An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature

Harsha Gurulingappa[*†], **Roman Klinger**[*], **Martin Hofmann-Apitius**[*†], **and Juliane Fluck**[*]

[*]Fraunhofer Institute for Algorithms and Scientific Computing
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[†]Bonn-Aachen International Center for Information Technology
Dahlmannstraße 2, 53113 Bonn, Germany
harsha.gurulingappa@scai-extern.fraunhofer.de,
{roman.klinger, martin.hofmann-apitius, and juliane.fluck}@scai.fraunhofer.de

Abstract

The mentions of human health perturbations such as the diseases and adverse effects denote a special entity class in the biomedical literature. They help in understanding the underlying risk factors and develop a preventive rationale. The recognition of these named entities in texts through dictionary-based approaches relies on the availability of appropriate terminological resources. Although few resources are publicly available, not all are suitable for the text mining needs. Therefore, this work provides an overview of the well known resources with respect to human diseases and adverse effects such as the MeSH, MedDRA, ICD-10, SNOMED CT, and UMLS. Individual dictionaries are generated from these resources and their performance in recognizing the named entities is evaluated over a manually annotated corpus. In addition, the steps for curating the dictionaries, rule-based acronym disambiguation and their impact on the dictionary performance is discussed. The results show that the MedDRA and UMLS achieve the best recall. Besides this, MedDRA provides an additional benefit of achieving a higher precision. The combination of search results of all the dictionaries achieve a considerably high recall. The corpus is available on `http://www.scai.fraunhofer.de/disease-ae-corpus.html`

1. Introduction

In the field of biomedical sciences, a huge amount of unstructured textual data is generated every year in the form of research articles, patient health records, clinical reports, medical narratives and patents (Karsten and Suominen, 2009; Cohen and Hersh, 2005). Enormous efforts have been invested in parallel to extract potentially useful information from these textual records (Wang et al., 2009; Chen et al., 2008). Therefore, automatic processing of literature data has gained popularity since over a decade, for example named entity recognition or key concept identification (Smith et al., 2008).

Named entity recognition serves as a basis for biomedical text mining in order to have key entities tagged before they can be subjected to relationship mining or semantic text interpretation. It deals with the identification of boundaries of terms in the text that represent biologically meaningful objects of interest such as genes, proteins, or diseases. Quite a lot of work has been done for the recognition of gene and protein names. For example, the BioCreAtIvE competitions address the challenges associated with the gene name recognition and normalization (Krallinger et al., 2008). Nevertheless, some groups have proposed different solutions for the identification of other interesting classes of biomedical entities such as drug names (Segura-Bedmar et al., 2008; Hettne et al., 2009) or disease names (Jimeno et al., 2008). However, in comparison to the gene and protein name recognition, only a little work has been invested for the recognition of disease names and particularly adverse effects in the free texts. This is partly due to a fact that the availability of annotated corpora is limited and they are of high cost for generation.

A disease in the context of human health is an abnormal condition that impairs the bodily functions and is associated with physiological discomfort or dysfunction. Similarly, an adverse effect is a health impairment that occurs as a result of intervention of a drug, treatment or therapy (Ahmad, 2003). The severity of adverse effects can range from mild signs or symptoms such as *nausea* and *abdominal discomfort* to irreversible damage such as *perinatal death*. Therefore, the mentions of both diseases and adverse effects in free texts denote special entity classes for the medical experts, clinical professionals as well as health care companies (Hauben and Bate, 2009; Forster et al., 2005). This not only helps in understanding the underlying hypothetical causes but also provide rationale means to prevent or diagnose such abnormal medical conditions. Specially in the clinical scenario, recognizing the adverse effects in medical literature can support the clinical decision making (Stricker and Psaty, 2004).

Some research work has been done in the past for the identification of diseases and adverse effects. Jimeno et al. (2008) proposed a statistical solution for the identification of diseases in a corpus of annotated sentences. They reused the corpus that was provided by Ray and Craven (2001) but the corpus has a limitation of being restricted to OMIM[1] diseases only that mostly include genetic disorders. Neveol et al. (2009) utilized the same corpus as well as PubMed[2] user queries for the detection of disease names. They adapted a statistical model and a natural language processing algorithm within their framework. Leaman et al. (2009) proposed a machine learning based technique for the identification of diseases in a corpus containing over

[1] Online Mendelian Inheritance in Man (OMIM): http://www.ncbi.nlm.nih.gov/omim/

[2] http://www.ncbi.nlm.nih.gov/pubmed/

2,500 sentences from PubMed. This corpus is made publicly available as the Arizona Disease Corpus (AZDC)[3] but the annotations are restricted to the diseases only and do not contain information about adverse effects. Curino et al. (2005) proposed a machine learning based solution for mining adverse effects of specific drugs from the web pages. They generated an adverse effect dictionary from the resources provided by the FDA[4]. However, the corpus utilized by Curino et al. (2005) is not openly available. McCray et al. (2001) proposed a statistical solution for mapping the terms in the corpus to the UMLS concepts. They determined the likelihood of a given UMLS string being found or not found in the corpus. A classical example of a tool for mapping the text to biomedical concepts in UMLS[5] meta-thesaurus is the MetaMap program (Aronson, 2001).

