An improved corpus of disease mentions in PubMed citations

Rezarta Islamaj Doğan
National Center for Biotechnology Information
8600 Rockville Pike
Bethesda, MD 20894, USA
Rezarta.Islamaj@nih.gov

Zhiyong Lu
National Center for Biotechnology Information
8600 Rockville Pike
Bethesda, MD 20894, USA
Zhiyong.Lu@nih.gov

Abstract

The latest discoveries on diseases and their diagnosis/treatment are mostly disseminated in the form of scientific publications. However, with the rapid growth of the biomedical literature and a high level of variation and ambiguity in disease names, the task of retrieving disease-related articles becomes increasingly challenging using the traditional keyword-based approach. An important first step for any disease-related information extraction task in the biomedical literature is the disease mention recognition task. However, despite the strong interest, there has not been enough work done on disease name identification, perhaps because of the difficulty in obtaining adequate corpora. Towards this aim, we created a large-scale disease corpus consisting of 6900 disease mentions in 793 PubMed citations, derived from an earlier corpus. Our corpus contains rich annotations, was developed by a team of 12 annotators (two people per annotation) and covers all sentences in a PubMed abstract. Disease mentions are categorized into Specific Disease, Disease Class, Composite Mention and Modifier categories. When used as the gold standard data for a state-of-the-art machine-learning approach, significantly higher performance can be found on our corpus than the previous one. Such characteristics make this disease name corpus a valuable resource for mining disease-related information from biomedical text. The NCBI corpus is available for download at http://www.ncbi.nlm.nih.gov/CBBresearch/Fellows/Dogan/disease.html.

1 Introduction

Identification of biomedical entities has been an active area of research in recent years (Rinaldi et al., 2011, Smith et al., 2008, Yeh et al., 2005). Automatic systems, both lexically-based and machine learning-based, have been built to identify medically relevant concepts and/or their relationships. Biomedical entity recognition research covers not only gene/protein mention recognition (Tanabe et al., 2005, Campos et al., 2012), but also other medically relevant concepts such as disease names, chemical/drug names, treatments, procedures etc. Systems capable of achieving high performance on these tasks are highly desirable as entity recognition precedes all other information extraction and text mining tasks.

Disease information is sought very frequently in biomedical search engines. Previous PubMed log usage analysis (Islamaj Dogan et al., 2009) has shown that disease is the most frequent non-bibliographic information requested from PubMed users. Furthermore, disease information was often found to be queried together with Chemical/Drug or Gene/Protein information. Automatic recognition of disease mentions therefore, is essential not only for improving retrieval of relevant documents, but also for extraction of associations between diseases and genes or between diseases and drugs. However, prior research shows that automatic disease recognition is a challenging task due to variations and ambiguities in disease names (Leaman et al., 2009, Chowdhury and Lavelli 2010).

Lexically-based systems of disease name recognition, generally refer to the Unified Medical Language System (UMLS) (Burgun and Bodenreider

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 91–99,
Montréal, Canada, June 8, 2012. ©2012 Association for Computational Linguistics

Table 1 AZDC corpus characteristics

Characteristics of the corpus	
Selected abstracts	793
Sentences	2,783
Sentences with disease mentions	1,757
Total disease mentions	3,224

2008). UMLS is a comprehensive resource of medically relevant concepts and relationships and METAMAP(Aronson and Lang 2010) is an example of a natural language processing (NLP) system that provides reliable mapping of the text of a biomedical document to UMLS concepts and their semantic types.

Machine learning systems, on the other hand, have been employed in order to benefit from the flexibility they allow over the rule-based and other statistical systems. However, machine learning systems are strongly dependent on the data available for their training; therefore a comprehensive corpus of examples representing as many variations as possible of the entity of interest is highly favorable.

To our best knowledge, there is one corpus of disease mentions in MEDLINE citations developed by Leaman et al., 2009. This corpus, AZDC corpus, was inspired by the work of Jimeno et al., 2008 and its overall characteristics are given in Table 1. This corpus has been the study of at least two different groups in building automatic systems for disease name recognition in biomedical literature (Leaman et al., 2009, Chowdhury and Lavelli, 2010). They both reported F-scores around 80% in 10-fold cross-validation experiments.

One common encountered difficulty in this domain is the fact that “disease” as a category has a very loose definition, and covers a wide range of concepts. “Disease” is a broadly-used term that refers to any condition that causes pain, suffering, distress, dysfunction, social problems, and/or death. In UMLS, the “disease” concept is covered by twelve different semantic types as shown in Table 2. The disease definition issue has been discussed extensively in other studies (Neveol et al., 2009, Neveol and Lu 2012).

