Trigger words were used in the verb-based language pattern. The verb predicate also involves a target biological object to complete the sentence, so the most basic language pattern of the predicate can be “<TRIGGER><TARGET>”. The following two examples show the actual pattern that can be frequently seen in text:

“inhibit protein kinase”

“activate the transcription of DNA”

The passive verb type of sentence shown here therefore has a relatively simple structure. The target of the bioactivity is the object of the sentence, and the small molecule is the subject of the sentence. The pattern is denoted VP1 in Table 3.6.

As an alternative, when the sentence is in the passive form, the small molecule is often in a prepositional phrase that follows the trigger verb. Therefore the basic language pattern can be defined as “<TARGET><be><TRIGGER><prep><SM>” where <prep> denotes a preposition, and, as before, <be> represents all inflectional morphological variants of the verb *be*. This is denoted pattern VP2.

After language patterns were generated, they were further extended to include alternative patterns which match the normalised biomedical text. The pre-processing step of the verb phrase extraction system will now be further explained, including essential components such as sentence normalisation and noun phrase filtering.

Table 3.6 Patterns for bioactivity verb phrases

VP1	<SM><TRIGGER><TARGET>
VP2	<TARGET><be><TRIGGER><prep><SM>

The implementation of the verb phrase rules also consists of two major steps, which are pre-processing and the main extraction component.

3.5.2 Pre-processing

After documents were selected, a pre-processing step was applied to the document collection in order to remove noise that may affect the extraction performance. As shown in Figure 3.6, the pre-processing flow first starts with a ‘sentenciser’, to split the document into sentences, and a sentence normalisation step to replace entity mentions in the sentence with placeholder tokens. Dependency parsing is used to detect noun phrases. Then the rule filtering can be applied to remove the false positive sentences. Definitions of the pre-processing stages are given below.

Sentenciser Extracting at the sentence level obtains results of higher accuracy, so the first step of pre-processing is to split each selected document into sentences.

Sentence normalisation At this step, all the entity mentions in the sentence are normalised into a placeholder token representing their semantic type. There is evidence the normalised structure enhances the next step, dependency parsing, because variability of the biological and chemical names can introduce confusion and affect the performance of the parser. This process is also known as entity linking.

Parsing The Gdep parser [104] is employed to parse each sentence and assign syntactic labels to each token. This analysis is used in order to further rule out useless tokens. This step also provides the input for the construction of domain phrases.

Identifying domain noun phrases As the prior analysis of language patterns showed, the small molecule mention and target mention both have to be noun phrases. Identifying the head of the noun phrase is a crucial step to further filter out false positives. At this step, all noun phrases which contain one or more semantic labels are analysed.

Rule-based filtering This is a further filtering step, which aims to filter out false positive sentences. For instance, the sentence

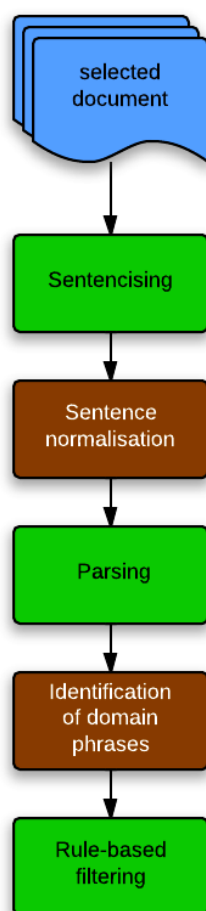


Fig. 3.6 Pre-processing work flow for bioactivity verb phrases

“Meningococcal penicillin-binding protein 2 activates dendritic cells through toll-like receptor 4”

has the three ingredient terms for a potential verb phrase bioactivity event, which are small molecule (*penicillin*), target biological entity (actually two, both proteins: *penicillin-binding protein 2* and *toll-like receptor 4*), and verb trigger word “activates”. However the small molecule mention is not the head of the noun phrase, so that the sentence does not present a bioactivity event of a small molecule. It should therefore be filtered out. The rule-based filtering takes three steps to remove sentences like this. The three steps are:

1. Small molecule noun phrase filtering. Once a sentence is detected where none of the SM mentions is the head of a noun phrase, the sentence is discarded. The justification for the filtering is that the small molecule mention has to be a noun phrase in order to induce a bioactivity.
2. Trigger verb phrase filtering. If none of the trigger words is the head of a verb phrase, the sentence is discarded. This filtering scans the fragment for the trigger words to make sure the trigger is the main verb in a predicate verb phrase.
3. If the object of the trigger word is not one of the target types listed in Section 3.3, the sentence is discarded. This filtering step ensures the sentence has at least one interesting biological target.

3.5.3 Special cases

In generating patterns, it was noticed that coordinate structures are particularly prevalent with verb forms of trigger words. Also allowances were necessary for verbs used in the active and passive voices. These were handled in the following way:

- Multiple small molecule mentions in a coordinate structure. If a bioactivity phrase contains more than one agent small molecule (as tagged by the NER), they will all be captured and presented in the same bioactivity. For example, from the sentence

*“Clomipramine, paroxetine and fluvoxamine did not significantly **inhibit** male sexual behaviour”*

the (negated) trigger word *inhibit* was first detected, followed by the target, a GO term (*male sexual behaviour*) and three small molecules (*Clomipramine*, *paroxetine* and *fluvoxamine*). Then three different bioactivities were derived from the sentence according to the three different agents. The three bioactivities are:

Agent: Clomipramine

Bioactivity Type: not inhibit

Target: male sexual behaviour

Target type: GO

Agent: Paroxetine

Bioactivity Type: not inhibit

Target: male sexual behaviour

Target type: GO

Agent: Fluvoxamine

Bioactivity Type: not inhibit

Target: male sexual behaviour

Target type: GO

- Multiple targets in a coordinate structure. Similar to the previous point, if NER returns multiple potential target entities in a bioactivity phrase, the targets will be included in the same bioactivity. For example, in the sentence

“Aspirin inhibits both COX-1 and COX-2”

two different bioactivities can be extracted:

Agent: Aspirin

Bioactivity Type: inhibit

Target: COX-1

Target type: protein

Agent: Aspirin

Bioactivity Type: inhibit

Target: COX-2

Target type: protein

- Active versus passive voice used in the bioactivity event. As mentioned in Section 3.5.1, the active format of bioactivity begins with one or more small molecules as the subject, and is completed with a verb predicate followed by its/their target(s). Another possible structure of the bioactivity event is the passive voice. In this structure type, a prepositional phrase usually starts with the word “by” after the bioactivity event trigger word, and the words prior to the trigger word can be one or more biological target(s).

3.6 Bioactivity triple extraction

To sum up, rule-based extraction uses a pattern-based approach to match basic syntactic frames of text. Four abstract patterns for bioactivity have been defined, as shown in Table 3.7. They are derived from the factors discussed in this chapter. The patterns were applied to MEDLINE abstracts to find bioactivity role term descriptions of ChEBI entities. The extraction process was designed to extract bioactivity triples which contain a) a small molecule term; b) a term of bioactivity type, which presents the relation type; and c) a term representing the target of the small molecule which is often contained in the role term noun phrase.

Table 3.7 Patterns for general bioactivity event extraction

ID	Pattern
NP1	<SM><TARGET><TRIGGER>
NP2	<SM>,<TARGET><TRIGGER>
VP1	<SM><TRIGGER><TARGET>
VP2	<TARGET><be><TRIGGER><preposition><SM>

Once a candidate event is obtained for a sentence the presence of negation words in the surrounding context must be detected. A simple rule was implemented by scanning the bioactivity phrase for a class of negation cues. Essentially the process is the same as described in Chapter 2. If a negation cue is detected in the immediate context of the bioactivity, it is recorded in the bioactivity triple. Empirically it was found that negated events count for approximately a third of all the bioactivity events found in text.

3.7 Chapter summary

In this chapter a rule-based system for extraction of small molecule bioactivity was presented. The system consists of two main units each based on two extraction rules. The two rules followed the basic principles of a lexical system, using a pattern language to describe the semantic structure of the bioactivity event. A description of the implementation of the system was provided. The next chapter will present results and evaluation from using this system on MEDLINE abstracts.

Chapter 4

Results and discussion

In the first half of this chapter, the results for extracting the relation between drugs and bacteria using a machine learning approach will be shown and discussed. These results are divided into two sections. Firstly, the result of the document classification task, including feature optimisation and the effect of different feature combinations on classification, will be presented. Secondly the relation extraction results based on the sentence classification will be provided and discussed.

The second half of the chapter will show the results from using the rule-based system to extract small molecule bioactivity events. The section will first compare feature distribution in MEDLINE, and further show the coverage of the defined patterns. Finally, a manual assessment of the rule-based extraction pipeline will be illustrated and discussed.

4.1 Performance parameters

The system performance is usually reflected using the following performance measures from information retrieval: precision, recall, and F-measure. Precision is the percentage of results that are correct. If TP is the number of true positives (correct results) and FP the number of false positives (incorrect results), then

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.1)$$

Recall is the percentage of true positives found relative to the total number of positives in the corpus. If FN is the number of false negatives (i. e. correct results that were missed),

then

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.2)$$

F-score is a weighted average of precision and recall. The general form is

$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \quad (4.3)$$

The parameter β can be tuned to prioritise precision or recall as required for a particular application. However the most common usage is the F_1 -score, which treats precision and recall equally. It is the F_1 -score which will be used here. In this case the calculation simplifies to the following:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.4)$$

Precision, recall, and F_1 -score will be reported for each experiment.

4.2 Results from the drug-bacterium study

Recall that in the machine learning experiment, the focus is on extracting a single type of bioactivity event: effects of drugs on bacteria. A support vector machine was used at two levels: firstly to select documents potentially bearing relevant bioactivity descriptions, and secondly to select relevant sentences within those documents. Selected sentences were then processed using a simple rule-based approach. The results of this experiment will be presented in this section.

