In: Ananiadou S, McNaught J, editors. Text mining for biology and biomedicine: Artech House; 2006. p. 43-66.

Chapter 3 Lexical, terminological and ontological resources for biological text mining

Olivier Bodenreider
U.S. National Library of Medicine, 8600 Rockville Pike, MS 43, Bethesda, Maryland 20894, USA

3.1 Introduction

Biomedical terminologies and ontologies are frequently described as enabling resources in text mining systems [e.g., 1, 2, 3]. These resources are used to supports tasks such as entity recognition (i.e., the identification of biomedical entities in text) and relation extraction (i.e., the identification of relationships among biomedical entities). Although a significant part of current text mining efforts focuses on the analysis of documents related to molecular biology, the use of lexical, terminological and ontological resources is mentioned in research systems developed for the analysis of clinical narratives (e.g., MedSyndikate [4]) or the biological literature (e.g., BioRAT [5], GeneScene [6], EMPathIE [7] and PASTA [7]). Of note, some systems initially developed for extracting clinical information have later been adapted to extract relations among biological entities (e.g., MedLEE [8] / GENIES [9], SemRep / SemGen [10]). Commercial systems such as TeSSI[i] also make use of such resources.

Entity recognition often draws on lists of entity names collected in lexicons, gazetteers and, more generally, terminology resources. Lists of disease names, for example, can be easily extracted from disease resources such as the International Classification of Diseases (ICD), from the disease component of general resources such as the Medical Subject Headings (MeSH) and from specialized resources such as the Online Multiple Congenital Anomaly / Mental Retardation (MCA/MR) Syndromes©. Relation extraction, on the other hand, may benefit from the relationships represented among terms in terminologies (e.g., *Parkinson's disease* ***child of*** *Neurodegenerative diseases* in MeSH) and in ontologies (e.g., *Basal ganglia* ***finding site of*** *Parkinson's disease* in SNOMED CT).

Biomedical *lexicons* such as the UMLS SPECIALIST lexicon collect lexical items (words and multi-word expressions) frequently observed in biomedical text corpora and record information about them, including part of speech (e.g., noun, adjective), inflectional variants (e.g., singular/plural), spelling variants (e.g., American vs. British English). This information is useful not only to natural language processing (NLP) tools such as part-of-speech taggers and parsers, but

also to entity recognition systems as it can help identify variants of entity names in text [11].
The purpose of biomedical *terminology* is to collect the names of entities employed in the biomedical domain. Most biomedical terminologies record synonymous terms (e.g., *Parkinson's disease* and *Paralysis agitans*) and have some kind of hierarchical organization, often tree or graph-like [12]. Terminology-driven approaches to text mining have been explored in [13].
In contrast, biomedical *ontology* aims to study not names, but kinds of entities (i.e., substances, qualities and processes) of biomedical significance and the relations among them. Examples of such entities include substances such as the mitral valve and glucose, qualities such as the diameter of the left ventricle and the catalytic function of enzymes, and processes such as blood circulation and secreting hormones. Fundamental relations in biomedical ontologies include not only ***is a*** and ***part of***, but also ***instance of***, ***adjacent to***, ***derives from***, etc. [14].
In practice, the distinction between lexicons, terminologies and ontologies is not always sharp. On the one hand, although ontologies mostly focus on relations among entities, some of them also record the names by which entities are referred to. On the other, although terminologies essentially collect the names of entities, their hierarchical organization also reflects relations among such entities. Finally, the very names of these resources can be misleading. For example, despite its name, the Gene Ontology (GO) defines itself as a controlled vocabulary (i.e., a terminological resource), but like ontologies, its terms are linked by relationships such as ***is a*** and ***part of***. However, the definition and use of such relations is not consistent throughout GO [15], as would be expected from ontologies.

The objective of this chapter is to present some of the resources (lexicons, terminologies and ontologies) of interest for entity recognition and relation extraction tasks. Providing an exhaustive list of these resources is beyond the scope of this paper. Moreover, many of these resources are highly specialized and would therefore be of little interest to most readers. Instead, we selected general, publicly available resources that have been shown to be useful for biomedical text mining. Furthermore, this review is purposely limited to resources in English.
We start by presenting an extended example illustrating biomedical terms in two pieces of text. We then give a brief description of the major resources available, with a particular emphasis on the Unified Medical Language System® (UMLS®) [16]. Finally, we discuss some issues related to biomedical terms and biomedical relations. The reader is referred to chapters 4 and 6 for a detailed presentation of the tasks of entity recognition and relation extraction.

3.2 Extended example

In this example, we consider two short pieces of text related to the genetic disease neurofibromatosis 2. *Neurofibromatosis 2* is an autosomal dominant disease characterized by tumors called *schwannomas* involving the acoustic nerve, as well as other features [17]. The disorder is caused by mutations of the *NF2* gene

resulting in absence or inactivation of the protein product. The protein product of *NF2* is commonly called *merlin* (but also *neurofibromin 2* and *schwannomin*) and functions as a tumor suppressor. The first fragment of text (1) is extracted from the abstract of an article[ii]. The second is the definition of neurofibromatosis 2 in the Medical Subject Headings (MeSH) vocabulary[iii].

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22. (1)

Neurofibromatosis type 2: An autosomal dominant disorder characterized by a high incidence of bilateral acoustic neuromas as well as schwannomas of other cranial and peripheral nerves, and other benign intracranial tumors including meningiomas, ependymomas, spinal neurofibromas, and gliomas. The disease has been linked to mutations of the NF2 gene on chromosome 22 (22q12) and usually presents clinically in the first or second decade of life. (2)

3.2.1Entity recognition

Many biomedical entities can be identified in these two fragments. Underlined expressions correspond to terms present in the UMLS Metathesaurus. This is the case, for example, of the disease *neurofibromatosis 2* and the protein *merlin*. Interestingly, *vestibular schwannomas* in (1) and *acoustic neuromas* in (2), although lexically distinct, name the same tumor. While a lexicon is useful to identify these disease names, a terminology (or ontology) is required to identify them as synonymous. These two terms are names for the same disease concept in the UMLS Metathesaurus (C0027859). The list of UMLS concepts that can be identified in the two text fragments is given in Table 3.1.

Table 3.1 – UMLS concepts (identifier [CUI], preferred name and semantic types [see Table 3.4 for the full names]) identifiable in text fragments (1) and (2). Column 'M' indicates the type of match (s: single simple match, m: multiple simple matches, a: approximate match, -: no direct match)

Source	String in text	M	CUI	Preferred name	S. Types
(1) (2)	Neurofibromatosis type 2	s	C0027832	Neurofibromatosis 2	neop
(1)	NF2	s	C0085114	Neurofibromatosis 2 genes	gngm
(1)	peripheral neurofibromatosis	s	C0027831	Neurofibromatosis 1	neop
(1)	{intracranial	-	C0007682	Central Nervous	dsyn

	condition}			System Diseases	
(1)	[bilateral] vestibular schwannomas	a	C0027859	Neuroma, Acoustic	neop
(1) (2)	mutation / mutations	s	C0026882	Mutation	genf
(1)	gene	s	C0017337	Genes	gngm
(1)	merlin	m	C0254123	Neurofibromin 2	aapp, bacs
(1) (2)	chromosome 22	s	C0008665	Chromosomes, Human, Pair 22	celc
(2)	autosomal dominant disorder	a	C0265385	Autosomal dominant hereditary disorder	dsyn
(2)	bilateral acoustic neuromas	s	C1136042	Neuroma, Acoustic, Bilateral	neop
(2)	schwannomas	s	C0027809	Neurilemmoma	neop
(2)	cranial and peripheral nerves	-	C0010268 C0031119	• Cranial Nerves • Peripheral Nerves	bpoc bdsy
(2)	[benign] intracranial tumors	a	C0750978	Neoplasms, Intracranial	neop
(2)	meningiomas	s	C0025286	Meningioma	neop
(2)	ependymomas	s	C0014474	Ependymoma	neop
(2)	neurofibromas	s	C0027830	Neurofibroma	neop
(2)	gliomas	s	C0017638	Glioma	neop
(2)	disease	s	C0012634	Disease	dsyn
(2)	NF2 gene	s	C0085114	Neurofibromatosis 2 genes	gngm