Several terminological resources are available that provide information about diseases and adverse effects. Few well known examples include the MeSH[6] thesaurus, the UMLS[7] meta-thesaurus, the ICD-10[8], and the NCI[9] thesaurus. These resources serve as a good basis for the dictionary-based named entity recognition in text but not all of them essentially suit the text mining needs. Although some of these resources have been utilized individually in the past for the detection of disease names (Jimeno et al., 2008; Chun et al., 2006), there is no common platform where most of these resources have been collectively evaluated.

The aim of this work is to provide an overview of the different data sources and evaluate the general usability of the contained disease and adverse effect terminology for named entity recognition. Although, a small set of corpus is available that contain sentences annotated with disease names, there is no freely available corpus containing the PubMed abstracts that are annotated with diseases as well as adverse effects. Therefore, a newly annotated corpora is made publicly available.

2. Terminological Resources

Dictionary-based named entity recognition approaches rely on comprehensive terminologies containing frequently used synonyms and spelling variants. Such resources include databases, ontologies, controlled vocabularies and thesauri. This section gives an overview of the available data sources for diseases and adverse effects. Examples of synonyms and term variants associated with the MeSH disease concepts are provided in Table 1.

Different resources have been designed to meet the needs of different user groups whereas some of them include certain disease specific information. For example, the NCI thesaurus serves as a reference terminology and an ontology providing a broad coverage of cancer domain including cancer related diseases, findings, abnormalities, gene products, drugs, and chemicals. Similarly, there are databases that include very specific organ or disease class related information such as the autoimmune disease database (Karopka et al., 2006) and the DSM-IV Codes[10] which is specific to mental disorders. On the other hand, sources such as the ICD-10, the UMLS and the MedDRA[11] provide a wider coverage of diseases, signs, symptoms, and abnormal findings irrespective of any kind of disease or any affected organ system. All these resources have their own advantages and areas of applicability. Therefore, the survey made here includes only those resources that encompass information about medical abnormalities that are associated with the entire human physiology.

From all the resources introduced here, individual dictionaries were generated and evaluated over a manually annotated corpus. Although, the MeSH, ICD-10, MedDRA, and SNOMED CT are already included as source vocabularies within the UMLS, these resources were separately downloaded from their respective official websites. The main reason is because when the terms from the source vocabularies are imported into the UMLS, they undergo a series of term modification steps [12]. This generates an impression that the terms present in the UMLS may not be identical to the terms present in the source vocabularies. Therefore, in order to validate the hypothesis of suitability of the individual resources for text mining, they were treated as independent terminologies.

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus from the NLM[13]. It is used by NLM for indexing articles from the PubMed database as well as books, documents, and audiovisuals acquired by the library (Coletti and Bleich, 2001). In MeSH, the terms are arranged in a hierarchical order that are associated with synonyms and term variants. A subset of MeSH that corresponds to the category *Diseases* (tree concepts with node identifiers starting with ‘C’) was extracted to generate a dictionary covering diseases and adverse effects. The MeSH dictionary contains over 4,500 entries.

Medical Dictionary for Regulatory Activities (MedDRA) is a standardized medical terminology that was developed to share regulatory information internationally about medical products used by human (Merrill, 2008). It provides a hierarchical structure of terms that include signs, symptoms, diseases, diagnosis, therapeutic indications, medical procedures, and familial histories. The MedDRA dictionary contains over 20,000 entries associated with synonyms and term variants.