4.2.1 Document classification results

Quantification of performance of the machine learning approach was initially computed using the DrugPro corpus we created for this task. The DrugPro corpus was originally created using 400 documents, however on inspection it was decided to remove five documents because they were unsuitable. In total 4,464 distinct features were generated for each document. The resulting training vector therefore consisted of $4,464 \times 395$ elements. The results from using this vector with the SVM classifier were a recall of 82.11%, precision of 83.87%, and F-score 82.98%. In order to improve the performance of the classifier, as well as to find the optimal values for the size of the feature vector used, a tuning step was introduced by thresholding the χ^2 of each feature.

Table 4.1 Classification performance using various BOW vector sizes

Feature set	Precision	Recall	F-score
Top 1,000 terms	93.68%	86.41%	89.90%
Top 1,200 terms	93.68%	87.25%	90.35%
Top 1,400 terms	91.58%	90.62%	91.10%
Top 1,500 terms	90.53%	90.53%	90.53%
Top 1,600 terms	90.53%	90.53%	90.53%
Top 1,700 terms	91.49%	90.53%	91.01%
Top 1,800 terms	91.49%	90.53%	91.01%
Top 2,000 terms	90.53%	92.47%	91.49%
Top 2,500 terms	90.53%	89.58%	90.05%
All terms (4,464)	82.11%	83.87%	82.98%

Feature selection using TF*IDF

A χ^2 value was computed for each feature using equation 2.3 on page 39. The features were ranked in order of descending χ^2 value. A number of features was then selected by applying a cut-off threshold to the ranked feature list. The TF*IDF values of the selected features were submitted into the SVM classifier to obtain another set of results. There were nine feature sets generated using this method, ranging from 1,000 to 2,500 features. The set sizes and associated performance parameters are shown in Table 4.1. The trendline of performance is plotted in Figure 4.1.

The results show that, in this study, using TF*IDF scores as a feature for the SVM provided good performance, but there is a strong case for thresholding. The first attempt using the entire set of features achieved an F-score of 82.98%. Using χ^2 as the cutoff feature improved the performance of the SVM, with performance increases of 10% at a feature vector size of 2,000. Performance slightly decreased again for feature vector sizes from 2,000 to 2,500.

Domain-specific features

As described in Section 2.5.2, other domain-specific features were also used. These were the frequency of drug and bacterium mentions, the presence of keywords, and co-occurrence of drug and bacteria names. These domain specific features were computed and used as input to the SVM classifier. Various experiments were performed and the results are shown in Table 4.2.

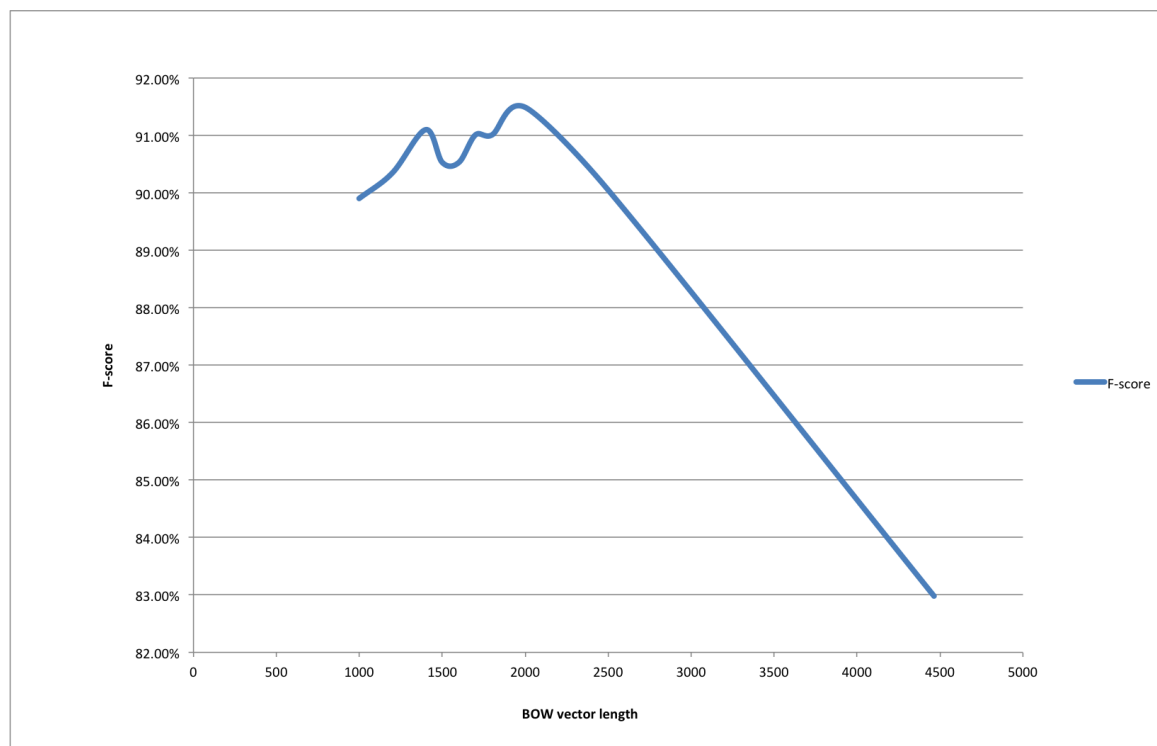


Fig. 4.1 Trendline of SVM performance based on BOW vector size

Table 4.2 Classification performance using other features.

Feature set	Precision	Recall	F-score
Drug mentions	72.00%	76.60%	74.23%
Bacterium mentions	70.00%	77.78%	73.69%
Keyword TF*IDF	51.58%	72.06%	60.12%
Drug-bacterium co-occurrence (document level)	83.00%	91.21%	86.91%
Drug-bacterium co-occurrence (sentence level)	88.42%	86.60%	87.50%

Table 4.3 Classification performance using other concatenated features

Feature set	Precision	Recall	F-score
TF*IDF + drug mention	90.53%	92.47%	91.49%
TF*IDF + bacterium mention	90.53%	92.47%	91.49%
TF*IDF + drug mention + bacterium mention	91.40%	93.12%	92.25%
TF*IDF + key word	90.53%	92.47%	91.49%
TF*IDF + co-occurrence of pair (document level)	92.00%	93.22%	92.60%
TF*IDF + co-occurrence of pair (sentence level)	92.56%	93.22%	92.88%

Feature ablation testing

In [66], it was found that when term weights were combined with other domain-specific features, precision decreased and recall increased. A similar approach was applied here by combining each of the single domain-specific features with the 2,000 TF*IDF feature set which previously showed the best performance. Each domain-specific feature is tried individually and systematically in order to assess its impact. The results are shown in Table 4.3.

Somewhat surprisingly, according to these results a minimal improvement of performance can be seen by concatenating domain-specific features with TF*IDF features. The combination of features does not affect SVM. The reason could firstly be that TF*IDF features are the dominant features of the classifier, and the domain-specific features were not good enough. Secondly, it is possible that the 2,000 feature set run reached the highest performance that can be achieved with this SVM model for this classification task.

4.2.2 Sentence classification results

For the second stage of the cascaded classification approach, the task is to extract relations found between drugs and bacteria at a sentence level. The sentence classifier was trained on the DrugPro corpus, which contains 400 manually annotated abstracts and has 3,800 sentences. Among these 3,800 sentences, only 760 are qualified as positive sentences in which one or more relations exist. The whole set of 3,800 sentences was parsed using the Gdep dependency parser and dependency trees were constructed as features, then used as input to the SVM classifier. The performance of the classifier was very poor with only 32.4% for precision and 32.0% for recall. To remedy this, the training set was balanced by randomly selecting 800 negative (i. e. without bioactivity) sentences to go with the positive sentences. This created a balanced training set of around 1600 sentences. With this training set the performance of the classifier increased for both precision and for recall, with 56.7% and 51.9% respectively. It is therefore concluded that a balanced training set is crucial and majorly affects the performance of the classifier.

Table 4.4 Sentence classification performance using other concatenated features

Feature set	Precision	Recall	F-score
Parse tree feature + drug mention	56.7%	51.9%	54.2%
Parse tree feature + bacterium mention	56.7%	51.9%	54.2%
Parse tree feature + co-occurrence of drug and bacterium	56.7%	51.9%	54.2%
Parse tree feature + co-occurrence of drug, bacterium and key word	58.9%	57.4%	58.1%

The SVM classifier is further trained by using a combination feature set. The effect of each feature was determined by combining it with the features from the dependency parse tree. As can be seen in Table 4.4, there is no change when the parse tree features were concatenated with the occurrences of drug mentions in the sentence, nor bacteria mentions. When the parse tree feature is combined with the binary feature “co-occurrence of drug, bacteria and key word”, there are improvements to precision (+2.2%) and recall (+5.5%). Therefore the conclusion was reached that domain-specific features do not substantially contribute to sentence classification.

4.2.3 Results of relationship extraction

In this subsection, the results of the rule-based relationship extraction are reported. The extraction was based on the language patterns constructed through analysis of bioactivity keywords, as described in Section 2.7. Four groups of patterns were drawn from the Drug-Pro corpus using four sets of keywords proposed by domain experts. The corpus consists of 2,931 sentences, in which 801 sentences (including duplicates) are system had been extracted by the sentence classifier as relationship-bearing sentences. The sentences were organised into four groups according to the keyword used in the depiction of bioactivity.

Based on these groups, the corpus was manually annotated by two curators. This required all the relationships in the set of likely positive sentences to be manually identified by hand; that is any drug-bacteria relationship found in a sentence was explicitly labelled together with the participating entities. It is important to note that this manual annotation of this group of sentences is the benchmark for recall calculation. That is, performance (specifically recall) is measured relative to the findings of the manual annotators when evaluating the 801 sentences extracted by the sentence classifier. It cannot be assumed that the results are transferrable to the whole MEDLINE collection.