Many expressions extracted from the two text fragments can be mapped to the UMLS Metathesaurus through a simple match (i.e., exact match or after normalization). Except for *merlin*, which maps to both a protein and a bird, the mapping is unambiguous. In contrast, expressions in the dotted boxes also correspond to biomedical entities, but the name found in the text cannot be mapped directly to a UMLS concept. Expressions such as *intracranial condition* in (1) are vague compared to the corresponding concept names in the UMLS (e.g., *central nervous system diseases*). Complex phrases such as *cranial and peripheral nerves* in (2) refer to two concepts (i.e., *cranial nerves* and *peripheral nerves*) present in the Metathesaurus. Conversely, some expressions in the text convey more precision than the corresponding concepts found in biomedical terminologies (e.g., *bilateral vestibular schwannomas* in (1) vs. *vestibular schwannomas* and *benign intracranial tumors* in (2) vs. *intracranial tumors*). In these cases, while terminological resources are useful for identifying entities in text, they may not be sufficient for capturing all nuances present in the text. Term variation and management issues are discussed extensively in chapter 2.

3.2.2Relation extraction

Once entities have been identified in text fragments, the next step consists of identifying the relationships among them such as *vestibular schwannomas*

manifestation of *neurofibromatosis 2* and *NF2 gene* ***located on*** *chromosome 22*. Such relations may be explicitly represented in biomedical ontologies. For example the relation *schwannomas* ***associated morphology of*** *neurofibromatosis 2* is asserted in SNOMED CT. However, ontologies do not necessarily contain such fine-grained assertions but may rather represent higher-level facts such as *gene* ***located on*** *chromosome*. A relation extraction system would first identify *NF2 gene* as a kind of gene and *chromosome 22* as a kind of chromosome before inferring that a particular gene (*NF2 gene*) is located on a particular chromosome (*chromosome 22*).

The use of ontologies to support relation extraction often requires the system to identify in the text not only entities, but also potential relationships. Clues for identifying relationships include lexical items (e.g., the preposition *on* for the relationship ***located on***) and syntactic structures (e.g., *intracranial tumors including meningiomas* for *meningiomas* ***is a*** *intracranial tumors*), as well as statistical and pattern based clues (not presented here). Relations may span several sentences and their identification often requires advanced linguistic techniques such as anaphora and co-reference resolution. For example, from the last sentence of (2), the relation *disease* ***associated with*** *mutation* can be extracted. While accurate, this relation is incomplete in this context because *disease* actually refers not to any disease, but to *neurofibromatosis 2* (anaphoric relation). Similarly, *mutations of the NF2 gene* (not mutations in general) is the entity associated with the disease. Therefore, the complete relation to be extracted is *neurofibromatosis 2* ***associated with*** *mutations of the NF2 gene*. The potential relations extracted from the text can then be validated against the relations explicitly represented in the ontology or inferred from it.

3.3 Lexical resources

The resources presented under this category provide the lexical and lexico-syntactic information needed for parsing text. The major resource for biomedical text is the SPECIALIST lexicon. Additionally, specialized resources can be useful for analyzing subdomains of biomedicine (e.g., lists of gene names for molecular biology corpora). Conversely, general resources such as WordNet can also help analyze the literature written for less specialized audiences (e.g., for patients).

3.3.1WordNet

WordNet® is an electronic lexical database developed at Princeton University that serves as a resource for applications in natural language processing and information retrieval [18]. The core structure in WordNet is a set of synonyms (synset) that represents one underlying concept. For example, the synset representing *hemoglobin* also contains the lexical entries *haemoglobin* (British-English spelling) and *Hb* (abbreviation). A definition is provided for the synset: “a hemoprotein composed of globin and heme that gives red blood cells their

characteristic color; function primarily to transport oxygen from the lungs to the body tissues". There are separate structures for each linguistic category covered: nouns, verbs, adjectives, and adverbs. For example, the adjective "renal" and the noun "kidney," although similar in meaning, belong to two distinct structures, and a specific relationship, "pertainymy," relates the two forms. The current version of WordNet (2.0) contains over 114,000 noun synsets. In addition to being a lexical resource, WordNet has some of the features of ontologies. For example, each synset in the noun hierarchy belongs to at least one ***is a*** tree (e.g., *hemoglobin* ***is a*** *protein*) and may additionally belong to several ***part of***-like trees (*hemoglobin* ***substance of*** *red blood cell*). Because of its modest coverage of the biomedical domain [19, 20], WordNet has been used only in a limited number of projects in biomedicine [21] where resources such as the UMLS usually play a more prominent role. WordNet is available free of charge from http://wordnet.princeton.edu/. Application programming interfaces (APIs) have been developed for the major programming languages, making it relatively easy for developers to integrate WordNet in applications.

3.3.2UMLS SPECIALIST lexicon

The SPECIALIST lexicon is one of three knowledge sources developed by the National Library of Medicine (NLM) as part of the Unified Medical Language System (UMLS) project. It provides the lexical information needed for processing natural language in the biomedical domain [22]. The lexicon entry for each word or multi-word term records syntactic (part of speech, allowable complementation patterns), morphological (base form, inflectional variants) and orthographic (spelling variants) information. It is in fact a general English lexicon that includes many biomedical terms. Lexical items are selected from a variety of sources, including lexical items from MEDLINE/PubMed® citation records, the UMLS Metathesaurus and a large set of lexical items from medical and general English dictionaries. Contrary to WordNet, the SPECIALIST lexicon does not include any information about synonymy or semantic relations among its entries. This information, however, is present in the Metathesaurus, another component of the UMLS (see 3.4.3). The record for *hemoglobin* in the SPECIALIST lexicon, shown in Figure 3.1, indicates the base form, one spelling variant, and two inflectional classes as hemoglobin is used as both an mass noun (e.g., in "Hemoglobin concentration is reported as grams of hemoglobin per deciliter of blood.") and a countable (e.g., in "the study of hemoglobins, both normal and mutant, […]"). Additionally, the abbreviation *Hb* and the acronym *Hgb* are cross-referenced to *hemoglobin*. The SPECIALIST lexicon is distributed as part of the UMLS and can be queried through application programming interfaces for Java and XML. It is also available as an open source resource as part of the SPECIALIST NLP tools (http://SPECIALIST.nlm.nih.gov).

```
{
  base=hemoglobin          (base form)
  spelling_variant=haemoglobin
  entry=E0031208           (identifier)
  cat=noun                 (part of speech)
  variants=uncount         (no plural)
  variants=reg             (plural: hemoglobins , haemoglobins)
}
```