International Classification of Diseases (ICD-10) is

[3] http://diego.asu.edu/downloads/AZDC/

[4] Food and Drug Administration (FDA): http://www.fda.gov/

[5] http://www.nlm.nih.gov/research/umls/

[6] Medical Subject Headings (MeSH): http://www.nlm.nih.gov/mesh/

[7] Unified Medical Language System (UMLS): http://www.nlm.nih.gov/research/umls/

[8] International Classification of Diseases Edition-10 (ICD-10): http://apps.who.int/classifications/ apps/icd/icd10online/

[9] National Cancer Institute (NCI): http://nciterms.nci.nih.gov/

[10] Diagnostic and Statistical Manual of Mental Disorders (DSM) 4th Edition: http://www.psych.org/mainmenu/research/dsmiv/dsmivtr.aspx

[11] Medical Dictionary for Regulatory Activities (MedDRA): http://www.meddramsso.com/

[12] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/source_faq.html#what_involved

[13] National Library of Medicine (NLM): http://www.nlm.nih.gov/

maintained by WHO[14] and it is used to classify diseases and heath problems recorded in many types of health and vital reports including death certificates and health records. The ICD-10 provides terms that are hierarchically ordered according to the organ system that is being affected. Unlike other resources, the ICD provides a flat list of terms and does not include synonyms or term variants. The complete ICD-10 was used for generating the dictionary and it contains over 70,000 entries altogether.

Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) [15] is a comprehensive clinical terminology that is maintained and distributed by IHTSDO[16] (Cornet, 2009). It covers most areas of clinical information such as diseases, findings, procedures, microorganisms, pharmaceuticals etc. The SNOMED CT concepts are organized into hierarchies and the sub-hierarchy that corresponds to *Disorder* was used to generate a dictionary. The SNOMED CT dictionary contains over 90,000 concepts associated with synonyms and term variants.

Unified Medical Language System (UMLS) is a very large, multipurpose, and multilingual meta-thesaurus that contains information about biomedical and health related concepts (Browne et al., 2003). Overall, the UMLS has more than 2 million concepts that are associated with synonyms and relationships between them. The concepts in the UMLS are categorized into semantic groups. The semantic group *Disorders* contains semantic subgroups such as *Acquired Abnormality*, *Disease or Syndrome*, *Mental or Behavioral Dysfunction*, *Sign or Symptom*, etc. Although, the downloadable subset of the UMLS enclose large subsets of concepts from sub-thesauri such as the ICD-9, ICD-10, SNOMED CT, and MeSH, the level of ambiguity it contains has been well demonstrated (Aronson, 2000; Rindflesch and Aronson, 1994). Therefore, we presumed to test the UMLS separately in addition to its constituent sources. All concepts in the *Disorders* semantic group of the UMLS were used to generate a dictionary. This dictionary contains over 120,000 entries altogether.

3. Dictionary Characteristics

The dictionaries generated for the recognition of diseases and adverse effects were analyzed with regard to the following properties:

- Total number of entries,
- Number of synonyms provided, and
- Availability of mappings to other data sources

Table 2 provides a quantitative estimate of the entities present in the raw dictionaries. The UMLS has the largest collection of disease and adverse effect data followed by the SNOMED CT. Figure 1 shows the distribution of synonyms for all the analyzed dictionaries. Since the ICD-10 does not provide synonyms and term variants, it is visible only as a point in Figure 1. A large part of all the dictionaries contain less than 20 synonyms. Few entries in the UMLS, MeSH, and MedDRA[17] are associated with as much as more than 60 synonyms. Resources with high number of synonyms are of great value for dictionary-based named entity recognition approaches. They help to overcome a high false negative rate but may pose a risk of high number of false positives requiring a dedicated curation.

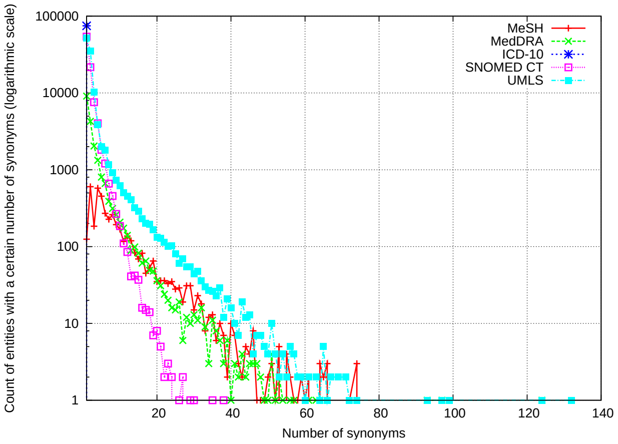

Figure 1: Plot of the synonym count distribution for all the analyzed dictionaries

Since UMLS is the largest resource, a survey was conducted to check the percentage of synonyms that overlap with synonyms in rest of the resources. The synonym comparison between the different resources was performed using a simple case-insensitive string match (i. e. only complete string matches were accepted). About 96 % of the MeSH and 23 % of the MedDRA synonyms are present in UMLS. Only 4 % of the ICD-10 and 13 % of the SNOMED CT synonyms are covered by UMLS. Hence, the outcome of this survey showed that integrating the smaller resources with UMLS would account for an enhanced terminology coverage.