Once the annotated corpus was available then the relationships extracted using the machine learning algorithm were validated. The results are shown in Table 4.5. The “Resistant/susceptible/sensitive” category is dominant, extracting 473 relationships; the “Resistance/susceptibility/sensitivity” category shows fewer relationships extracted from the cor-

Table 4.5 Performance of the keyword-based extraction on the filtered sentences

Keyword	Annotated by curators	True positives	False positives	Precision	Recall
Resistant / susceptible / sensitive	473	414	0	100.0%	87.5%
Resistance / susceptibility / sensitivity	76	32	24	57.1%	42.1%
Against	52	40	7	85.0%	76.9%
Other key-words	43	36	3	92.3%	77.7%
No relationship detected	157	33	11	75.0%	21.0%
Totals	801	555	45		

pus. There were 157 sentences extracted by the sentence classifier which had no detected relationships (false negatives).

The table also summarises the performance of the technique for each keyword. The grouping of “Resistant/susceptible/sensitive” has the best performance compared to the others. The language patterns that were used in this category capture a highly accurate level of relationship, however a portion of the relationships were misannotated by the system. The reason for this resulting in false negative cases will be further investigated in a later section. Surprisingly, the “Resistance/susceptibility/sensitivity” group did not have good performance, with only 57.1% precision and 42.1% recall. However in the “Against” group, both precision and recall are good, at 85% and 76.9% respectively.

4.2.4 Analysis of errors

The relationship extraction results were manually examined to determine the main factors which led to the false positives and false negatives. Overall it can be said that the adopted approach showed high precision but low recall. To speculate about the general reasons for errors, most false negatives are caused by sentences that were incorrectly classified by the upstream processing (i. e. the sentence-level SVM). A second reason might be the complexity of natural language which was inadequately covered by language patterns. The following is an interpretation of the error analysis for each category.

In the “Resistant/susceptible/sensitive” category, the false positives are mainly caused by the performance of the SVM classifier, in particular when the sentence classifier fails to discard an incorrect relationship-bearing sentence. In total, 70 sentences were incorrectly

identified as positive relationship-bearing, therefore the relationship extraction subsystem could not recognise a bioactivity pattern between the entities in the sentence. For instance, the classifier judged the sentence

*“Twenty-one patients with multidrug-resistant (MDR) **Acinetobacter baumannii** and **Pseudomonas aeruginosa** pneumonia were treated with nebulized polymyxin E (colistin).”*

as a positive relationship-bearing sentence, however the extraction system did not identify a relationship between bacteria *Acinetobacter baumannii*, *Pseudomonas aeruginosa* pneumonia and drug *polymyxin E*. This is correct, because the sentence only states the experiment of testing multidrug-resistant patients with a certain drug, but the outcome of the test is not mentioned.

The “Resistance/susceptibility/sensitivity” rule shows poor performance on both precision and (particularly) recall. This has been caused by a relatively high number of false negatives, i. e. where an event was present in a sentence but was not extracted. False negatives are often caused by pattern limitations. For a pattern to be successful it must be able to find the relationship keyword together with named entities in expected positions in the surrounding context. In many false negative cases the patterns match a short phrase but are unable to fully process the annotation of the entities, perhaps due to natural language variations.

In the “Against” category, the constructed language patterns produced a reasonable outcome. For example, the sentence

*“The MIC of seven antimicrobials (**ampicillin**, **erythromycin**, **roxithromycin**, **clindamycin**, **tetracycline**, **minocycline**, **nadifloxacin**) against **P. acnes** was under 3.13 micrograms/ml”*

describes an “against” relationship between a group of antimicrobials and *P. acnes*. First the sentence was classified as a positive relationship-bearing sentence by the SVM classifier, then the rule-based system detected the “against bacteria” pattern and confirmed the relationship type, “against”.

The false negatives yielded in this category are often from an incorrect judgement by the upstream SVM classifier. For instance, the sentence

*“**Cefepime** had excellent activity against **Enterobacteriaceae**, including eight presumptive extended-spectrum **beta-lactamase** producers and 33 stably derepressed mutants of natural **cephalosporinases**”*

was classified by the SVM as a negative relationship-bearing sentence, so it was not processed by the downstream relationship extraction subsystem. This is unfortunate because the relationship is presented in a form that appears to be suited to the extraction patterns.

False positives within the category are mainly caused by the part-of-speech variation of words. As one example,

*“Among the flavonoids examined, four flavonols (*myricetin*, *datiscetin*, *kaempferol* and *quercetin*) and two flavones (*flavone* and *luteolin*) exhibited **inhibitory** activity against *methicillin-resistant Staphylococcus aureus* (MRSA).”*

does not exhibit an “against” relationship between drugs mentioned and *MRSA*. Instead the prepositional phrase headed by “against” modifies the word “activity”; it is identified as a preposition rather than a verb.

The rest of the keywords appear with lower frequency. The rule applied to this set was extracting a basic “subject-verb-object” pattern out of the normalised text. The results showed promising performance in terms of precision, however relatively low performance on recall. Again, this is because the language patterns were not flexible enough to cover the phenomena occurring in real text. In one instance, the sentence

*“LAMPs from *M. penetrans*, *M. genitalium*, *M. salivarium*, *M. pneumoniae*, and *M. orale* also had an **inhibitory** effect on *glucocorticoid receptor (GR)* response to hormone *dexamethasone* (Dex) in TSU transfectants.”*

contains a bioactivity event: an inhibitory relation between the listed bacteria and *dexamethasone*. However the noun phrases of bacteria and drug are indirectly related.

In summary, the error analysis provides some explanation for the observed performance of the relationship extraction system on each type of keyword. Most false negatives were generated by false negatives from the sentence SVM, and the incorrect judgements would lead to missing results in the relationship detection. Secondly, rule-based systems often have a problem with recall, and in this particular task the limitation of pattern-based detections showed lack of coverage of relationship-bearing sentences, which also led to the system missing correct annotations.

4.2.5 Drug-bacteria bioactivity in the DrugPro corpus

Through the identification of relations between drugs and bacteria, very few bioactivity events were detected in relation-depicting phrases (i. e. those containing event trigger words).

The text from the corpus presents drug/antibiotic bioactivity by interaction with one bacterium. Often bioactivity is described in terms of the biological effects at a higher level of abstraction, for example on the organism as a whole. In total, only six bioactivity events were detected from the corpus.

It can be concluded that the bioactivity event of a drug with bacteria as a target is not a representative event type. This is shown by the insignificant quantity collected from the corpus. Even though bacteria hypothetically are the main species target of drugs, there is a lack of evidence of this type of bioactivity through the methods employed here. On the other hand, it might be that the data collection resulted in insufficient bioactivity data of the type we were looking for. Many descriptions of bioactivity in the corpus may be implicit, or may be speculative. This highlights a weakness of the current system and demonstrates the complexity of the task.

4.3 Results from the general bioactivity study

The results in this section will be presented as follows. Firstly an analysis of bioactivity targets will be presented, showing the results from the NER processing and the distribution of entity types found to be associated with the bioactivity trigger words.

Secondly, the analysis of the small molecule NER will be presented. This will include the previously described hybrid approach of customised OSCAR3 and Jochem chemical recognisers.

Then an initial data set of 20 hand-annotated abstracts was used to manually evaluate the performance of the rule-based extraction system. Analysis of system performance against this data set will initially be presented, followed by the results from running the system on a downloaded copy of the whole MEDLINE collection.

4.3.1 Analysis of target entity recognition

Table 4.6 shows the results after pre-processing the collection of abstracts. The data provide an overview of the co-occurrence of target type (from NER) and trigger words. Each cell shows the unique count of semantic tagging for a certain feature. Both nested and exact matching on the modifier of the bioactivity terms are considered.

In general, it was found that protein names were the most common entity found with a biological role term. Co-occurrence of UniProtKB-based NER tagging and the keyword ‘inhibitor’ gives a high number of unique hits: 336,420. This represents more than a quarter

Table 4.6 Co-occurrence of bioactivity role term and NER tag, as found in MEDLINE abstracts

Bioactivity role term	Target				Total
	Protein	Organ	Organism	Biological process	
stimulator	2,526	3,303	500	1,808	8,137
adaptor	3,729	100	133	1,016	4,978
modulator	7,847	1,468	536	4,204	14,055
messenger	10,056	1,186	1,151	3,876	16,269
agent	10,522	10,292	19,374	8,744	48,932
blocker	13,588	1,371	9,235	4,203	28,397
toxin	16,890	1,583	10,265	3,276	32,014
suppressor	18,534	1,301	2,382	2,988	25,205
regulator	27,724	5,469	2,802	27,270	63,265
factor	40,427	21,959	11,152	77,670	151,208
agonist	48,973	3,633	13,154	12,353	78,113
activator	71,165	1,745	3,895	19,376	96,181
antagonist	80,932	9,483	11,740	19,486	121,641
inhibitor	336,420	12,102	30,839	142,289	521,650
Total	689,333	74,995	117,158	328,559	1,210,045

of the total results found. These data therefore suggest that bioactivity descriptions in text most commonly refer to processes inhibiting a protein or enzyme. Two such examples are:

*“Other lysosomal hydrolases are not **inhibited** by N-bromoacetyl-beta-D-galactosylamine, with the exception of ‘neutral’ beta-glucosidase glucohydrolase.”*

*“At the biochemical level cardiac guanylyl cyclase activity is enhanced 2–3 times with acetylcholine and this enhancement is completely **blocked** by atropine.”*

There are not as many hits in the organ and organism groups. A few true positive examples, such as *bothrops jararaca* inhibitor and thyroid stimulator can be found. However, there are many examples in which the organ or organism appears in the sentence only to denote the location of the bioactivity being described. For example:

“Caesium ion: antagonism to chlorpromazine- and L-dopa-produced behavioural depression in mice.”