Figure 3.1 – Representation of *hemoglobin* in the SPECIALIST lexicon

3.3.3Other specialized resources

While general resources such as WordNet and the SPECIALIST lexicon provide a good coverage of the general biomedical language, they (purposely) fail to cover in detail specialized subdomains such as gene and protein names or chemical and drug names. The syntactic analyzers and parsers relying on these resources may therefore give suboptimal results when analyzing specialized corpora (e.g., molecular biology abstracts). One approach to solving this problem is to use machine learning techniques to identify the names of specialized entities. Alternatively or in conjunction with these techniques, resources such as lists of gene, protein, chemical and drug names can be exploited [23]. In molecular biology, for example, the Human Genome Organization (HUGO) has established through its Gene Nomenclature Committee (HGNC) a list of over 20,000 approved gene names and symbols, called Genew [24]. Recorded in this database are the symbol *NF2* and the name *neurofibromin 2 (bilateral acoustic neuroma)* for the gene *merlin*, whose mutation causes the disease *neurofibromatosis 2*. More generally, lists of names for specialized entities can be extracted from specialized resources. Examples of publicly available specialized resources for genes, proteins, chemical entities, and drugs are given in Table 3.2. Finally, acronyms and abbreviations harvested from the biomedical literature [25, 26] and collected in databases [27] can also benefit entity recognition applications. This issue is discussed extensively in chapter 5.

Table 3.2 -- Examples of publicly available specialized resources for genes, proteins, chemical entities, and drugs

Domain	Resources	URL
Genes and proteins	Genew	http://www.gene.ucl.ac.uk/nomenclature/
	Entrez Gene[iv]	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
	UniProt	http://www.ebi.uniprot.org/index.shtml
Chemical entities	PubChem	http://pubchem.ncbi.nlm.nih.gov/
	ChemIDplus	http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp
	ChEBI	http://www.ebi.ac.uk/chebi/

Drugs	RxNorm	http://www.nlm.nih.gov/research/umls/rxnorm_main.html
	National Drug Code	http://www.fda.gov/cder/ndc/

3.4 Terminological resources

The purpose of terminology is to collect the names of entities employed in the biomedical domain [28]. Terminologies typically provide lists of synonyms for the entities in a given subdomain and for a given purpose. As such, they play an important role in entity recognition. Additionally, most terminologies have some kind of hierarchical organization that can be exploited for relation extraction purposes. Many terminologies consist of a tree where nodes are terms and links represent parent-to-child or more-general-to-more-specific relationships. Some terminologies allow multiple inheritance and have the structure of a directed acyclic graph. The Gene Ontology and MeSH provide examples of terminological systems created to support different tasks. Because it integrates a large number of terminologies, the UMLS Metathesaurus is the terminological system most frequently used in the analysis of biomedical text.

3.4.1Gene Ontology

The Gene Ontology™ (GO) is a controlled vocabulary developed by the Gene Ontology Consortium for the annotation of gene products in model organisms. GO is organized in three separate hierarchies for molecular functions (6,933 terms), biological processes (9,053 terms) and cellular components (1,414 terms), as of February 1, 2005 [29]. For example, annotations for the gene *NF2* in the GOA database[v] include the molecular function term *cytoskeletal protein binding,* the biological process term *negative regulation of cell proliferation* and the cellular component terms *plasma membrane* and *cytoskeleton*. Each of the three hierarchies is organized in a directed acyclic graph in which the nodes are GO terms and the edges represent the GO relationships ***is a*** and ***part of***. For example, as illustrated in Figure 3.2, the relations of the cellular component *cytoskeleton* to its parent terms include *cytoskeleton* ***is a*** *intracellular non-membrane-bound organelle* and *cytoskeleton* ***part of*** *intracellular*. GO terms may have synonyms (e.g., synonyms for *plasma membrane* include *cytoplasmic membrane* and *plasmalemma*). Most terms have a textual definition (e.g., for *plasma membrane*: "The membrane surrounding a cell that separates the cell from its external environment. It consists of a phospholipid bilayer and associated proteins.").
Both the names and the relations comprised in the Gene Ontology can benefit text mining applications. The names of molecular functions, biological processes and cellular components are frequently used in the biomedical literature [30]. For example, the biological process *activation of MAPK* and the cellular component *adherens junction* can be identified in the title "Erbin regulates MAP kinase activation and MAP kinase-dependent interactions between merlin and adherens

junction protein complexes in Schwann cells". As illustrated in the following text fragment, hierarchical relations can help resolve anaphora and interpret associative relations.

The organization of the actin cytoskeleton in prefusion aligning myoblasts is likely to be important for their shape and interaction. We investigated actin filament organization and polarity by transmission electron microscopy (TEM) in these cells. (3)

The terms *actin cytoskeleton* and *actin filament* identified in the first two sentences of (3) are present in GO. Moreover, a relation between them is explicitly recorded in GO (*actin filament* ***part of*** *actin cytoskeleton*), which helps link together the two sentences. However, many concepts and relations are not represented in GO or other biomedical terminologies. For example, a relation between *myoblasts* and *these cells* – namely *myoblast* ***is a*** *cell* – is needed to resolve the anaphoric relation between the two terms in (3). Such a relation cannot be found in GO where the term *myoblast* is not even represented.

Finally, GO terms constitute an entry point to annotation databases, providing a wealth of relations between gene products and the molecular functions, biological processes and cellular components with which they are associated (e.g., *NF2* ***has biological process*** *negative regulation of cell proliferation*). GO is available from http://geneontology.org/ and is distributed in various formats including XML and database formats. Perl and Java application programming interfaces are also available. GO is one of the source vocabularies included in the UMLS Metathesaurus. GO is a member of a family of controlled vocabularies called Open Biomedical Ontologies (OBO). These resources can be useful in text mining applications as a source of specialized vocabulary (e.g., for chemicals or experimental conditions). OBO resources are available at http://obo.sourceforge.net/.