Disease mentions are also heavily abbreviated in biomedical literature (Yeganova et al., 2010). These abbreviations are not always standard; the same abbreviated form may represent different defining strings in different documents. It is therefore, unclear whether these ambiguities could be resolved by an abbreviation look-up list from UMLS Metathesaurus and other available databases.

In this study, we present our efforts in improving the AZDC corpus by building a richer, broader and more complete disease name corpus. The NCBI corpus reflects a more representative view of what constitutes a disease name as it combines the decisions of twelve annotators. It also provides four different categories of disease mentions. Our work was motivated by the following observations:

- The need of a pool of experts:

The AZDC corpus is the work of one annotator. While in terms of consistency this is generally a good thing, a pool of annotators guarantees a more representative view of the entity to be annotated and an agreement between annotators is preferred for categories with loose definitions such as “disease”. Moreover, this would ensure that there would be fewer missed annotations within the corpus.

- The need of annotating all sentences in a document:

The AZDC corpus has disease mention annotations of selected sentences in a collection of PubMed abstracts. In order to be able to perform higher level text mining tasks that explore relationships between diseases and other types of information such as genes or drugs, the disease name annotation has to include all sentences, as opposed to selected ones.

Our work is also related to other corpus annotation projects in the biomedical domain (Grouin et al., 2011, Tanabe at al., 2005, Thompson et al., 2009, Neveol at al., 2009, Chapman et al., 2012). These studies generally agree on the need of multiple experienced annotators for the project, the need of detailed annotation guidelines, and the need of large scale high-quality annotation corpora. The production of such annotated corpora facilitates the development and evaluation of entity recognition and information extraction systems.

2 Methods

Here we describe the NCBI corpus, and its annotation process. We discuss the annotation guidelines and how they evolved through the process.

2.1 The NCBI disease corpus

The AZDC corpus contains 2,783 sentences chosen from 793 PubMed abstracts. These selected

Table 2 The set of UMLS semantic types that collectively cover concepts of the “disease” category

UMLS sematic types	*Disease name example*
Acquired Abnormality	Hernia, Varicose Veins
Anatomical Abnormality	Bernheim aneurysm, Fistula of thoracic duct
Congenital Abnormality	Oppenheim's Disease, Ataxia Telangiectasia
Cell or Molecular Dysfunction	Uniparental disomy, Intestinal metaplasia
Disease or Syndrome	Acute pancreatitis, Rheumatoid Arthritis
Experimental Model of Disease	Collagen-Induced Arthritis, Jensen Sarcoma
Injury or Poisoning	Contusion and laceration of cerebrum
Mental or Behavioral Dysfunction	Schizophrenia, anxiety disorder, dementia
Neoplastic Process	Colorectal Carcinoma, Burkitt Lymphoma
Pathologic Function	Myocardial degeneration, Adipose Tissue Atrophy
Sign or Symptom	Back Pain, Seizures, Skeletal muscle paralysis
Finding	Abnormal or prolonged bleeding time

sentences were annotated for disease mentions, resulting in 1,202 unique mentions and 3,224 total mentions. The NCBI corpus starts with this original corpus; however, it is expanded to cover all the sentences in all the 793 PubMed abstracts.

2.2 Annotation guidelines

One fundamental problem in corpus annotation is the definition of what constitutes an entity to be tagged. Following the lead of the AZDC annotations, the group of annotators working on the NCBI corpus decided that a textual string would be annotated as a disease mention if it could be mapped to a unique concept in the UMLS Metathesaurus, if it corresponded to at least one of the semantic types listed in Table 2, and if it contained information that would be helpful to physicians and health care professionals.

Annotators were invited to use their common knowledge, use public resources of the National Library of Medicine such as UMLS or PubMed Health, Disease Ontology (Warren et al., 2006) and Wikipedia and consider the viewpoint of an average user trying to find information on diseases.

Initially, a set of 20 randomly chosen PubMed abstracts was used as a practice set for the development of annotation guidelines. After each annotator worked individually on the set, the results were shared and discussed among all annotators. The final annotation guidelines are summarized below and also made available at the corpus download website.

What to annotate?