This sentence is actually referring to how a Caesium ion was found to lessen the effects of chlorpromazine and L-dopa on mouse behaviour. Therefore despite the co-occurrence of all the required entities, it is not a bioactivity but a false positive.

*“The changes in the contents of glycolytic intermediates in the **livers** indicate that the **phosphoenolpyruvate carboxykinase** [EC 4.1.1.32] reaction is **inhibited** by **tryptophan** administration in all groups of **rats**.”*

This is a true positive but particularly challenging to process because the sentence contains multiple target entities. The bioactivity of interest is the inhibitory effect of tryptophan on the enzyme phosphoenolpyruvate carboxykinase.

*“The oral administration of **metaproteranol** increased the leukocyte **adenyl cyclase** activity which was **stimulated** by **NaF** and decreased the count of peripheral eosinophils in some of the **monkeys**.”*

There are actually two bioactivities here: the separate effects of metaproteranol and sodium fluoride (NaF) on the protein adenyl cyclase. Once again the organism in the sentence is not the target, and so in this context is essentially noise.

In sum, organs and organisms most commonly provide the contextual information about where a bioactivity takes place, rather than being themselves the target of the bioactivity. This makes Organ and Organism categories particularly problematic for rule-based extraction. However the organ/organism information is still likely to be of interest to practitioners. This conclusion could be used to influence future ontology models (in the manner described in Section 1.1.5), to support capture of this additional level of detail beyond the basic bioactivity triple.

Cases where GO terms were tagged in bioactivity terms were also analysed, for example, *inhibitor of DNA transcription*. Here, a biological process is the target of the bioactivity term.

As expected from these types of text mining approaches, there are also typical false positives in the result caused by natural language, such as ‘hand’ being tagged as a body part in sentences containing *On the other hand*, and ‘dialysis’ being incorrectly tagged as a species in the sentence

“Influence of peritoneal dialysis on factors affecting oxygen transport”

(Dialysis is also the name of a species of insect, but this is not meant here.)

As usual, care needs to be taken regarding negation, in that some of the results reflect sentences where the bioactivity being described is explicitly reported as not taking place. We tried to cover these cases by applying a negation filter.

4.3.2 Analysis of small molecule recognition

To identify the small molecule (chemical) entities in the phrases found by the previous stage, the dictionary-based approach using Jochem was compared with the results generated using OSCAR3 (which is able to identify novel chemical names in text using a machine learning approach). Table 4.7 shows the frequency of each triple mentioned in the text together with the unique count of triples before and after the rule-based filtering described in Section 3.4.

OSCAR3 yields many more triples than Jochem does. This is expected, since OSCAR3 recognises any chemical-like string. However, OSCAR3's approach also results in a considerable number of false positives due to its recognition of chemical-like nomenclature appearing as a component in larger stretches of text (such as protein names). Furthermore, a smaller number of triples identified by UniProtKB and OSCAR3 are observed compared to the set identified by UniProtKB and Jochem. This is because OSCAR3 produces annotations that nest within a protein mention in the sentence and thus lowers the subsequent annotation protein mentions. Jochem performs more long-form matching than OSCAR3 does, therefore the subsequent protein identification has a higher likelihood of identifying a protein term within the sentence, hence yielding a greater number of triples.

The comparison of before and after filtering shows whether the triple mention is accidental and the association between the chemical and the other semantic group is more than contextually related. Between chemicals and proteins the ratio is smaller than the other groups. The non-unique number of triples is fewer than twice the number of unique ones, while it is more than this ratio in other groups (specifically in the chemical organ group). The number of non-unique triples identified by Jochem after filtering is almost three times the unique count.

As pointed out in [30], one of primary types of object of OSCAR3 NER is ontology terms from ChEBI, that is OSCAR3 was meant to be developed as an appropriate tool for the extraction and verification of chemistry in the ChEBI ontology. However OSCAR tends to identify chemical name fragments and gives good coverage over the ChEBI ontology. Jochem is a NER that is more domain-specific, since it combines one of the largest drug dictionary, DrugBank, as well as other biomedical drug type dictionaries, such as UMLS. It is a dictionary developed to identify small molecules and drugs in free text.

Table 4.7 Analysis of triples from MEDLINE

Chemical tagger	Filtering	UniProtKB		Organ		Organism		GO	
		Unique	Non-unique	Unique	Non-unique	Unique	Non-unique	Unique	Non-unique
Jochem	Before	4,114,286	7,853,314	2,666,468	7,148,677	1,785,771	4,076,253	1,244,099	2,947,289
	After	2,912,756	5,457,529	1,632,855	5,302,115	1,394,310	3,085,056	935,864	2,089,163
OSCAR3	Before	11,599,131	23,988,686	4,344,247	11,855,944	2,672,206	5,836,725	1,864,403	4,607,315
	After	7,827,737	12,776,542	2,222,450	4,598,353	1,347,442	2,338,487	945,320	1,804,411

Table 4.8 Filtering result

Filter	Number of sentences remaining
Raw input	207
Anchor filter	70
Agent and target filter	49
SM and target noun phrase Filter	29
Bioactivity role term filter	24
Pattern filter	16

4.3.3 Manual assessment of the rule-based extraction pipeline

Before applying the rule-based processing to the whole MEDLINE abstract collection, 20 abstracts were initially processed and the results manually verified. The 20 abstracts were selected at random. The idea was to have a relatively small set to ‘pipeclean’ the process.

The resulting annotations of bioactivity output were manually checked by curators in order to judge the performance of the system. The small molecule bioactivity event evaluation takes into account the Agent-Target pair and the bioactivity type. Lee *et al.* [49] introduced a per-pair evaluation method, which was used to validate the result of extracted protein-protein interaction pairs. A bioactivity event can result through interaction between multiple Agent-Target pairs, therefore judging the identification of the relation type between each pair can be a viable solution to vote based on the prediction labels assigned to each pair instance. Precision and recall are computed based on the number of distinct pairs (not instances) that are classified correctly. Consequently, the per-pair precision, recall and F-score were calculated. Manual evaluation meanwhile brings benefit to improve the extraction system since an error analysis was also given and interpreted. This process is believed to be helpful for the future development of the extraction system.

Results were retrieved from each stage of processing for analysis. Additionally an error analysis of false negatives and false positives was performed.

The stages of processing act as cascaded filters. Table 4.8 illustrates the effect of each filtering stage. As can be seen, the 20 abstracts were sentencised into 207 sentences in total. After the first stage, two types of event anchors were applied as introduced in Chapter 3. In this way 137 sentences were eliminated due to lack of a signal of small molecule bioactivity event. These sentences were further classified according to whether they had a “bioactivity role term” or a “verb type event anchor”, with 23 sentences found in the former group and 47 sentences in the latter.

The second filter removed sentences lacking a pair of event participants. In this particular case, the participants are Agent and Target. Using this filter 21 sentences were removed.

The system then judges the agent and target noun phrase. If the participants of the bioactivity event do not have a small molecule or predefined target noun as the head of the phrase, the sentence is rejected. For example, in the sentence

*“cyclosporin A, synergistically **suppresses** the T cell response to MiHC and MHC in mouse and in human”*

the system has correctly identified the event anchor *suppresses* and the small molecule mention *cyclosporin A*. The filter correctly identified *cyclosporin A* as the head of a noun phrase. However, neither of the valid target entities *mouse* or *human* appeared as the head of a noun phrase, so the system judges the target is absent in the sentence. Therefore the sentence was correctly removed by the filter. In this way, 20 sentences were removed.

This filtering rule is based on the fact that often when a sentence is identified by multiple semantic taggers, entities of different semantic types are nested or overlapped with each other. This makes such sentences difficult to deal with. The adopted solution makes use of the concept of a noun phrase having a head noun. The head of a noun phrase denotes the main subject or object of a relation. When the small molecule is the head of a noun phrase and a valid target entity is also the head of a separate noun phrase in the same sentence, the probability of the sentence containing a bioactivity representation is increased. The sentence is likely to be in the form of “subject predicate” or “subject-predicate-object”.

Once the trigger word is detected, the system analyses whether it meets the criteria of a bioactivity role term. If not, the sentence is discarded. Applying this resulted in elimination of another eight sentences from the test data set. For instance, for the sentence

*“Gangliosides are potent **inhibitors** of the antiviral activity of mouse fibroblasts and other eta-interferons”*

the system detected the small molecule mention *Gangliosides* and the species mention *mouse*, and qualified them as head nouns of their corresponding noun phrases. However, the filter is designed around the premise that the bioactivity target is the “role term anchor”. This sentence was therefore filtered out because the head of the modifier (in this case *antiviral*) is not one of the defined target types, which means that the target of the bioactivity event must be missing from the sentence.

In the final step, filtering with the patterns defined in was used to find specific matching patterns in the sentences. At the output of this stage there were 16 sentences from the original test set of 207 (20 abstracts) that were confirmed to contain a bioactivity event.

Manual inspection found that these 16 sentences contained a total of 21 Agent-Target pairs. Two sentences each derived two Agent-Target pairs, and one sentence contained three Agent-Target pairs. According to the evaluation schema, these sentences were replicated in order to carry out per-pair based evaluation. In addition, one false positive and one false negative were found from the manual evaluation, and 183 true negatives were identified. Overall, the performance of the system was estimated as an F1-measure of 94.7% with 94.7% recall and 94.7% precision relative to the test data set.