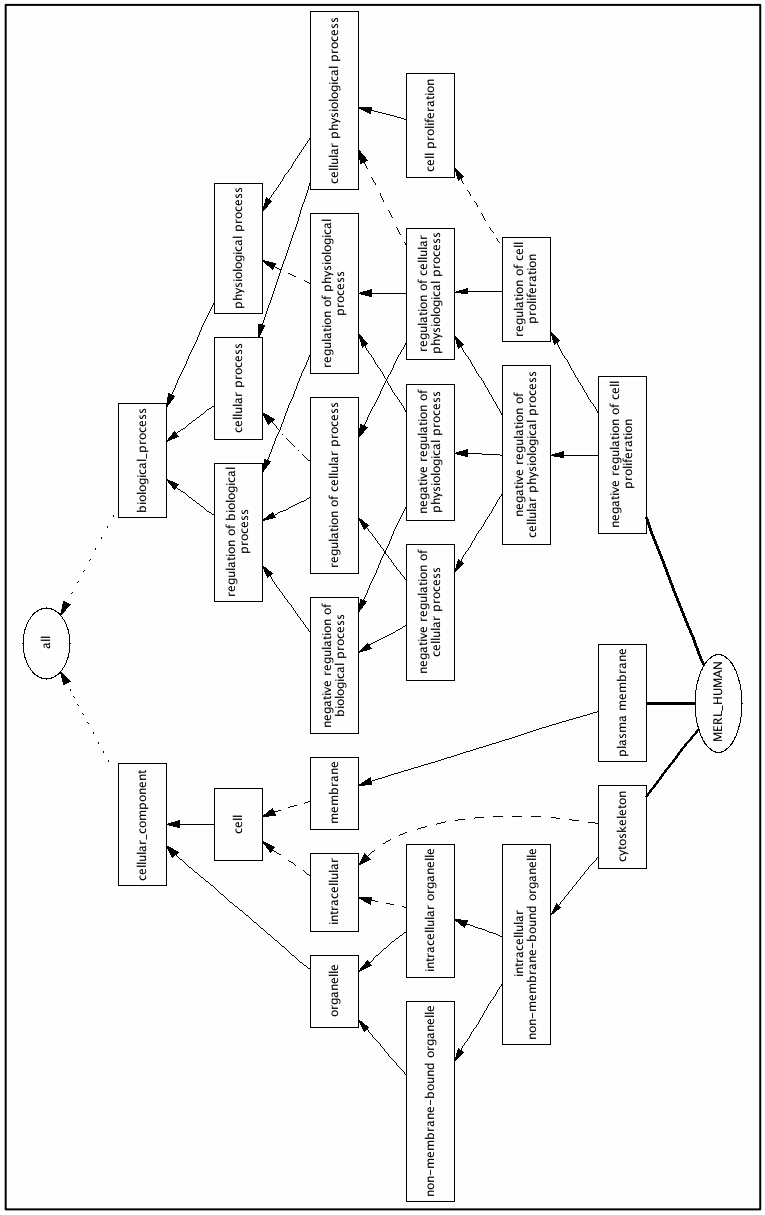

Figure 3.2 – Representation of the gene product *merlin* (*MERL_HUMAN*). Solid lines represent some of its annotations to the Gene Ontology; solid and dashed arrows represent the Gene Ontology relationships ***is a*** and ***part of***, respectively.

3.4.2Medical Subject Headings

The Medical Subject Headings (MeSH®) thesaurus is a controlled vocabulary produced by the National Library of Medicine and used for indexing, cataloging and searching for biomedical and health-related information and documents [31]. It consists of 22,995 descriptors (main headings) organized in fifteen hierarchies. Additionally, a set of about 150,000 "supplementary concept records" provides a finer-grained representation of biomedical entities including chemicals and proteins. A list of entry terms (synonyms or closely related terms) is given for each descriptor. Entry terms for the disease *Neurofibromatosis 2* include *Neurofibromatosis Type II*, *Bilateral Acoustic Neurofibromatosis*, *Bilateral Acoustic Schwannoma* and *Familial Acoustic Neuromas*. A scope note often provides a definition of the descriptor (see (2) for an example). In the MeSH thesaurus, descriptors are related by parent/child relations; each descriptor has at least one parent and may have several. For example, *Neurofibromatoses* and *Neuroma, Acoustic* are the two parents of the descriptor *Neurofibromatosis 2*. The arrangement of MeSH descriptors in hierarchies is intended to serve the purpose of indexing and information retrieval and does not always follow strict classificatory principles. In addition to hierarchical relations, cross-references may link a descriptor to descriptors from other hierarchies. For example, the disease *Neurofibromatosis 2* is linked to the protein *Neurofibromin 2* and to the gene *Genes, Neurofibromatosis 2*. The MeSH thesaurus is used by NLM for indexing articles from 4,600 biomedical journals for the MEDLINE/PubMed database. Like GO, MeSH can be used in text mining applications for the many names and relations it provides. Its scope is broader than GO's, but its granularity is coarser. MeSH is available from http://www.nlm.nih.gov/mesh/ in various formats including XML. MeSH is one of the source vocabularies included in the UMLS Metathesaurus.

3.4.3UMLS Metathesaurus

The UMLS Metathesaurus is one of three knowledge sources developed and distributed by the National Library of Medicine as part of the Unified Medical Language System (UMLS) project [16]. Version 2005AA of the Metathesaurus contains information about over 1 million biomedical concepts and 5 million concept names from more than 100 controlled vocabularies and classifications (some in multiple languages) used in patient records, administrative health data, bibliographic and full-text databases and expert systems. The Metathesaurus also records over 16 million relations among these concepts, inherited from the source vocabularies or specifically generated. While the Metathesaurus preserves the names, meanings, hierarchical contexts, attributes, and inter-term relationships present in its source vocabularies, it also integrates existing terminologies into a common semantic space. Like in WordNet, synonymous names are clustered together to form a concept. Additionally, the Metathesaurus assigns a unique

identifier to each concept and establishes new relations between terms from different source vocabularies as appropriate. Each concept is also categorized with at least one semantic type from the UMLS Semantic Network (see 3.5.2), independently of its hierarchical position in the source vocabularies. The scope of the Metathesaurus is determined by the combined scope of its source vocabularies, including – in addition to Gene Ontology and MeSH – disease vocabularies (e.g., International Classification of Diseases), clinical vocabularies (e.g., SNOMED CT), nomenclatures of drugs and medical devices, as well as the vocabularies of many subdomains of biomedicine (e.g., nursing, psychiatry, gastrointestinal endoscopy).

Examples of Metathesaurus concepts are given in Table 3.1. C0254123 identifies the protein *neurofibromin 2*, whose synonyms include *merlin*, *NF2 protein*, and *schwannomin*. Its semantic types are *Amino Acid, Peptide, or Protein* and *Biologically Active Substance*. The following source vocabularies contributed names to this concept: MeSH, SNOMED CT and the NCI Thesaurus. Once integrated in the Metathesaurus, *neurofibromin 2* has multiple parents including *membrane proteins* (from MeSH), *tumor suppressor proteins* (from both MeSH and SNOMED CT) and *signaling protein* (from the NCI Thesaurus). Its only descendant is *merlin, Drosophila* (from MeSH). Beside hierarchical relations, associative relations link the protein *neurofibromin 2* to the gene *neurofibromatosis 2 genes* and to the disease *neurofibromatosis 2*. Also recorded in the Metathesaurus are the frequencies of co-occurrence of MeSH descriptors in MEDLINE/PubMed citations. For example, during the last ten years, the descriptors *Neurofibromin 2* and *Neurofibromatosis 2* occurred together 13 times as major descriptors. Other descriptors frequently co-occurring with *Neurofibromin 2* include *Membrane Proteins* (8 times), *Phosphoproteins* and *NF2 gene* (7 times) and *Cell Transformation, Neoplastic* (5 times).
Section 3.2 illustrated how the Metathesaurus can be used in entity recognition and relation extraction tasks. Used in many biomedical entity recognition studies, the MetaMap (MMTx) program has been specially designed to take advantage of the features of the UMLS Metathesaurus and SPECIALIST lexicon [32]. MMTx is available from http://mmtx.nlm.nih.gov/. Besides text mining, the Metathesaurus is used in a wide range of applications including linking between different clinical or biomedical vocabularies, information retrieval and indexing, and biomedical language processing. The Metathesaurus is available from http://umlsks.nlm.nih.gov/ (or on DVD) in relational database format. Users are required to complete the *License Agreement for the Use of UMLS Metathesaurus*. Java and XML application programming interfaces are available for the Metathesaurus.