Although, there is an enormous variation in size of the dictionaries used, their adaptability for finding terms in the text is questionable. A manual survey was performed concerning the quality of information contained in each of these dictionaries. The UMLS and SNOMED CT contained over 20,000 terms each that had special characters such as '@', '#&', '[X]', etc. enclosed within the terms. Examples of such ambiguous terms found in the UMLS are *5-@FLUOROURACIL TOXICITY* and *Congestive heart failure #&124*. A large subset of terms were too long and descriptive composed of more than 10 words. Such synonyms are seldom found in the text. An example of such descriptive term found in ICD-10 is *Nondisplaced fracture of lateral condyle of right femur, initial encounter for closed fracture*. ICD-10 has nearly 35,000 long descriptive terms

[14] World Health Organization (WHO): http://www.who.int/en/

[15] http://www.nlm.nih.gov/research/umls/Snomed

[16] International Health Terminology Standards Development Organisation (IHTSDO): http://www.ihtsdo.org/

[17] MedDRA, the Medical Dictionary for Regulatory Activities terminology is the international medical terminology developed under the auspices of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). MedDRA is a registered trademark of the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA)

ID	Concept	Synonyms
D000292	Pelvic Inflammatory Disease	Adnexitis, Inflammatory Disease; Pelvic, Inflammatory Pelvic Disease; Pelvic Disease, Inflammatory
D002534	Brain Hypoxia	Anoxia, Brain; Anoxic Brain Damage; Brain Anoxia; Brain Hypoxia; Cerebral Hypoxia; Encephalopathy, Hypoxic; Hypoxic Brain Damage; Hypoxic Encephalopathy

Table 1: Examples of synonyms and term variants associated with the concepts in the MeSH database.

	MeSH	MedDRA	ICD-10	SNOMED CT	UMLS
No. of entries	4,350	20,515	74,830	92,376	112,341
No. of synonyms (incl. concepts)	42,631	69,121	74,830	170,561	295,773
Percentage of synonyms covered by UMLS	96 %	23 %	4 %	13 %	100 %
Mappings	no	yes	no	yes	yes

Table 2: A quantitative analysis of the dictionaries generated for the disease and side effect named entity recognition. Total number of entries, number of synonyms, percentage of synonyms covered by UMLS, and the availability of inter data source mappings for individual dictionaries are reported. For the UMLS coverage, all synonyms of all the entries were compared.

which constitutes nearly 50 % of the entire dictionary. According to the experience of curators, MeSH and MedDRA were regarded as the specialized resources with considerably low level of ambiguity. Nevertheless, few vague entries such as *Acting out*, *Alcohol Consumption*, and *Childhood* were encountered in these dictionaries.

4. Corpus Characteristics and Annotation

For evaluating the performance of named entity recognition systems, an annotated corpus is necessary. Since, there is no freely available corpus that contains annotations of disease and adverse effect entities, a corpus containing 400 randomly selected MEDLINE abstracts was generated using 'Disease OR Adverse effect' as a PubMed query. This evaluation corpus was annotated by two individuals who hold a Master's degree in life sciences. All the abstracts were annotated with two entity classes, i. e., *disease* and *adverse effect*. In order to obtain a good estimate of the level of agreement between the annotators, they were insisted to carry out the task independently. First, one annotator participated in the development of a guideline for annotation. The corpus was iteratively annotated by this person along with the standardization of the annotation rules. Later, the second person annotated the whole corpus based on the annotation guideline generated by the first annotator. This procedure formed an evaluation corpus of 400 abstracts containing 1428 disease and 813 adverse effect annotations. Recognizing the boundaries without considering the different classes in the evaluation corpus, the inter-annotator agreement F_1 score and kappa (κ) between the two annotators are 84 % and 89 % respectively which indicates a substantial agreement.

The annotation of disease and adverse effect entities were performed very sensitively taking the context into account. Several instances occurred where the disease names and adverse effect names were the same. For example, in the sentence *Hypersensitivity reactions including fever, rash and (more seriously) agranulocytosis are associated with procainamide, and a frequent adverse effect requiring cessation of therapy is the development of systemic lupus erythematosus. (PMID: 2285495)*, the term *systemic lupus erythematosus* occurs as an adverse effect associated with procainamide treatment. In contrary, the sentence *IL-17 expression was found to be associated with many inflammatory diseases in humans, such as rheumatoid arthritis, asthma, systemic lupus erythematosus and allograft rejection and many in vitro studies have indicated a proinflammatory function for IL-17. (PMID: 20338742)* contains *systemic lupus erythematosus* as a disease associated with certain gene function. In such cases, the annotators were strictly insisted to use the contextual information for annotating the entities. Entities that overlap with semantic classes *disease* and *adverse effect* are difficult to be recognized unless a context-based disambiguation is performed. Altogether, there were 178 annotated entities had an overlap with the classes *disease* and *adverse effect*.