Consider the one false positive case,

*“In a series of three experiments we compared various **specific serotonin reuptake inhibitors** (SSRIs) for their ability to **suppress** sexual behaviour in male **rats**.”*

The small molecule name contains the word *inhibitors* which happens to also be one of the trigger words. It is therefore likely that the sentence was classified as a false positive because either *serotonin reuptake* or just *reuptake* was in the NER dictionary.

The false negative sentence was the following

*“It is concluded that **droperidol** is a competitive **inhibitor** of the vascular **alpha-adrenoceptors**, leaving the **beta-adrenoceptors** of the **heart** unaffected, and that it **inhibits** the neuronal uptake mechanism of **noradrenaline**.”*

This sentence is particularly challenging, with three clauses and a variety of entity types. However it was found that the failure of the system was for a relatively simple reason: the NER stage had failed to annotate the protein mention *alpha-adrenoceptors*.

If we consider the entire pipeline, no other bioactivity event was found in the 207 sentences during manual inspection. We can therefore say the approach appears to be high-precision, with the caveat that the test data set is small.

It can be concluded (as might originally be assumed) that the event detection system heavily depends on the quality of the named entity recognition; true positive annotations can only be achieved if the entities are correctly identified. The NER module in the system heavily relies on the input dictionaries created from existing terminologies, which introduces the limitation that new terms in the text cannot be identified. As we have seen, NER for generalised small molecule bioactivity (i. e. against a broad set of targets) is particularly challenging.

It is suggested that a good training corpus is required to improve the performance of the rule-based system, allowing continuous assessment of rules as they are developed, and development of more sophisticated NER approaches (e. g. using machine learning). However

Table 4.9 Extraction of bioactivity events from all of MEDLINE

Bioactivity type	First rule	Second rule	Total
stimulate	135	140	274
adapt	7	2	9
modulate	297	150	447
messenger	29	0	29
agent	67	0	67
block	103	81	184
toxicate	18	0	18
suppress	30	68	98
regulate	573	106	679
factor	840	0	840
agonise	225	0	225
activate	192	149	341
antagonise	298	0	298
inhibit	3,035	384	3,419
Total	5,849	1,080	6,929

the danger of using a corpus for rule development is that the rules become too specialised (improving precision at the expense of recall).

4.3.4 Extraction of small molecule bioactivity from MEDLINE

After the manual verification of the system had been carried out, the system was applied to a downloaded copy of the whole MEDLINE abstracts collection. Processing took around 4-5 hours on a server farm. Table 4.9 shows the results of the two different rule-based extraction modules against each bioactivity anchor. From the 14 types of bioactivity events, “inhibit” is the most frequent at 3,419, out of which 3,035 bioactivity events were extracted using the noun-phrase rule, and only about a tenth of that yielded by the verb-phrase rule. “Factor” appears to be another popular bioactivity event type in the literature. Compared to other bioactivity types, “factor” seems to be less precise in denoting a bioactivity effect, unlike the event type “inhibit” which appears to be subject to less variability in how it is used. (Or, to put it another way, “factor” appears frequently but it is often used in contexts that do not denote bioactivity.) The terms “regulate” and “modulate” are the other popular bioactivity event types detected from the biomedical text. There were few mentions of “adapt”, “agent”, “messenger” and “toxicate” detected in the MEDLINE text.

Further analysis of the bioactivity event target type was performed, in order to understand which type of target is the major target of small molecule bioactivity events in the

Table 4.10 Bioactivity target type distribution

Bioactivity type	Target				Total
	Protein	Organ	Organism	Biological process	
stimulate	40	39	3	123	205
adapt	0	2	0	0	2
modulate	12	24	3	174	213
messenger	7	2	0	9	18
agent	23	0	0	44	67
block	23	22	24	47	116
toxicate	15	1	2	9	27
suppress	2	12	2	55	71
regulate	53	39	4	326	422
factor	86	34	120	225	465
agonise	268	7	57	12	344
activate	289	171	27	564	1,051
antagonise	139	4	42	10	195
inhibit	1,469	90	52	783	2,394
Total	2,426	447	336	2,381	5,560

literature. Table 4.10 shows the distribution of four different types of biological targets which were collected from the entire MEDLINE. Protein entities are by far the major target type, and “inhibit” is the dominant bioactivity event type.

It was found that the major portion of target proteins are enzymes and protein receptors. For instance, from the sentence

*“d-tubocurarine and procaine have been shown to **inhibit** the acetylcholinesterase of rat brain homogenate by coupled and non-competitive mechanism respectively, which suggests binding to enzyme peripheral sites.”*

the agent small molecule *procaine* is detected by the system, and the target is detected as *acetylcholinesterase* which is a serine protease (i. e. a protein).

Alternatively, “biological process” is also a popular target type for small molecule bioactivities. For example, the biological process mention, *secretion of PSA* is identified as the target of the small molecule *isoflavones* in the sentence

*“Previous studies have documented that isoflavones can **inhibit** the secretion of PSA in the androgen-dependent prostate cancer cell line, LNCaP, however, the effects of genistein on androgen-independent PSA expression has not been explored.”*

4.4 Qualitative comparison of the two approaches

The work has provided details of the text mining task via two different approaches. The first hybrid approach focused on drug-bacteria (prokaryote) relationship extraction from the DrugPro corpus, and used a combination of machine learning and rule-based processing. The second approach had a more general bioactivity event extraction focus and used an entirely rule-based system. The two systems are compared according to their advantages and disadvantages.

The hybrid system used feature-based SVM classifiers both on a document level and a sentence level. The SVM classifiers are trained using a combination of syntactic and semantic features. They act as filters to reduce the corpus to a collection of relationship-depicting sentences. This procedure of SVM classification is equivalent to the filtering approach carried out in the rule-based system.

As is the case in any pipelined system, the performance of the SVM directly affects the overall performance of the hybrid system. The system has provided a promising performance in term of F-score on the training corpus. However through incorrect judgement of relationship-bearing sentences, some errors still occur which limit the performance of the downstream rule-based relationship extraction. Many false negatives resulted from incorrect performance in the SVM step, when the SVM mistakenly filters out a positive sentence. On the other hand, when the SVM classification results in a false positive sentence, the performance is also affected, because the downstream pattern-matching might mistakenly identify a spurious pattern from the false positive.

The second experiment was significantly more general in terms of types of bioactivity. The approach taken was completely rule-based, and is mainly based on patterns of syntactic graphs. In contrast to the SVM approach, the rule-based approach is essentially bottom up, i. e. it starts with identifying single terms and then uses parsing to explore relationships at higher levels of abstraction. The pattern-based approach is limited, because some patterns are not transferable to different sentence structures. Because of this, the system suffers from low recall even though it shows high precision when evaluated on a test data set.

In both cases, named entity recognition represents the critical processing step. It was noted that named entity recognition was a particular challenge for the general class of bioactivity, due to the very broad scope of potential target entities.

4.5 Chapter summary

The chapter has presented the relation extraction results of the two methods of bioactivity event extraction. The first method, based on SVM classifiers, was focused on finding bioactivity relationships between drugs and bacteria. It has been shown that the performance of the SVM classifier can be improved by adding domain-specific features to bag-of-word features. The classifier was used with simple rules for triple extraction, however the system exhibited low recall due to the inability of the patterns to accurately cope with the variety of bioactivity representations in natural language.

The second method aimed to be more general, and extract various types of small molecule bioactivity events using a rule-based extraction system. The result has implied that the method is a high precision approach, however it is still difficult to prove the coverage of the rules defined. The approach has only two rules and a limited number of patterns, which has led to a low recall system. Therefore, a training corpus is needed in order to extend the system.

Chapter 5

Conclusion

Online resources have the potential to significantly change the nature of scientific research. However at present some of the key resources in the biomedical field are isolated. It is the motivation of this work that an extended version of the ChEBI ontology would enable connection with other biological ontologies, such as ontologies of genes and proteins, which it is believed could facilitate exciting advances in biomedical research. There are several reasons for this belief. Firstly, the extended version of the ChEBI ontology provides an organised view of ChEBI data resources by introducing new biological role assignments to each ChEBI entity. Secondly, chemical entities can be well characterised over additional biological dimensions in relation with other biological types. Thirdly, the newer version of the ChEBI ontology can enhance semantic reasoning over different biological resources.

This thesis has taken the view that the most appropriate way of connecting ChEBI with other ontologies is via the bioactivity of small molecules. In order to identify bioactivity in text, two different text mining approaches have been developed.

In this final chapter, a summary will be presented, the contributions of this research will be reviewed, and there will be discussion of directions for future research.

5.1 Summary of the research

Chapter 1 introduced the motivation behind connecting ChEBI with other biomedical informatics resources, and explained why text mining has been selected as an appropriate method for doing so. An overview of biomedical text mining was provided together with an overview of the ongoing work in the biomedical text mining field. The importance of text mining in the biomedical domain was emphasised, along with the positive influences which have resulted from the technique. Secondly, the current state of biomedical text min-

Table 5.1 Review of language patterns used to extract drug-bacterium relationships

Keyword	Example patterns
Resistant/susceptible/ sensitive/tolerant	<NP1><KEYWORD><NP2> <NP1><be><KEYWORD><preposition><NP2>
Resistance/susceptibility/ sensitivity/tolerance	<KEYWORD><preposition><NP1><preposition><NP2>
Against	<NP1><KEYWORD><NP2>
Other	<NP1><verb><NP2>

ing technology has been explained, including the general tasks which are involved as well as the applications and technologies that have been developed to support biomedical text mining. The chapter also provided the background of the project, highlighting the aim of improving the utility of the current ChEBI ontology. The limitations and drawbacks of the current version of the ChEBI ontology were discussed and analysed. Then the prospective contributions of the project were discussed, in addition to proposals on how to construct a future ChEBI ontology based on evidence extracted from the scientific literature.