3.5 Ontological resources

Biomedical ontology aims to study the kinds of entities (i.e., substances, qualities and processes) of biomedical significance. Unlike biomedical terminology, biomedical ontology is not primarily concerned with names, but with the

principled definition of biological classes and their interrelations. In practice, however, as most terminologies have some degree of organization and many ontologies also collect names for their entities, the distinction between ontological and terminological resources is somewhat arbitrary. Because they share many characteristics with ontologies, we will list under this rubric two broad resources (SNOMED CT and the UMLS Semantic Network). Other ontologies will be briefly discussed.

3.5.1SNOMED CT

The Systematized Nomenclature of Medicine (SNOMED®) Clinical Terms® (SNOMED CT), developed by the College of American Pathologists, was formed by the convergence of SNOMED RT and Clinical Terms Version 3 (formerly known as the Read Codes). SNOMED CT is the most comprehensive biomedical terminology recently developed in native description logic formalism[vi]. The version described here (January 31, 2004) contains some 270,000 concepts, named by over 400,000 names. SNOMED CT consists of eighteen independent hierarchies reflecting, in part, the organization of previous versions of SNOMED into *axes*, such as *Diseases*, *Drugs*, *Living organisms*, *Procedures* and *Topography*. Each SNOMED CT concept is described by a variable number of elements. For example, the concept *Neurofibromatosis, type 2* has a unique identifier (92503002), several names (*Bilateral acoustic neurofibromatosis, BANF - Bilateral acoustic neurofibromatosis, Neurofibromatosis, type 2* and *Neurofibromatosis type 2*) and has multiple ***is a*** parents including *Congenital anomaly of inner ear*, *Neoplasm of uncertain behavior of cranial nerve* and *Acoustic neuroma*. Additionally, *Neurofibromatosis, type 2* participates in a complex network of associative relations to other concepts. The relations (called *roles*), shown in Table 3.3, indicate, for example, that the lesions encountered in *Neurofibromatosis, type 2* include neurofibromatosis of the vestibulocochlear nerve (group 1) and neurilemoma of the vestibular nerve (group 3). SNOMED CT is available as part of the UMLS (from http://umlsks.nlm.nih.gov/), at no charge for UMLS licensees in the U.S. The structure of the UMLS Metathesaurus has been modified to accommodate the level of detail provided by ontological resources like SNOMED CT. Because SNOMED CT has only become available through the UMLS in 2004, the number of studies reporting its uses is still limited.

Table 3.3 – Some of the roles present in the definition of *Neurofibromatosis, type 2*

Group	Role	Value
1	***Associated morphology***	*Neurofibromatosis*
	Finding site	*Skin structure*
	Finding site	*Vestibulocochlear nerve structure*
3	***Associated morphology***	*Neurilemoma*
	Finding site	*Vestibular nerve structure*

3.5.2UMLS Semantic Network

The UMLS Semantic Network is one of three knowledge sources developed and distributed by the National Library of Medicine as part of the Unified Medical Language System (UMLS) project. It was created in an effort to provide a semantic framework for the UMLS and its constituent vocabularies [33]. Unlike the Metathesaurus, the Semantic Network is a small structure composed of 135 high-level categories called semantic types. It is organized in two single-inheritance hierarchies: one for *Entity* and one for *Event*. In addition to ***is a***, 53 kinds of relationships are defined in the Semantic Network, which are used to represent over 6,700 relations – hierarchical and associative – among semantic types. Semantic types from the Semantic Network are linked to Metathesaurus concepts by the categorization link established by the Metathesaurus editors: Each concept is categorized with at least one semantic type from the Semantic Network, independently of its hierarchical position in the source vocabularies. Fifteen collections of semantic types, called *semantic groups*, have been defined in order to partition Metathesaurus concepts into a smaller number of semantically consistent groups [34].

Semantic types for the Metathesaurus concepts listed in Table 3.1 are presented in Table 3.4, along with the corresponding semantic groups. For example, the concept *Neurofibromatosis 2* is categorized as *Neoplastic Process*, a semantic type from the semantic group *Disorders*. In addition to *mutation*, Metathesaurus concepts categorized with *Genetic Function* include *alternative splicing*, *loss of heterozygosity* and *ribonuclease activity*. Examples of relations among semantic types include *Body Part, Organ, or Organ Component* ***location of*** *Neoplastic Process*, *Pharmacologic Substance* ***treats*** *Neoplastic Process* and *Neoplastic Process* ***manifestation of*** *Genetic Function*. A relationship between two semantic types indicates a *possible link* between the concepts categorized with these semantic types. In natural language processing and text mining applications, Semantic Network relations are typically used as supporting evidence for the candidate predicates (i.e., <$concept_1$, relationship, $concept_2$> structures) extracted from the text [35]. For example, in "schwannomas of cranial nerves", after identifying the concepts *neurilemmoma* (from "schwannoma") as a *Neoplastic Process* and *cranial nerves* as a *Body Part, Organ, or Organ Component*, the preposition *of* can be interpreted as indicating the location of the neoplastic process to the body part. This candidate predicate is supported by the Semantic Network relation *Body Part, Organ, or Organ Component* ***location of*** *Neoplastic Process*. Many relation extraction systems rely on correspondences established between semantic relations and linguistic phenomena [e.g., 36]. Semantic Network relations can also be exploited in conjunction with relations among concepts in the Metathesaurus [e.g., 37]. The Semantic Network is distributed as part of the UMLS and is available from http://umlsks.nlm.nih.gov/. Like the other UMLS knowledge

sources, it can be queried through application programming interfaces for Java and XML.

Table 3.4 – Semantic types and semantic groups for the Metathesaurus concepts listed in Table 3.1

ST abbr.	ST name	Semantic group
aapp	Amino Acid, Peptide, or Protein	Chemicals & Drugs
bacs	Biologically Active Substance	Chemicals & Drugs
bdsy	Body System	Anatomy
bpoc	Body Part, Organ, or Organ Component	Anatomy
celc	Cell Component	Anatomy
dsyn	Disease or Syndrome	Disorders
genf	Genetic Function	Physiology
gngm	Gene or Genome	Genes & Molecular Sequences
neop	Neoplastic Process	Disorders

3.5.3Other ontological resources

In addition to SNOMED CT and the UMLS Semantic Network, several ontological resources can be used to support text mining. The Foundational Model of Anatomy[vii] (FMA) is a large reference ontology of anatomy developed at the University of Washington [38]. In addition to NLP applications [39], the FMA has been used in entity recognition tasks [40] as well as relation extraction tasks [41]. Ontologies such as OpenGALEN[viii] have been developed to support terminological services [42] and may be less useful for text mining applications. For example, unlike terminologies, OpenGALEN does not record lists of synonyms for biomedical entities. For more information about biomedical ontologies, we refer the interested reader to [43].

3.6 Issues related to entity recognition

The biomedical domain has a long tradition of collecting and organizing terms as well as building classifications, dating back the seventeenth century. The dozens of terminological resources resulting from this effort now benefit entity recognition tasks. Moreover, the terminology integration system Unified Medical Language System (UMLS) mentioned earlier has contributed to make existing terminologies both easier to use by providing a common format and distribution mechanism and more useful by identifying synonymy and other semantic relations across them. As part of this effort, the National Library of Medicine (NLM) also developed the lexical resources (lexicon and lexical programs) used to detect lexical similarity among biomedical terms and, more generally, to process biomedical text. This is the reason why the UMLS is used in a large number of text mining systems in biomedicine.