5. Results of Dictionary Performance

For the identification of named entities in text, the ProMiner (Hanisch et al., 2005) system was used along with different dictionaries. The text searching with ProMiner was performed using the raw or unprocessed dictionaries as well as with the processed dictionaries. The search was performed using case-insensitive, word order-sensitive and the longest string match as constraints.

The performance of the ProMiner runs with different dictionaries was evaluated using the Precision and Recall. The evaluations were performed for the complete match as well as partial match between the annotated entities and the dictionary terms. A partial match is a situation where either the left boundary or the right boundary of the annotated entity and the ProMiner search result are matched.

The results with raw dictionaries and such a simple search strategy gives a rough estimate of the coverage of different

	MeSH	MedDRA	ICD-10	SNOMED CT	UMLS
No. of entries	4,335	18,273	37,263	84,292	100,871
No. of synonyms (incl. concepts)	42,531	57,017	37,263	146,545	243,602

Table 3: A quantitative analysis of the curated dictionaries applied for the disease and side effect named entity recognition. Total number of entries and number of synonyms present within the individual dictionaries are reported.

		Raw			Curated			Disambiguation		
Dictionary	Match type	All	DIS	AE	All	DIS	AE	All	DIS	AE
MeSH	*Complete*	0.54/0.43	0.46	0.40	0.61/0.43	0.46	0.40	0.61/0.43	0.46	0.40
	Partial	0.73/0.58	0.64	0.51	0.80/0.57	0.62	0.51	0.80/0.57	0.62	0.51
MedDRA	*Complete*	0.48/0.62	0.64	0.59	0.57/0.61	0.63	0.59	0.60/0.61	0.62	0.59
	Partial	0.55/0.72	0.76	0.68	0.67/0.72	0.75	0.68	0.69/0.71	0.74	0.68
ICD-10	*Complete*	0.46/0.10	0.10	0.10	0.57/0.15	0.10	0.19	0.57/0.15	0.10	0.19
	Partial	0.59/0.15	0.15	0.14	0.66/0.19	0.14	0.23	0.57/0.19	0.14	0.23
SNOMED CT	*Complete*	0.38/0.18	0.18	0.18	0.40/0.20	0.22	0.18	0.43/0.18	0.20	0.15
	Partial	0.66/0.28	0.33	0.23	0.69/0.34	0.39	0.28	0.71/0.34	0.39	0.28
UMLS	*Complete*	0.18/0.58	0.60	0.55	0.33/0.57	0.60	0.54	0.36/0.57	0.60	0.54
	Partial	0.25/0.73	0.74	0.71	0.43/0.72	0.73	0.71	0.46/0.72	0.73	0.71
Combined	*Complete*	0.12/0.75	0.80	0.70	0.18/0.76	0.81	0.71	0.19/0.76	0.80	0.71
	Partial	0.14/0.92	0.92	0.91	0.21/0.91	0.92	0.89	0.22/0.91	0.92	0.89

Table 4: Comparison of the performance of different dictionaries tested over the evaluation corpus. The results are reported for the *complete matches* and *partial matches* of annotated classes disease (DIS), adverse effect (AE) and a combination of both the classes (All). For a combination of both the classes, i. e. *All*, the precision and recall values are reported. For the classes DIS and AE, only the recall values are reported. 'Combined' indicates the performance achieved by combining the results of all the dictionaries.

dictionaries and the effort that has to be invested to curate them. Table 4 shows the search results obtained with every individual dictionary when complete matches and partial matches were considered. The highest recall for complete matches were achieved by the MedDRA dictionary (62 %) and the UMLS dictionary (58 %). The recall of ICD-10 was the lowest of all dictionaries covering only 10 % of the entities annotated in the corpus. Unlike the other dictionaries, ICD-10 lacks information about the synonyms and term variants which hinders it from covering different types of variants mentioned in the text. The combination of results of all the dictionaries lead to a promising recall of 75 %.

Another important observation is the low recall (18 %) attained by the SNOMED CT dictionary. Although, this dictionary contains over 90,000 entries with 170,561 different terms, its usability for finding entities in the text seems extremely limited. One reason is because of the descriptive nature of most of the terms present in the SNOMED CT vocabulary such as *Spastic paraplegia associated with T-cell lymphotropic virus - 1 infection*. Although such long descriptive terms provide substantial information about the medical condition, they are not quite often used in the literature. Additional reasons are the perception of named entities in annotator's mind as well as the style adopted by the annotation guideline. Perhaps, our principle annotators would annotate such a textual description with *Spastic paraplegia* and *T-cell lymphotropic virus - 1 infection* as two distinct entities rather than annotating the entire phrase as a single entity.