Two alternative approaches have been developed to achieve this goal, as presented in Chapter 2 and Chapter 3.

The approach of Chapter 2 was based on the use of cascaded machine learning classifiers. This work had a specific focus on a single class of bioactivity, that between drugs and bacteria. The chapter described the motivation for relation extraction between these two types of entities. The focus on a specific case of bioactivity facilitates analysis of vagueness in the way bioactivity is presented. Partly this made possible through development of a corpus. The corpus is used to train machine learning classifiers. These were support vector machines which were used to perform binary classification to find documents and then sentences likely to contain descriptions of bioactivity.

The new corpus created for this work is called DrugPro. The DrugPro corpus contains 400 manually annotated abstracts, supporting the task of named entity recognition of drugs and prokaryotes, bioactivity relation extraction, and development of tools for tasks such as NER.

The chapter moved on to present the relation extraction system that was developed around the classifiers. Once sentences have been found a simple pattern matching approach is used to extract the bioactivity information. The derived patterns work for a range of bioactivity types, as shown in Table 5.1.

Chapter 3 discussed an alternative system that looked at small molecule bioactivity in general, against a whole range of target entities. This system used only rule-based pro-

cessing, without the machine learning classifiers (which were not an option due to the lack of a suitable training corpus). The rule-based approach outlines the bioactivity event extraction system based on two different rule sets, designed for text with different syntactic structures. The key differentiator between the rule sets was whether the bioactivity event was represented as a noun phrase or a verb phrase. Analysis of MEDLINE revealed the distribution of target types for small molecule bioactivity. It was shown how this makes NER particularly difficult.

The rule-based system first detects the event anchor terms and discriminates them according to their part-of-speech categories. Keyword patterns of bioactivity description in the literature were derived, as summarised in Table 5.2.

Table 5.2 Review of language patterns used for general bioactivity event extraction

ID	Pattern	Example
NP1	<SM><TARGET><TRIGGER>	<i>Aspirin is a COX-1 inhibitor</i>
NP2	<SM>,<TARGET><TRIGGER>	<i>Aspirin, as a COX-1 inhibitor</i>
VP1	<SM><TRIGGER><TARGET>	<i>Aspirin inhibits COX-1</i>
VP2	<TARGET><be><TRIGGER><preposition><SM>	<i>COX-1 is inhibited by Aspirin</i>

The results from both approaches were reported in Chapter 4, along with discussion and interpretation. Firstly the results returned from the machine learning system were evaluated. It was shown how the selection of features presented to the SVM algorithm determines the quality of the results. It was found that feature tuning (by using statistical thresholds to limit the quantity of features) made an improvement. The importance of feature selection was also shown. It was found that concatenating some simple domain-specific features to the bag-of-words vector, such as the frequency of drug/bacterium names in a particular paper, also marginally improved the classification results.

The general rule-based extraction approach was manually evaluated using a small data set to allow for manual inspection of the results, and performance evaluation including analysis of errors. This also enabled provisional calculation of precision and recall although it is noted that the data set is very small. The patterns were subsequently used in conjunction with other filters to match and retrieve relevant information from MEDLINE. Of course there is no bioactivity reference data for the whole MEDLINE which can be used for performance evaluation.

There was some comparison of the two approaches. Generally it can be said that the rule-based system suffers from poorer recall, because of the difficulty in constructing patterns that can represent the huge variability present in natural language descriptions of biological phenomena.

5.2 Contributions

The following are the main research contributions of this study. Some minor contributions are omitted. The contribution are listed in order of descending significance.

- The concept of small molecule bioactivity events is a new idea which takes into account a list of specific event types and a wide range of target types. This information is used to create a new annotation schema, which provides a new way to connect ChEBI with other resources, and potentially facilitating knowledge discovery and new scientific research. This could also help with future work on the creation of resources for the text mining community.
- A rule-based system was designed to extract general small molecule bioactivity events from the biomedical literature. The developed system has delivered promising results against the small manually annotated data set available for evaluation purposes. If these results were reproduced against the whole MEDLINE the system would have good precision, although recall appears still low. This has confirmed the feasibility of using this approach to link ChEBI with resources from other biological domains, supporting the original motivation for this project.
- A training corpus which supports drug-bacterium relation extraction has been developed. In order to develop the corpus it was necessary to create annotation guidelines, which instruct practitioners in the process of manually annotating drug and bacterium entities and relationships. This contributes to three possible text mining tasks: NER of drug names, NER of bacterium names, and relation extraction of drug-bacterium associations.
- A classification system using cascaded SVMs, trained using the corpus mentioned above, was presented. The results show that a small but potentially valuable increase in system performance can be seen through adding domain-specific features (including dependency features) to bag-of-word features. The training scheme can be further explored and extended to other relation type learning and extraction, and possibly to other domains.

5.3 Future research

The planned development of the ChEBI ontology has provided a natural guide for future research. Typically, the new reasoning mechanism would bring more biological character-

istics into a small molecule record. The developed systems can be further investigated to adapt more feature inputs to make it scalable.

Future research can be outlined as the following points.

5.3.1 Implementing changes to ChEBI

This thesis aims to contribute to extending the ChEBI ontology, which has led the research in a direction to investigate the limitations and problems which can happen when using the old version of the ontology. One of the objectives for future research can be to implement the changes to the ChEBI ontology, by expanding the *has_role* and *has_target* links and connecting to external resources such as UniProt. It is believed that the new ontology can lead to a better understanding of small molecule characteristics. This research has found evidence from the literature supporting an extension branch of the ontology. These benefits may also hold for the more general population of user or ontology user with direct interests in investigating more interesting traits of small molecule and finding a new direction for biological research. The semantic relationships proposed in this work provides the fundamental backbone supporting the creation of a ChEBI sub-ontology, however finding new semantic relations can require heuristic methods.

5.3.2 Development of a general corpus for small molecule bioactivity

In Chapter 4, the limitation of the rule-based extraction system has been discussed. Ultimately the system, like most rule-based systems, will sacrifice recall for precision. Elsewhere it has been argued that precision (i.e. accuracy of results) is more important than recall for biomedical text mining [49]. However this is likely to depend on the particular application.

The discussion has identified the potential to further tune the extraction system. Further improvements to the rules could be made, but realisable rule-based approaches will probably always represent a trade-off between precision and recall.

It is believed that when extracting or tagging natural-language text, machine learning approaches will perform substantially better than rule-based approaches. To some extent this can be verified by looking at the performance of systems entered into community shared tasks, such as BioNLP and ConLL (noting that these are very specific tasks). Rule-based systems are generally heuristic, however, because the features to be used have to be identified by the machine learning algorithms. Machine learning approaches are better equipped to deal with the variability present in natural language. The reference training corpus is

therefore crucial, since it is a pre-requisite for supervised machine learning and determines the abilities of the approach being employed.

Ultimately the rule-based approach was adopted due to the substantial overhead of developing a suitable training corpus for general small molecule bioactivity. Whilst it was feasible to create a corpus for the constrained field of drug-prokaryote relationships, expanding this to small molecule bioactivity in general (and expecting to have good coverage) would be a huge challenge. Unfortunately corpus creation is a manual process which relies on the skill, knowledge, and diligence of human annotators. Creating a large, generalised, corpus would take a huge amount of effort, and consequently represents a large financial cost. Therefore perhaps the pragmatic approach is to build on work already completed. The GENIA corpus is made up of 2,000 MEDLINE abstracts already annotated with many aspects of interest to small molecular bioactivity applications. Extension of GENIA probably represents the best starting point for a future small molecule bioactivity corpus.

5.3.3 Solving the NER problem for small molecule bioactivity

NER is arguably the most critical process in text mining. Reasons for its complexity in the biomedical field were described in Chapter 1. In Chapter 3 we saw how it was necessary to combine NER tools to provide coverage of the various entities involved in small molecule bioactivity. We also saw how the order in which NER tools are applied can have a major effect. It is clear that a single, unified approach to NER for small molecule bioactivity would be a significant step forward.

There are multiple approaches to NER. An excellent summary is provided in [15]. Dictionary-based approaches, of the type used here, can be limited in recall. Hence the application of machine learning to the NER task is a natural progression. Once again this highlights the need for a suitable training corpus with annotation of all the relevant target entities. Ultimately it may be possible to train machine learning classifiers without an annotated corpus. For now, hybrid approaches combining the best of dictionaries, hand-crafted rules, and machine learning, offer possibly the greatest potential.

5.3.4 Automatic pattern generation

The current version of the extraction system only aims for high precision, and subsequently does not achieve high recall. The pattern matching component restricts the recall of the general bioactivity extraction system. The problem is caused by the patterns being overfitted to a particular language form. In order to see improvements in recall more patterns would

be required, but this is difficult and tedious, and ultimately infeasible.

Consequently, further development of the system is needed. To avoid manually creating patterns, several research projects have investigated the feasibility of automated pattern learning (e. g. [107]). Patterns are loosely expressed in regular-expression style language, and are then generated based on a previously annotated training corpus. Riloff [38] developed AutoSlog, a system for automatically constructing a dictionary (essentially lexical-semantic patterns) for information extraction that achieved 90% of the performance of a hand-crafted alternative, but was developed in a fraction of the time. Similarly, Califf & Mooney [108] designed a bottom-up learning algorithm that uses techniques from inductive logic programming. Patterns can be purely lexical, based on named entities, or use the output of dependency parsing. Perhaps the most interesting opportunities lie in simultaneous exploitation of all of these features.

The ability to automatically learn patterns was briefly investigated during the course of this research, and resulted in a conference publication [58]. The approach was called bootstrapping, because it started with a small number of high precision ‘seed’ pairs and then used machine learning to discover new patterns. The work is immature but early results were promising. The work was inspired by the Snowball system [109].