The properties of biomedical terms have been studied. For example, [44, 45] found matches for 10-34% of the UMLS strings in MEDLINE/PubMed (depending on the matching criteria used) and [44] developed a model for identifying the UMLS terms useful in natural language processing (NLP) applications. In the domain of molecular biology, researchers have investigated the lexical properties of the Gene ontology (GO): 35% of GO terms have been found in the biomedical literature [30] and 66% of GO terms are composed of other GO terms [46]. A model of compositionality in GO has even been proposed [47]. These studies have confirmed the interest of using existing terminological resources in entity recognition tasks.

There are, however, some remaining challenges in biomedical entity recognition, including limited coverage of terminological resources and ambiguity in biomedical names.

3.6.1Limited coverage

First, some subdomains remain only partially covered by existing resources. One example is given by genes and proteins and, more generally, chemical entities. Names for such entities have proved difficult to compile in terminologies in an exhaustive manner. Vocabularies extracted from specialized databases may complement traditional terminologies here. Moreover, while variant formation has been studied and effectively modeled for clinical terms [48], normalization techniques for the less regular names of entities employed in genomics have only be recently researched [49]. For these reasons, entity recognition techniques in this subdomain often include machine learning approaches rather than the rule-based approach traditionally employed in biomedical NLP. Many gene names identification systems have been developed in the last five years (see [50-53] for examples). Entity recognition systems in molecular biology texts may include algorithms rather than (or in addition to) static resources [23]. However, the product of some of these algorithms is made available to the research community by their authors. For example, [54] share the lexicon of over one million gene and protein names they have extracted from the biomedical literature. Coverage issues have been explored in clinical terminologies as well [55], and techniques have been developed to extend the coverage of terminologies to specialized subdomains [e.g., 56] or from specific corpora [e.g., 57]. More generally, relation extraction may also benefit from term extraction techniques resulting from research in terminology [58].

3.6.2Ambiguity

The second issue is the ambiguity of many names in biology. This phenomenon is common in natural language but poses specific challenges to biomedical entity recognition. Polysemy (several meanings for the same name) is illustrated by *NF2*, which simultaneously names the gene, the protein it produces and the disease

resulting from its mutation. While polysemy does not usually pose problems for domain experts, it makes it difficult for entity recognition systems to select the appropriate meaning. The ambiguity resulting from polysemous gene names has been quantified by [59]. These authors found modest ambiguities with general English words (0.57%) and medical terms (1.01%), but high ambiguity across species (14.20%). Ambiguity across species may be difficult to resolve, for example when only capitalization conventions differentiate between gene names in various model organisms (e.g., *NF2* in Homo sapiens vs. *Nf2* in Mus musculus). Various disambiguation strategies have been applied to biomedical language processing [e.g., 60, 61]. But further research is needed to develop strategies adapted to the specificity of molecular biology (e.g., ambiguity across species). Moreover, the limited availability of annotated resources such as the GENIA corpus [62] hinders the development of unsupervised disambiguation techniques.

3.7 Issues related to relation extraction

3.7.1Terminological vs. ontological relations

Not only do terminologies contain a large number of names for biomedical entities useful for entity recognition tasks, but they also represent a similarly considerable number of relations. For example, over sixteen million relations are recorded in the UMLS Metathesaurus. While not all of them represent well-defined predicates or assertions as would be expected from ontologies, these relations are essentially beneficial to applications such as relation extraction, especially when used in combination with lexico-syntactic clues and additional ontological relations.

The relations found in the most recent terminologies – often developed using knowledge representation techniques such as description logics – are generally better specified and principled, and therefore more directly useful for relation extraction. However, a careful inspection of these and other ontological resources through the prism of formal ontology reveals some limitations, especially in terms of consistency [15, 63, 64]. Applying formal ontological principles to biomedical ontologies results in clarifying the relations [65], which, in turn, is expected to result in more consistent ontologies and more accurate inferences.

Recent experiments in reengineering terminologies have shown both the benefit and the cost (in terms of human resources) of such efforts [66, 67]. However, improving ontologies is likely to benefit relation extraction as the candidate assertions extracted from text must be checked not necessarily against relations explicitly represented in ontologies, but most often against inferred relations.

3.7.2Interactions between text mining and terminological resources

This chapter deliberately looks at ontologies and other resources as enabling resources for text mining and relation extraction in particular. It is worth mentioning that, conversely, the relations extracted from text corpora and other

knowledge sources (e.g., annotation database) can help identify additional ontological relations. For example, lexico-syntactic patterns have been used to extract hypernymy relations from text corpora [68] and statistical methods have helped identify associative relations among Gene Ontology terms [69]. In other words, the relations between text mining techniques and terminological resources are not unilateral: there is rather a virtuous cycle in which applications and resources benefit from one another. Studying this symbiotic relation is, however, beyond the scope of this chapter. More generally, various existing resources can be combined in order to create new resources. For example, semantic lexicons have been derived from lexicons, terminologies and text corpora [70, 71].

3.8 Conclusion

This chapter has presented the various kinds of enabling resources used in biomedical text mining applications. Lexicons support basic natural language processing tasks such as parsing. Additionally, along with terminologies, lexicons provide lists of names (including variants) for biological entities, supporting entity recognition tasks. Finally, the relations represented in ontologies and terminologies often serve as a reference for relation extraction algorithms.

Because it integrates these three kinds of resources, the Unified Medical Language System (UMLS) plays a central role in biomedical text mining. This chapter illustrated the use of its three components (SPECIALIST lexicon, Metathesaurus and Semantic Network) in entity recognition and relation extraction tasks. The role of other resources, either more specialized or more general, was also discussed.

Despite the existence of these resources, there remain many challenges to entity recognition and relation extraction in biology. Existing biomedical lexicons and terminologies fail to provide adequate coverage of specialized subdomains (e.g., genes and proteins for the various model organisms). Approaches to normalizing the names of genomic entities and to resolving the ambiguity introduced by some of them need to be further researched. Finally, the development of large, consistent, principled sources of biomedical knowledge – namely ontologies – will benefit not only text mining applications, but more generally the wide range of tasks relying upon biomedical knowledge (e.g., database interoperability, decision support, etc).

3.9 Acknowledgments

The author would like to thank Tom Rindflesch, Pierre Zweigenbaum and Karin Verspoor for useful comments on a previous version of this manuscript.

3.10 References

[1] de Bruijn, B., and J. Martin. "Getting to the (C)Ore of Knowledge: Mining Biomedical Literature." *Int J Med Inform*, Vol. 67, No. 1-3, 2002, pp. 7-18.

[2] Shatkay, H., and R. Feldman. "Mining the Biomedical Literature in the Genomic Era: An Overview." *J Comput Biol*, Vol. 10, No. 6, 2003, pp. 821-55.

[3] Yandell, M. D., and W. H. Majoros. "Genomics and Natural Language Processing." *Nat Rev Genet*, Vol. 3, No. 8, 2002, pp. 601-10.

[4] Hahn, U., M. Romacker, and S. Schulz. "Medsyndikate--a Natural Language System for the Extraction of Medical Information from Findings Reports." *Int J Med Inform*, Vol. 67, No. 1-3, 2002, pp. 63-74.

[5] Corney, D. P., B. F. Buxton, W. B. Langdon, et al. "BioRAT: Extracting Biological Information from Full-Length Papers." *Bioinformatics*, Vol. 20, No. 17, 2004, pp. 3206-13.