Comparison of the results of complete matches and partial matches in Table 4 shows the granularity of information covered by different data sources and the textual explications. The UMLS and MedDRA achieved an overall recall of 73 % and 72 % respectively for the partial matches whereas the combined results of all the dictionaries achieved a highest recall of 92 %. This provides an indication that the terms contained in these dictionaries cover the head nouns associated with the disease and adverse effect entities but does not include different enumerations used in the literature. For example, in the case of *progressive neurodegenerative disorder*, only *neurodegenerative disorder* was identified whereas the adjective *progressive* was not covered. Based on the experience of the curators and the results from Table 4, nearly 10 % of the mismatches are caused by the medical adjectives such as *chronic*, *acute*, and *idiopathic* that are frequently used in texts but not provided by the resources. Another source of mismatch is the anatomical information often attached to the disease entity in texts. For example, in the case of *vaginal squamous cell carcinoma*, only the *squamous cell carcinoma* was recognized whereas the remaining anatomical substring remained unidentified.

The highest precision rates for the complete matches were achieved by the MeSH dictionary (0.54) and the MedDRA dictionary (0.48) hence validating the curator's opinion about the quality of these resources. The lowest precision of 18 % was achieved by the UMLS dictionary. The precision after combining the results of different dictionaries was considerably low due to the overlapping false positives generated by different dictionaries. The low precision is due to the presence of noisy terms such as *disease* or *response* within the dictionaries. The amount of such noisy terms considerably varies among the different resources with UMLS having the highest. Therefore, the curation of dictionaries is necessary in order to achieve better performance. Experiences from the previously reported dictionary-based named entity approaches let us assume that the precision could be greatly improved by the dictionary curation.

Since the MedDRA dictionary achieved the highest recall, the true positive matches obtained with this dictionary were mapped to the MedDRA level-2 superclasses in order to analyze the distribution of disease and adverse effect terminology over the complete MedDRA hierarchy. The analysis of distribution of annotated entities over the MedDRA subhierarchies is shown in Table 5 and Table 6. From the MedDRA tree distribution of disease or adverse effect matches, it is difficult to understand whether the entity is of kind disease or an adverse event. Here an additional context will be necessary to classify the matches into their respective classes.

MedDRA Superclass	No. of annotated entities
Infections and infestations	110
Psychiatric disorders	83
Neoplasms benign, malignant and unspecified	83
Nervous system disorders	47
Blood and lymphatic system disorders	38

Table 5: Analysis of the top five most frequently occurring disease entities distributed over different MedDRA level-2 superclasses.

MedDRA Superclass	No. of annotated entities
Cardiac disorders	96
Infections and Infestations	93
Injury, poisoning and procedural complications	29
Vascular disorders	23
Gastrointestinal disorders	19

Table 6: Analysis of the top five most frequently occurring adverse effect entities distributed over different MedDRA level-2 superclasses.

5.1. Dictionary Curation

The dictionaries were processed and filtered based on a subset of pre-defined rules in order to reduce the level of ambiguity associated with them. Most of the rules were adapted from Hanisch et al. (2005) and Aronson (1999). The rules that were applied for processing the dictionaries are listed below. All the rules were used in common to all the analyzed dictionaries.

Remove very short tokens: Single character alphanumericals that appear as individual synonyms were removed. For example, '5' was mentioned as a synonym of the concept *Death Related to Adverse Event* in the UMLS.

Remove terms containing special characters: Remove all the terms that contain unusual special characters such as '@', ':' and '&#'. An examples of such term in SNOMED CT is *Heart anomalies: [bulbus/septum] [patent foramen ovale]* .

Remove underspecifications: Substrings such as *NOS*, *NES* and *not elsewhere classified* were removed away from the terms. Such strings were often encountered at endings of the dictionary terms. An example of such a term from MedDRA is *Congenital limb malformation, NOS*

Remove very long terms: Very long and descriptive terms that contains more than 10 words were removed. An example of such a term found in SNOMED CT is *Pancreas multiple or unspecified site injury without mention of open wound into cavity*. Although such long terms do not appear in the text, filtering them from the dictionary gradually reduces the run time of the process.

Remove unusual brackets: Unusual substrings that often appear within the brackets were removed from the terms. Examples of such terms found in SNOMED CT include *[X]Papulosquamous disorders* and *[D]Trismus*.