1. *Annotate all specific disease mentions.*

A textual string referring to a disease name may refer to a Specific Disease, or a Disease Class. Disease mentions that could be described as a family of many specific diseases were annotated with an annotation category called Disease Class. The annotation category Specific Disease was used for those mentions which could be linked to one specific definition that does not include further categorization.

e.g. <Specific Disease> Diastrophic dysplasia </> is an <Disease Class> autosomal recessive disease</> characterized by short stature, very short limbs and joint problems that restrict mobility.

2. *Annotate contiguous text strings*.

A textual string may refer to two or more separate disease mentions. Such mentions are annotated with the Composite Mention category.

e.g. The text phrase “Duchenne and Becker muscular dystrophy” refers to two separate diseases. If this phrase is separated into two strings: “Duchenne” and “Becker muscular dystrophy”, it results in information loss, because the word “Duchenne” on its own is not a disease mention.

3. *Annotate disease mentions that are used as modifiers for other concepts*

A textual string may refer to a disease name, but it may not be a noun phrase and this is better expressed with the Modifier annotation category.

e.g.: Although this mutation was initially detected in four of 33 <Modifier> colorectal cancer </> families analysed from eastern England, more extensive analysis has reduced the frequency to four of 52 English <Modifier> HNPCC </> kindreds analysed.

4. *Annotate duplicate mentions.*

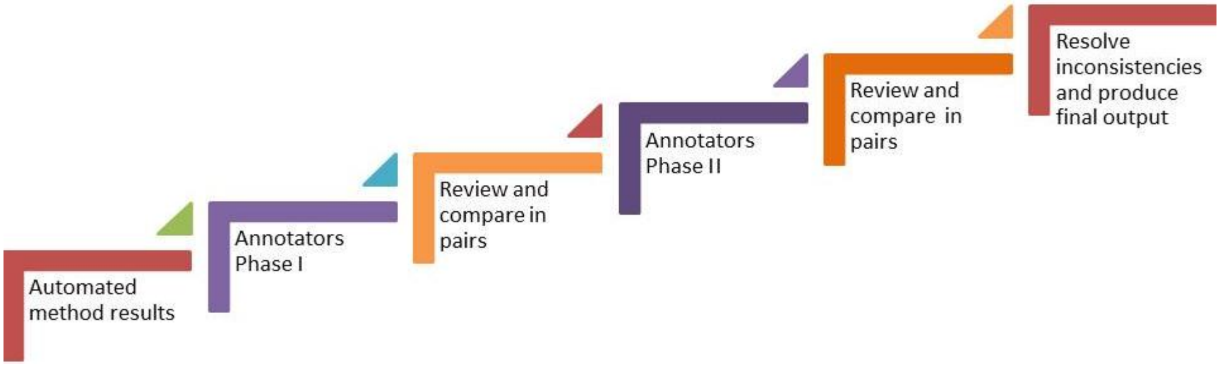

Automated method results
Annotators Phase I
Review and compare in pairs
Annotators Phase II
Review and compare in pairs
Resolve inconsistencies and produce final output

Figure 1. The annotation process

For each sentence in the PubMed abstract and title, the locations of all disease mentions are marked, including duplicates within the same sentence.

5. *Annotate minimum necessary span of text.*

The minimum span of text necessary to include all the tokens expressing the most specific form of the disease is preferred. For example, in case of the phrase "insulin-dependent diabetes mellitus", the disease mention including the whole phrase was preferred over its substrings such as "diabetes mellitus" or "diabetes".

6. *Annotate all synonymous mentions*.

Abbreviation definitions such as "Huntington disease" ("HD") are separated into two annotated mentions.

What not to annotate?

1. *Do not annotate organism names*.

Organism names such as "human" were excluded from the preferred mention. Viruses, bacteria, and other organism names were not annotated unless it was clear from the context that the disease caused by these organisms is discussed.

e.g. Studies of biopsied tissue for the presence of <Specific Disease> Epstein-Barr virus</> and <Specific Disease> cytomegalovirus </> were negative.

2. *Do not annotate gender*.

Tokens such as "male" and "female" were only included if they specifically identified a new form of the disease, for example "male breast cancer".

3. Do not annotate overlapping mentions.

For example, the phrase "von Hippel-Lindau (VHL) disease" was annotated as one single disease mention.

4. *Do not annotate general terms*.

Very general terms such as: disease, syndrome, deficiency, complications, abnormalities, etc. were excluded. However, the terms cancer and tumor were retained.

5. *Do not annotate references to biological processes*.

For example, terms corresponding to biological processes such as "tumorigenesis" or "cancerogenesis".