[6] Leroy, G., H. Chen, and J. D. Martinez. "A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text." *J Biomed Inform*, Vol. 36, No. 3, 2003, pp. 145-58.

[7] Humphreys, K., G. Demetriou, and R. Gaizauskas. "Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures." *Pac Symp Biocomput*, 2000, pp. 505-16.

[8] Friedman, C., P. O. Alderson, J. H. Austin, et al. "A General Natural-Language Text Processor for Clinical Radiology." *J Am Med Inform Assoc*, Vol. 1, No. 2, 1994, pp. 161-74.

[9] Friedman, C., P. Kra, H. Yu, et al. "GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles." *Bioinformatics*, Vol. 17 Suppl 1, 2001, pp. S74-82.

[10] Rindflesch, T. C., M. Fiszman, and B. Libbus. "Semantic Interpretation for the Biomedical Literature." In *Medical Informatics: Advances in Knowledge Management and Data Mining in Biomedicine*, pp. 399-422, H. Chen, S. Fuller, W.R. Hersh and C. Friedman (eds.): Springer-Verlag, 2005.

[11] Grabar, N., P. Zweigenbaum, L. Soualmia, et al. "Matching Controlled Vocabulary Words." *Stud Health Technol Inform*, Vol. 95, 2003, pp. 445-50.

[12] Bodenreider, O., and C. A. Bean. "Relationships among Knowledge Structures: Vocabulary Integration within a Subject Domain." In *Relationships in the Organization of Knowledge*, pp. 81-98, C.A. Bean and R. Green (eds.): Kluwer, 2001.

[13] Nenadic, G., I. Spasic, and S. Ananiadou. "Terminology-Driven Mining of Biomedical Literature." *Bioinformatics*, Vol. 19, No. 8, 2003, pp. 938-43.

[14] Smith, B., W. Ceusters, B. Klagges, et al. "Relations in Biomedical Ontologies." *Genome Biol*, Vol. 6, No. 5, 2005, pp. R46.
[15] Smith, B., J. Williams, and S. Schulze-Kremer. "The Ontology of the Gene Ontology." *AMIA Annu Symp Proc*, 2003, pp. 609-13.
[16] Bodenreider, O. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology." *Nucleic Acids Res*, Vol. 32, No. Database issue, 2004, pp. D267-70.
[17] Baser, M. E., D. G. Evans, and D. H. Gutmann. "Neurofibromatosis 2." *Curr Opin Neurol*, Vol. 16, No. 1, 2003, pp. 27-33.
[18] Fellbaum, C. *WordNet: An Electronic Lexical Database, Language, Speech, and Communication*. Cambridge, Mass: MIT Press, 1998.
[19] Bodenreider, O., A. Burgun, and J. A. Mitchell. "Evaluation of WordNet as a Source of Lay Knowledge for Molecular Biology and Genetic Diseases: A Feasibility Study." *Stud Health Technol Inform*, Vol. 95, 2003, pp. 379-84.
[20] Burgun, A., and O. Bodenreider. "Comparing Terms, Concepts and Semantic Classes in WordNet and the Unified Medical Language System." *Proceedings of the NAACL'2001 Workshop, "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"*, 2001, pp. 77-82.
[21] Leroy, G., and H. Chen. "Meeting Medical Terminology Needs--the Ontology-Enhanced Medical Concept Mapper." *IEEE Trans Inf Technol Biomed*, Vol. 5, No. 4, 2001, pp. 261-70.
[22] Browne, A. C., G. Divita, A. R. Aronson, et al. "UMLS Language and Vocabulary Tools." *AMIA Annu Symp Proc*, 2003, pp. 798.
[23] Krauthammer, M., and G. Nenadic. "Term Identification in the Biomedical Literature." *J Biomed Inform*, Vol. 37, No. 6, 2004, pp. 512-26.
[24] Wain, H. M., M. J. Lush, F. Ducluzeau, et al. "Genew: The Human Gene Nomenclature Database, 2004 Updates." *Nucleic Acids Res*, Vol. 32, No. Database issue, 2004, pp. D255-7.
[25] Pustejovsky, J., J. Castano, B. Cochran, et al. "Automatic Extraction of Acronym-Meaning Pairs from MEDLINE Databases." *Medinfo*, Vol. 10, No. Pt 1, 2001, pp. 371-5.
[26] Schwartz, A. S., and M. A. Hearst. "A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text." *Pac Symp Biocomput*, 2003, pp. 451-62.
[27] Wren, J. D., J. T. Chang, J. Pustejovsky, et al. "Biomedical Term Mapping Databases." *Nucleic Acids Res*, Vol. 33 Database Issue, 2005, pp. D289-93.
[28] Chute, C. G. "Clinical Classification and Terminology: Some History and Current Observations." *J Am Med Inform Assoc*, Vol. 7, No. 3, 2000, pp. 298-303.

[29] Ashburner, M., C. A. Ball, J. A. Blake, et al. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." *Nat Genet*, Vol. 25, No. 1, 2000, pp. 25-9.

[30] McCray, A. T., A. C. Browne, and O. Bodenreider. "The Lexical Properties of the Gene Ontology." *Proc AMIA Symp*, 2002, pp. 504-8.

[31] Nelson, S. J., D. Johnston, and B. L. Humphreys. "Relationships in Medical Subject Headings." In *Relationships in the Organization of Knowledge*, pp. 171-84, C.A. Bean and R. Green (eds.): Kluwer, 2001.

[32] Aronson, A. R. "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The Metamap Program." *Proc AMIA Symp*, 2001, pp. 17-21.

[33] McCray, A. T. "An Upper-Level Ontology for the Biomedical Domain." *Comp Funct Genom.*, No. 4, 2003, pp. 80-4.

[34] Bodenreider, O., and A. T. McCray. "Exploring Semantic Groups through Visual Approaches." *J Biomed Inform*, Vol. 36, No. 6, 2003, pp. 414-32.

[35] Rindflesch, T. C., and M. Fiszman. "The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text." *J Biomed Inform*, Vol. 36, No. 6, 2003, pp. 462-77.

[36] Libbus, B., H. Kilicoglu, T. Rindflesch, et al. "Using Natural Languge Processing, LocusLink, and the Gene Ontology to Compare OMIM to MEDLINE." *Proceedings of the HLT-NAACL Workshop on Linking the biological literature, ontologies and databases: Tools for users*, 2004, pp. 69-76.

[37] McCray, A. T., and O. Bodenreider. "A Conceptual Framework for the Biomedical Domain." In *The Semantics of Relationships: An Interdisciplinary Perspective*, pp. 181-98, R. Green, C.A. Bean and S.H. Myaeng (eds.), Boston: Kluwer Academic Publishers, 2002.

[38] Rosse, C., and J. L. Mejino, Jr. "A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy." *J Biomed Inform*, Vol. 36, No. 6, 2003, pp. 478-500.

[39] Distelhorst, G., V. Srivastava, C. Rosse, et al. "A Prototype Natural Language Interface to a Large Complex Knowledge Base, the Foundational Model of Anatomy." *AMIA Annu Symp Proc*, 2003, pp. 200-4.

[40] Sneiderman, C. A., T. C. Rindflesch, and C. A. Bean. "Identification of Anatomical Terminology in Medical Text." *Proc AMIA Symp*, 1998, pp. 428-32.