Remove noisy terms: The ProMiner with different dictionaries was run over an independent corpus of 100,000 abstracts that were randomly selected from MEDLINE. The 500 most frequently occurring terms matched with the individual dictionaries were manually investigated to remove the most frequently occurring false positives. This process will improve the precision of entity recognition during the subsequent runs.

In addition to dictionary curation, the configuration of the ProMiner system was readjusted to match the possessive terms (e. g. *Alzheimer's disease*) that contain '*'s*' substring at the word endings. After the end of the dictionary processing and filtering, the number of entries and synonyms that remained in the individual dictionaries can be found in Table 3. The MeSH dictionary sustained minimum changes with only 15 entries being removed whereas ICD-10 underwent a large noticeable change. The size of the ICD-10 dictionary was reduced to nearly half of the previously used raw dictionary. The search results obtained with every individual curated dictionary can be found in Table 4.

As the result of dictionary curation, the performance of all the dictionaries improved remarkably well. For the complete matches, the precision of UMLS dictionary raised by 15 % with a drop in recall by just 1 %. Other dictionaries that benefited well from the curation process are ICD-10 and MedDRA with raise in their precision by 11 % and 9 % respectively. SNOMED CT showed only 2 % increase in the precision. The recall of all the dictionaries changed marginally except for ICD-10. Processing the synonyms of ICD-10 increased its recall on adverse effect entities by 9 % with an overall raise in the recall by 5 % for both the annotated classes.

5.2. Acronym Disambiguation

In spite of processing the dictionaries by removing the noisy terms as well as lexical modification of the synonyms, the acronyms present in the dictionaries turned out to be another source of frequent false positives. For example, *ALL* which is an acronym for *Acute Lymphoid Leukemia* generated a considerable noise. Therefore, acronyms present in all the dictionaries that have two to four characters were collected in a separate acronym list. Whenever there is a match between the term in the acronym list and the text tokens, a rule was defined in order to accept or neglect the match. This disambiguation facility is available within the ProMiner system. The acronym disambiguation rule accepts the match based on two criteria and they are:

- The match should be case sensitive.
- The acronym as well as any one of its synonym in the respective dictionary should co-occur anywhere within in the same abstract.

For example, the term *ALL* is associated with 17 synonyms in the MedDRA dictionary. Any case sensitive match between the *ALL* and tokens in the text would be accepted if any one synonym of the *ALL* occurs within the same abstract. The search results obtained with the individual curated dictionaries in addition to the acronym disambiguation can be found in Table 4. Considering the complete matches, the acronym disambiguation raised the precision of MedDRA, SNOMED CT and UMLS dictionaries by 3 % each. The performance of MeSH and ICD-10 remain unaffected indicating the presence of less acronyms within them. There was a marginal decline (less than 2 %) in the recall of the dictionaries after applying the disambiguation rule.

In summary, the experiments demonstrated that the performance of a simple search strategy using individual dictionaries for the identification of diseases or adverse effects is low. However, the precision of the dictionary look-up can be improved with the help of curation as well as rule-based filtering (e. g. the one adopted here for disambiguating the acronyms). When the performance of different dictionaries was compared, the MeSH and the MedDRA showed the highest quality with comparably low false positive rate and low ambiguity. The UMLS and SNOMED CT having the size five times as greater than MedDRA or MeSH reported low precision although there was an improvement after the subsequent curation. Depending on the user-specific needs, the UMLS and MedDRA cover large parts of the elementary disease names but does not include sufficient medical adjectives and anatomical specifications within the terms. Although, a sufficient effort has been invested to curate the SNOMED CT and UMLS, the amount of noise they contain overweighs their performance. The MedDRA and UMLS dictionaries demonstrated a competitive recall but the MedDRA being substantially smaller than UMLS reported comparatively low false positive rate. Finally, a combination of all the dictionaries reported the highest recall indicating the diversity of terms provided by different resources.

6. Conclusions

A survey of the performance of different resources for the identification of diseases and adverse effects in texts was performed. An outcome of the survey upheld the MedDRA as a compatible resource for the text mining needs having its recall competitive to the UMLS meta-thesaurus with considerably fair precision upon processing. The UMLS being the largest resource does not include all the names that are covered by the smaller resources. Hence, the combination of the search results from all the terminologies lead to a high increase in recall. This indicates a need for intelligent ways to integrate and merge the information spread across different resources. The amount of work that needs to be invested to curate very large resources such as the SNOMED CT and UMLS is also shown.

In addition to the performance comparison, the effect of dictionary curation and a limited manual investigation of the noisy terms shows to be effective. A rule-based processing coupled with the dictionary curation can substantially improve the performance of the named entity recognition.

In future, we will investigate more enhanced dictionary curation methods for improving the performance of dictionaries. Nevertheless, the performance of rule-based and machine learning-based approaches for identifying the disease and adverse effect named entities needs to be tested.