Such a pattern learning approach can also be developed further in order to provide improved detection of negation and speculation, two particularly challenging problems for pattern development.

5.4 Final remarks

In all, the work has shown that the idea of using bioactivity to extend the current ChEBI ontology is feasible, and the resulting resource would contribute to the ontology community as well as the biomedical research domain. The potential benefits encourage research to characterise small molecules using a new set of criteria.

The very broad scope of small molecule bioactivity provides a significant challenge, potentially making this more demanding than many other biomedical text mining projects. When the full scope of small molecule bioactivity is considered, then critical processes such as named-entity recognition become extremely difficult, compounded by the lack of a suitable corpus. When the scope is constrained to specific instances of bioactivity (such as between drugs and bacteria), it becomes possible to create the resources required for supervised machine learning techniques. In comparison to basic rule-based approaches, machine learning is more able to handle the unpredictability and vagueness in natural language text,

with a corresponding improvement in results.

References

- [1] P. Lambrix, H. Tan, V. Jakoniene, and L. Strömbäck, “Biological ontologies,” in *Semantic Web*, pp. 85–99, Springer, 2007.
- [2] J. B. Bard and S. Y. Rhee, “Ontologies in biology: design, applications and future challenges,” *Nature Reviews Genetics*, vol. 5, no. 3, pp. 213–222, 2004.
- [3] P. de Matos, R. Alcantara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, and C. Steinbeck, “Chemical entities of biological interest: an update,” *Nucl. Acids Res.*, vol. 38, pp. D249–254, January 2010.
- [4] A. Guaadaoui, S. Benaicha, N. Elmajdoub, M. Bellaoui, and A. Hamal, “What is a bioactive compound? A combined definition for a preliminary consensus,” *International Journal of Nutrition and Food Sciences*, vol. 3, no. 3, pp. 174–179, 2014.
- [5] Y. Yan, J. Hastings, J.-H. Kim, S. Schulz, C. Steinbeck, and D. Rebholz-Schuhmann, “Use of multiple ontologies to characterize the bioactivity of small molecules,” in *ICBO*, 2011.
- [6] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, *et al.*, “ChEMBL: a large-scale bioactivity database for drug discovery,” *Nucleic acids research*, vol. 40, no. D1, pp. D1100–D1107, 2012.
- [7] M. Horridge, N. Drummond, J. Goodwin, A. L. Rector, R. Stevens, and H. Wang, “The Manchester OWL syntax,” in *OWLed*, vol. 216, 2006.
- [8] UniProt Consortium, “The universal protein resource (UniProt),” *Nucleic acids research*, vol. 36, no. suppl 1, pp. D190–D195, 2008.
- [9] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, “Gene ontology: tool for the unification of biology,” *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [10] D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp, “Database resources of the National Center for Biotechnology Information,” *Nucleic acids research*, vol. 28, pp. 10–14, Jan. 2000.
- [11] L. Dey, M. Abulaish, G. Sharma, *et al.*, “Text mining through entity-relationship based information extraction,” in *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops*, pp. 177–180, IEEE Computer Society, 2007.

- [12] J. Jiang and C. Zhai, "A systematic exploration of the feature space for relation extraction.," in *HLT-NAACL*, pp. 113–120, 2007.
- [13] M. Palakal, M. Stephens, S. Mukhopadhyay, R. Raje, and S. Rhodes, "A multi-level text mining method to extract biological relationships," in *Bioinformatics Conference, 2002. Proceedings. IEEE Computer Society*, pp. 97–108, IEEE, 2002.
- [14] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," *Database*, vol. 2011, 2011.
- [15] U. Leser and J. Hakenberg, "What makes a gene name? Named entity recognition in the biomedical literature," *Briefings in Bioinformatics*, vol. 6, no. 4, pp. 357–369, 2005.
- [16] A. Manconi, E. Vargiu, G. Armano, and L. Milanesi, "Literature retrieval and mining in bioinformatics: state of the art and challenges," *Advances in bioinformatics*, vol. 2012, 2012.
- [17] Y. Levy and T. J. Ellis, "A systems approach to conduct an effective literature review in support of information systems research," *Informing Science: International Journal of an Emerging Transdiscipline*, vol. 9, pp. 181–212, 2006.
- [18] I. Moszer, L. M. Jones, S. Moreira, C. Fabry, and A. Danchin, "SubtiList: the reference database for the *Bacillus subtilis* genome," *Nucleic acids research*, vol. 30, no. 1, pp. 62–65, 2002.
- [19] S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein, L. Ungar, S. Winters, and P. White, "Integrated annotation for biomedical information extraction," in *Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pp. 61–68, 2004.
- [20] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus—a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. suppl 1, pp. i180–i182, 2003.
- [21] H. Chen, *Medical informatics: knowledge management and data mining in biomedicine*, vol. 8. Springer, 2005.
- [22] R. Rodriguez-Esteban, "Biomedical text mining and its applications," *PLoS computational biology*, vol. 5, no. 12, p. e1000597, 2009.
- [23] R. A. Erhardt, R. Schneider, and C. Blaschke, "Status of text-mining techniques applied to biomedical text," *Drug discovery today*, vol. 11, no. 7, pp. 315–325, 2006.
- [24] R. J. Mooney and R. Bunescu, "Mining knowledge from text using information extraction," *ACM SIGKDD explorations newsletter*, vol. 7, no. 1, pp. 3–10, 2005.
- [25] R. Rodriguez-Esteban, "Biomedical text mining and its applications," *PLoS computational biology*, vol. 5, no. 12, p. e1000597, 2009.
- [26] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno, "Text processing through web services: calling Whatizit," *Bioinformatics*, vol. 24, no. 2, pp. 296–298, 2008.

- [27] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, 2005.
- [28] R. Leaman, G. Gonzalez, *et al.*, "BANNER: an executable survey of advances in biomedical named entity recognition.," in *Pacific Symposium on Biocomputing*, vol. 13, pp. 652–663, 2008.
- [29] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. Hendriksen, B. J. Schijvenaars, E. M. Van Mulligen, J. Kleinjans, and J. A. Kors, "A dictionary to identify small molecules and drugs in free text," *Bioinformatics*, vol. 25, no. 22, pp. 2983–2991, 2009.
- [30] D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy, and P. Murray-Rust, "OS-CAR4: a flexible architecture for chemical text-mining," *Journal of cheminformatics*, vol. 3, no. 1, pp. 1–12, 2011.
- [31] K. Sagae, "Gdep (GENIA dependency parser)." <http://people.ict.usc.edu/~sagae/parser/gdep/>. [Online; accessed 7-February-2015].
- [32] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski, "BioInfer: a corpus for information extraction in the biomedical domain," *BMC Bioinformatics*, vol. 8, no. 1, p. 50, 2007.
- [33] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong, "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial intelligence in medicine*, vol. 33, no. 2, pp. 139–155, 2005.
- [34] F. Guenther, "Electronic lexica and corpora research at CIS," *International Journal of Corpus Linguistics*, vol. 1, no. 2, pp. 287–301, 1996.
- [35] K. Fundel, R. Küffner, and R. Zimmer, "RelEx–relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.
- [36] M. Miwa, S. Pyysalo, T. Hara, and J. Tsujii, "A comparative study of syntactic parsers for event extraction," in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, BioNLP '10*, (Stroudsburg, PA, USA), pp. 37–45, Association for Computational Linguistics, 2010.
- [37] J. Cowie and W. Lehnert, "Information extraction," *Communications of the ACM*, vol. 39, no. 1, pp. 80–91, 1996.
- [38] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93)*, pp. 811–816, AAAI Press/The MIT Press, 1993.
- [39] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D684–D688, 2008.

- [40] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D198–D201, 2007.
- [41] T. Klein, J. Chang, M. Cho, K. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D. Oliver, *et al.*, "Integrating genotype and phenotype information: an overview of the PharmGKB project," *Pharmacogenomics J*, vol. 1, no. 3, pp. 167–170, 2001.
- [42] A. P. Davis, B. L. King, S. Mockus, C. G. Murphy, C. Saraceni-Richards, M. Rosenstein, T. Wieggers, and C. J. Mattingly, "The comparative toxicogenomics database: update 2011," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D1067–D1072, 2011.
- [43] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen, and P. Bork, "STITCH 3: zooming in on protein–chemical interactions," *Nucleic acids research*, vol. 40, no. D1, pp. D876–D880, 2012.
- [44] S. Ahmad and E. Patela, "SuperTarget: Applications software for oil refinery retrofit," tech. rep., American Institute of Chemical Engineers, New York, NY, 1987.
- [45] N. Hecker, J. Ahmed, J. von Eichborn, M. Dunkel, K. Macha, A. Eckert, M. K. Gilson, P. E. Bourne, and R. Preissner, "SuperTarget goes quantitative: update on drug–target interactions," *Nucleic acids research*, vol. 40, no. D1, pp. D1113–D1117, 2012.
- [46] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.
- [47] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D277–D280, 2004.
- [48] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in KEGG," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D354–D357, 2006.
- [49] J. Lee, S. Kim, S. Lee, K. Lee, and J. Kang, "High precision rule based PPI extraction and per-pair basis performance evaluation," in *Proceedings of the ACM sixth international workshop on data and text mining in biomedical informatics*, pp. 69–76, ACM, 2012.
- [50] M. He, Y. Wang, and W. Li, "PPI Finder: a mining tool for human protein-protein interactions," *PloS one*, vol. 4, no. 2, p. e4554, 2009.
- [51] T. Fayruzov, G. Dittmar, N. Spence, M. De Cock, and A. Teredesai, "A rapidminer framework for protein interaction extraction," in *RapidMiner community meeting*, 2010.
- [52] A. Munger and A. Teredesai, "Relation extraction from biomedical text using SVMs with dependency tree kernels," *Bioinformatics*, 2010.