[41] Bean, C. A., T. C. Rindflesch, and C. A. Sneiderman. "Automatic Semantic Interpretation of Anatomic Spatial Relationships in Clinical Text." *Proc AMIA Symp*, 1998, pp. 897-901.

[42] Nowlan, W. A., A. L. Rector, T. W. Rush, et al. "From Terminology to Terminology Services." *Proc Annu Symp Comput Appl Med Care*, 1994, pp. 150-4.

[43] Bodenreider, O., and A. Burgun. "Biomedical Ontologies." In *Medical Informatics: Advances in Knowledge Management and Data Mining in Biomedicine*, pp. 211-36, H. Chen, S. Fuller, W.R. Hersh and C. Friedman (eds.): Springer-Verlag, 2005.

[44] McCray, A. T., O. Bodenreider, J. D. Malley, et al. "Evaluating UMLS Strings for Natural Language Processing." *Proc AMIA Symp*, 2001, pp. 448-52.

[45] Srinivasan, S., T. C. Rindflesch, W. T. Hole, et al. "Finding UMLS Metathesaurus Concepts in MEDLINE." *Proc AMIA Symp*, 2002, pp. 727-31.

[46] Ogren, P. V., K. B. Cohen, G. K. Acquaah-Mensah, et al. "The Compositional Structure of Gene Ontology Terms." *Pac Symp Biocomput*, 2004, pp. 214-25.

[47] Mungall, C. "Obol: Integrating Language and Meaning in Bio-Ontologies." *Comparative and Functional Genomics*, Vol. 5, No. 7, 2004, pp. 509-20.

[48] McCray, A. T., S. Srinivasan, and A. C. Browne. "Lexical Methods for Managing Variation in Biomedical Terminologies." *Proc Annu Symp Comput Appl Med Care*, 1994, pp. 235-9.

[49] Morgan, A. A., L. Hirschman, M. Colosimo, et al. "Gene Name Identification and Normalization Using a Model Organism Database." *J Biomed Inform*, Vol. 37, No. 6, 2004, pp. 396-410.

[50] Collier, N., and K. Takeuchi. "Comparison of Character-Level and Part of Speech Features for Name Recognition in Biomedical Texts." *J Biomed Inform*, Vol. 37, No. 6, 2004, pp. 423-35.

[51] Koike, A., Y. Niwa, and T. Takagi. "Automatic Extraction of Gene/Protein Biological Functions from Biomedical Text." *Bioinformatics*, Vol. 21, No. 7, 2005, pp. 1227-36.

[52] Proux, D., F. Rechenmann, L. Julliard, et al. "Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction." *Genome Inform Ser Workshop Genome Inform*, Vol. 9, 1998, pp. 72-80.

[53] Yu, H., V. Hatzivassiloglou, C. Friedman, et al. "Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles." *Proc AMIA Symp*, 2002, pp. 919-23.

[54] Tanabe, L., and W. J. Wilbur. "Generation of a Large Gene/Protein Lexicon by Morphological Pattern Analysis." *J Bioinform Comput Biol*, Vol. 1, No. 4, 2004, pp. 611-26.

[55] Chute, C. G., S. P. Cohn, K. E. Campbell, et al. "The Content Coverage of Clinical Classifications. For the Computer-Based Patient Record Institute's Work Group on Codes & Structures." *J Am Med Inform Assoc*, Vol. 3, No. 3, 1996, pp. 224-33.

[56] Harris, M. R., G. K. Savova, T. M. Johnson, et al. "A Term Extraction Tool for Expanding Content in the Domain of Functioning, Disability,

and Health: Proof of Concept." *J Biomed Inform*, Vol. 36, No. 4-5, 2003, pp. 250-9.

[57] Bodenreider, O., T. C. Rindflesch, and A. Burgun. "Unsupervised, Corpus-Based Method for Extending a Biomedical Terminology." *Proceedings of the ACL'2002 Workshop "Natural Language Processing in the Biomedical Domain"*, 2002, pp. 53-60.

[58] Jacquemin, C. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, Mass.: MIT Press, 2001.

[59] Chen, L., H. Liu, and C. Friedman. "Gene Name Ambiguity of Eukaryotic Nomenclatures." *Bioinformatics*, Vol. 21, No. 2, 2005, pp. 248-56.

[60] Liu, H., Y. A. Lussier, and C. Friedman. "Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text: An Unsupervised Method." *J Biomed Inform*, Vol. 34, No. 4, 2001, pp. 249-61.

[61] Liu, H., V. Teller, and C. Friedman. "A Multi-Aspect Comparison Study of Supervised Word Sense Disambiguation." *J Am Med Inform Assoc*, Vol. 11, No. 4, 2004, pp. 320-31.

[62] Kim, J. D., T. Ohta, Y. Tateisi, et al. "Genia Corpus--Semantically Annotated Corpus for Bio-Textmining." *Bioinformatics*, Vol. 19 Suppl 1, 2003, pp. i180-2.

[63] Ceusters, W., B. Smith, A. Kumar, et al. "Ontology-Based Error Detection in SNOMED-Ct(R)." *Medinfo*, Vol. 2004, 2004, pp. 482-6.

[64] Smith, B., A. Kumar, and S. Schulze-Kremer. "Revising the UMLS Semantic Network." *Medinfo*, Vol. 2004, No. CD, 2004, pp. 1700.

[65] Smith, B., and C. Rosse. "The Role of Foundational Relations in the Alignment of Biomedical Ontologies." *Medinfo*, Vol. 2004, 2004, pp. 444-8.

[66] Hahn, U., and S. Schulz. "Towards a Broad-Coverage Biomedical Ontology Based on Description Logics." *Pac Symp Biocomput*, 2003, pp. 577-88.

[67] Wroe, C. J., R. Stevens, C. A. Goble, et al. "A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL." *Pac Symp Biocomput*, 2003, pp. 624-35.

[68] Hearst, M. A. "Automatic Acquisition of Hyponyms from Large Text Corpora." *Proceedings of COLING*, 1992, pp. 539-45.

[69] Bodenreider, O., M. Aubry, and A. Burgun. "Non-Lexical Approaches to Identifying Associative Relations in the Gene Ontology." In *Pacific Symposium on Biocomputing 2005*, pp. 91-102, R.B. Altman, A.K. Dunker, L. Hunter, T.A. Jung and T.E. Klein (eds.): World Scientific, 2005.

[70] Johnson, S. B. "A Semantic Lexicon for Medical Language Processing." *J Am Med Inform Assoc*, Vol. 6, No. 3, 1999, pp. 205-18.

[71] Verspoor, K. "Towards a Semantic Lexicon for Biological Language Processing." *Comparative and Functional Genomics*, Vol. 6, No. 1-2, 2005, pp. 61-66.

[i] Language & Computing (http://www.landcglobal.com/)
[ii] Uppal, S., and A. P. Coatesworth. "Neurofibromatosis Type 2." *Int J Clin Pract*, 57, no. 8, 2003, pp. 698-703.
[iii] http://www.nlm.nih.gov/mesh/
[iv] Formerly LocusLink
[v] http://www.ebi.ac.uk/GOA/
[vi] Although not distributed in any description logic (DL) language, SNOMED CT has been developed in native DL formalism, thus ensuring the consistency of its hierarchical relations.
[vii] http://fma.biostr.washington.edu/
[viii] http://www.opengalen.org/