7. References

S. R. Ahmad. (2003). Adverse drug event monitoring at the Food and Drug Administration. *Journal of General Internal Medicine*, 18(1), pp. 57–60.

A. R. Aronson. (1999). Filtering the UMLS Metathesaurus for MetaMap. Technical report, National Library of Medicine, MD, USA. Available at http://skr.nlm.nih.gov/papers/references/filtering99.pdf.

A. R. Aronson. (2000). Ambiguity in the UMLS Metathesaurus. Technical report, National Library of Medicine, MD, USA. Available at http://skr.nlm.nih.gov/papers/references/ambiguity00.pdf.

A. R. Aronson. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, pp. 17–21.

A. C. Browne, G. Divita, A. R. Aronson, and A. T. McCray. (2003). UMLS language and vocabulary tools. *Proceedings of the AMIA Symposium*, p. 798.

E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman. (2008). Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association*, 15(1), pp. 87–98.

H. Chun, Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. (2006). Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. *Pacific Symposium on Biocomputing*, pp. 4–15.

A. M. Cohen and W. H. Hersh. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1), pp. 57–71.

M. H. Coletti and H. L. Bleich. (2001). Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, 8(4), pp. 317–323.

R. Cornet. (2009). Definitions and qualifiers in SNOMED CT. *Methods of Information in Medicine*, 48(2), pp. 178–183.

C. A Curino, Y. Jia, B. Lambert, P. M. West, and C. Yu. (2005). Mining officially unrecognized side effects of drugs by combining web search and machine learning. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 365–372.

A. J. Forster, J. Andrade, and C. van Walraven. (2005). Validation of a discharge summary term search method to detect adverse events. *Journal of the American Medical Informatics Association*, 12(2), pp. 200–206.

D. Hanisch, K. Fundel, H. Mevissen, R. Zimmer, and J. Fluck. (2005). Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 Suppl 1, pp. S14.

M. Hauben and A. Bate. (2009). Decision support methods for the detection of adverse events in post-marketing data. *Drug Discovery Today*, 14(7-8), pp. 343–357.

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. (2009). A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, 25(22), pp. 2983–2991.

A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9 Suppl 3, pp. S3.

T. Karopka, J. Fluck, H. Mevissen, and A. Glass. (2006). The Autoimmune Disease Database: a dynamically compiled literature-derived database. *BMC Bioinformatics*, 7, pp. 325.

H. Karsten and H. Suominen. (2009). Mining of clinical and biomedical text and data: editorial of the special issue. *International Journal of Medical Informatics*, 78(12), pp. 786–787.

M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biology*, 9 Suppl 2, pp. S1.

R. Leaman, C. Miller, and G. Gonzalez. (2009). Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark. In *Handbook of the 3rd International Symposium on Languages in Biology and Medicine*.

A. T. McCray, O. Bodenreider, J. D. Malley, and A. C. Browne. (2001). Evaluating umls strings for natural language processing. *AMIA Annual Symposium Proceedings*, pp. 448–452.

G. H. Merrill. (2008). The MedDRA paradox. *AMIA Annual Symposium Proceedings*, pp. 470–474.

A. Neveol, W. Kim, J. W. Wilbur, and Z. Lu. (2009). Exploring two biomedical text genres for disease recognition. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pp. 144–152.

S. Ray and M. Craven. (2001). Representing sentence structure in hidden markov models for information extraction. In *IJCAI'01: Proceedings of the 17th international joint conference on Artificial intelligence*, pp. 1273–1279, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

T. C. Rindflesch and A. R. Aronson. (1994). Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, pp. 240–244.

I. Segura-Bedmar, P. Martinez, and M. Segura-Bedmar. (2008). Drug name recognition and classification in biomedical texts. A case study outlining approaches underpinning automated systems. *Drug Discovery Today*, 13(17-18), pp. 816–823.

L. Smith, L. K. Tanabe, R. J. Ando, C. J. Kuo, I. F. Chung, C.N. Hsu, Y. S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. A. Struble, R. J. Povinelli, A. Vlachos, W. A. Baumgartner, L. Hunter, B. Carpenter, R. T. Tsai, H. J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Mana-Lopez, J. Mata, and W. J. Wilbur. (2008). Overview of BioCreative II gene mention recognition. *Genome Biology*, 9 Suppl 2, pp. S2.

B. H. Stricker and B. M. Psaty. (2004). Detection, verification, and quantification of adverse drug reactions. *British Medical Journals*, 329(7456), pp. 44–47.

X. Wang, G. Hripcsak, M. Markatou, and C. Friedman. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3), pp. 328–337.