- [53] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09 shared task on event extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 1–9, Association for Computational Linguistics, 2009.
- [54] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09 shared task on event extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 1–9, Association for Computational Linguistics, 2009.
- [55] S. Pyysalo, T. Ohta, R. Rak, D. Sullivan, C. Mao, C. Wang, B. Sobral, J. Tsujii, and S. Ananiadou, "Overview of the infectious diseases (ID) task of BioNLP shared task 2011," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 26–35, Association for Computational Linguistics, 2011.
- [56] S. Riedel, D. McClosky, M. Surdeanu, A. McCallum, and C. D. Manning, "Model combination for event extraction in BioNLP 2011," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 51–55, Association for Computational Linguistics, 2011.
- [57] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 28–34, Association for Computational Linguistics, 2011.
- [58] Y. Yan, J.-H. Kim, S. Croset, and D. Rebholz-Schuhmann, "Finding small molecule and protein pairs in scientific literature using a bootstrapping method," in *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp. 172–175, Association for Computational Linguistics, 2012.
- [59] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," 2007.
- [60] S. Zhao and R. Grishman, "Extracting relations with integrated information using kernel methods," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 419–426, Association for Computational Linguistics, 2005.
- [61] Z. GuoDong, S. Jian, Z. Jie, and Z. Min, "Exploring various knowledge in relation extraction," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 427–434, Association for Computational Linguistics, 2005.
- [62] C. Boulis and M. Ostendorf, "Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams," in *Proc. of the International Workshop in Feature Selection in Data Mining*, pp. 9–16, 2005.
- [63] T. Joachims, *Learning to classify text using support vector machines: methods, theory and algorithms*. Kluwer Academic Publishers, 2002.
- [64] S. Kim, J. Yoon, and J. Yang, "Kernel approaches for genic interaction extraction," *Bioinformatics*, vol. 24, no. 1, pp. 118–126, 2008.

- [65] R. McDonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, and P. White, “Simple algorithms for complex relation extraction with applications to biomedical IE,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 491–498, Association for Computational Linguistics, 2005.
- [66] Ş. Kafkas, E. Varoğlu, D. Rebholz-Schuhmann, and B. Taneri, “Functional variation of alternative splice forms in their protein interaction networks: a literature mining approach,” *BMC Bioinformatics*, vol. 11, no. Suppl 5, p. P1, 2010.
- [67] T. Mitsumori, M. Murata, Y. Fukuda, D. Kouichi, and D. Hirohumi, “Extracting protein-protein interaction information from biomedical text with SVM,” *IEICE Transactions on Information and Systems*, vol. 89, no. 8, pp. 2464–2466, 2006.
- [68] R. Sætre, K. Sagae, and J. Tsujii, “Syntactic features for protein-protein interaction extraction,” in *LBM (Short Papers)*, Citeseer, 2007.
- [69] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, “Support vector machine learning for interdependent and structured output spaces,” in *Proceedings of the twenty-first international conference on Machine learning*, p. 104, ACM, 2004.
- [70] Y. Yan, S. Kafkas, M. Conroy, and D. Rebholz-Schuhmann, “Towards generating a corpus annotated for prokaryote-drug relations,” in *SMBM 2012 - Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine*, pp. 100–101, 2012.
- [71] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, “Drugbank: a comprehensive resource for in silico drug discovery and exploration,” *Nucleic acids research*, vol. 34, no. suppl 1, pp. D668–D672, 2006.
- [72] N. Fierer, J. A. Jackson, R. Vilgalys, and R. B. Jackson, “Assessment of soil microbial community structure by use of taxon-specific quantitative pcr assays,” *Applied and Environmental Microbiology*, vol. 71, no. 7, pp. 4117–4120, 2005.
- [73] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, “Text mining techniques for patent analysis,” *Information Processing & Management*, vol. 43, no. 5, pp. 1216–1247, 2007.
- [74] E. Gabrilovich and S. Markovitch, “Feature generation for text categorization using world knowledge,” in *IJCAI*, vol. 5, pp. 1048–1053, 2005.
- [75] P. Kolari, T. Finin, and A. Joshi, “SVMs for the blogosphere: Blog identification and splog detection,” in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 92–99, 2006.
- [76] Y. Yang, “An evaluation of statistical approaches to text categorization,” *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.
- [77] P. Soucy and G. W. Mineau, “Beyond TFIDF weighting for text categorization in the vector space model,” in *IJCAI*, vol. 5, pp. 1130–1135, 2005.
- [78] N. Indurkha and T. Zhang, *Text mining: predictive methods for analyzing unstructured information*. Springer, 2005.

- [79] L. M. Manevitz and M. Yousef, “One-class SVMs for document classification,” *The Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2002.
- [80] E.-H. S. Han and G. Karypis, *Centroid-based document classification: Analysis and experimental results*. Springer, 2000.
- [81] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.
- [82] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, “Mismatch string kernels for discriminative protein classification,” *Bioinformatics*, vol. 20, no. 4, pp. 467–476, 2004.
- [83] L.-P. Jing, H.-K. Huang, and H.-B. Shi, “Improved feature selection approach TFIDF in text mining,” in *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on*, vol. 2, pp. 944–946, IEEE, 2002.
- [84] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, *et al.*, “Overview of BioCreative II gene normalization,” *Genome biology*, vol. 9, no. Suppl 2, p. S3, 2008.
- [85] T. Joachims, *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer/Springer, 2002.
- [86] A. Moschitti, “A study on convolution kernels for shallow semantic parsing,” in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 335, Association for Computational Linguistics, 2004.
- [87] S. Pyysalo, F. Ginter, K. Haverinen, J. Heimonen, T. Salakoski, and V. Laippala, “On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BioInfer and GENIA,” in *Proceedings of the Workshop on BioNLP 2007: Biological, translational, and clinical language processing*, pp. 25–32, Association for Computational Linguistics, 2007.
- [88] G. Schneider, F. Rinaldi, K. Kaljurand, and M. Hess, “Steps towards a GENIA dependency treebank,” in *Third Workshop on Treebanks and Linguistic Theories (TLT) 2004*, pp. 137–149, 2004.
- [89] R. C. Bunescu and R. J. Mooney, “A shortest path dependency kernel for relation extraction,” in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 724–731, Association for Computational Linguistics, 2005.
- [90] X. Zhou, M.-C. J. Kao, and W. H. Wong, “Transitive functional annotation by shortest-path analysis of gene expression data,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12783–12788, 2002.
- [91] R. Girju and D. I. Moldovan, “Text mining for causal relations,” in *FLAIRS Conference*, pp. 360–364, 2002.

- [92] R. Feldman, Y. Regev, M. Finkelstein-Landau, E. Hurvitz, and B. Kogan, "Mining biomedical literature using information extraction," *Current Drug Discovery*, vol. 2, no. 10, pp. 19–23, 2002.
- [93] R. D. Van Valin and R. J. LaPolla, *Syntax: Structure, meaning, and function*. Cambridge University Press, 1997.
- [94] P. Skehan, *A cognitive approach to language learning*. Oxford University Press, 1998.
- [95] A. B. Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach," *Journal of biomedical semantics*, vol. 2, no. Suppl 5, p. S4, 2011.
- [96] H. Kilicoglu and S. Bergler, "Syntactic dependency based heuristics for biological event extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 119–127, Association for Computational Linguistics, 2009.
- [97] Q.-C. Bui, E. M. van Mulligen, D. Campos, and J. A. Kors, "A fast rule-based approach for biomedical event extraction," *ACL 2013*, p. 104, 2013.
- [98] Q.-C. Bui and P. Sloot, "Extracting biological events from text using simple syntactic patterns," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 143–146, Association for Computational Linguistics, 2011.
- [99] J. R. Hobbs, "Information extraction from biomedical text," *Journal of Biomedical Informatics*, vol. 35, no. 4, pp. 260 – 264, 2002. Sublanguage - Zellig Harris Memorial.
- [100] E. Ritter and M. Wiltchko, "Anchoring events to utterances without tense," in *Proceedings of WCCFL*, vol. 24, pp. 343–351, 2005.
- [101] S. Ananiadou, S. Pyysalo, J. Tsujii, and D. B. Kell, "Event extraction for systems biology by text mining the literature," *Trends in biotechnology*, vol. 28, no. 7, pp. 381–390, 2010.
- [102] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucl. Acids Res.*, vol. 32, pp. D267–270, Jan. 2004.
- [103] C. Kolárik, R. Klinger, C. M. Friedrich, M. Hofmann-Apitius, and J. Fluck, "Chemical names: terminological resources and corpora annotation," in *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*, vol. 36, 2008.
- [104] E. Buyko and U. Hahn, "Evaluating the impact of alternative dependency graph encodings on solving event extraction tasks," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 982–992, Association for Computational Linguistics, 2010.

- [105] J. R. Hobbs, D. E. Appelt, J. Bear, D. J. Israel, M. Kameyama, M. E. Stickel, and M. Tyson, "FASTUS: A cascaded finite-state transducer for extracting information from natural-language text," *CoRR*, vol. cmp-lg/9705013, 1997.
- [106] M. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, *et al.*, "The gene ontology (GO) database and informatics resource.," *Nucleic acids research*, vol. 32, no. Database issue, pp. D258–61, 2004.
- [107] Q.-C. Bui and P. M. Sloot, "A robust approach to extract biomedical events from literature," *Bioinformatics*, vol. 28, no. 20, pp. 2654–2661, 2012.
- [108] M. E. Califf and R. J. Mooney, "Bottom-up relational learning of pattern matching rules for information extraction," *The Journal of Machine Learning Research*, vol. 4, pp. 177–210, 2003.
- [109] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proceedings of the fifth ACM conference on Digital libraries*, pp. 85–94, ACM, 2000.