6. *Do not annotate disease mentions interrupted by nested mentions*.

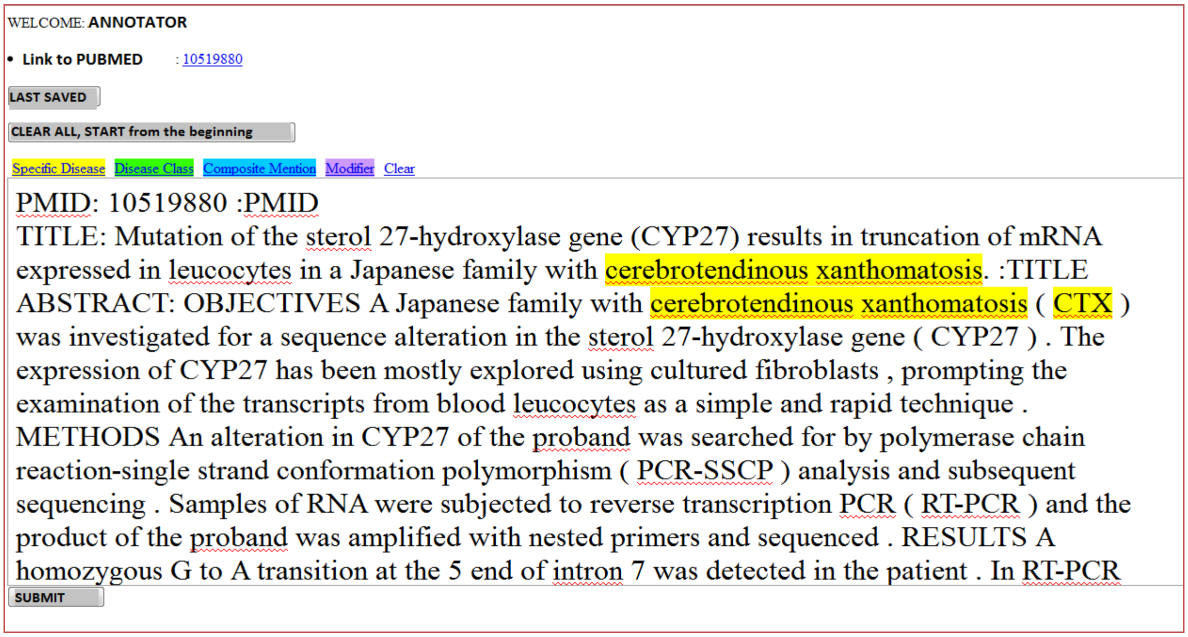
Basically, do not break the contiguous text rule. E.g. WT1 dysfunction is implicated in both neoplastic (Wilms tumor, mesothelioma, leukemia, and breast cancer) and nonneoplastic (glomerulosclerosis) disease.

In this example, the list of all disease mentions includes: "neoplastic disease" and "nonneoplastic disease" in addition to the underlined mentions. However, they were not annotated in our corpus, because other tokens break up the phrase.

2.3 Annotators and the annotation process

The annotator group consisted of 12 people with background in biomedical informatics research and experience in biomedical text corpus annotation. The 793 PubMed citations were divided into sets of 25 PubMed citations each. Every annotator worked on 5 or 6 sets of 25 PubMed abstracts. The sets were divided randomly among annotators. Each set was shared by two people to annotate. To avoid annotator bias, pairs of annotators were chosen randomly for each set of 25 PubMed abstracts.

As illustrated in Figure 1, first, each abstract was pre-annotated using our in-house-developed CRF disease mention recognizer trained on the AZDC corpus. This process involved a 10-fold



WELCOME: ANNOTATOR

- Link to PUBMED : 10519880

PMID: 10519880 :PMID
TITLE: Mutation of the sterol 27-hydroxylase gene (CYP27) results in truncation of mRNA expressed in leucocytes in a Japanese family with cerebrotendinous xanthomatosis. :TITLE
ABSTRACT: OBJECTIVES A Japanese family with cerebrotendinous xanthomatosis (CTX) was investigated for a sequence alteration in the sterol 27-hydroxylase gene (CYP27) . The expression of CYP27 has been mostly explored using cultured fibroblasts , prompting the examination of the transcripts from blood leucocytes as a simple and rapid technique . METHODS An alteration in CYP27 of the proband was searched for by polymerase chain reaction-single strand conformation polymorphism (PCR-SSCP) analysis and subsequent sequencing . Samples of RNA were subjected to reverse transcription PCR (RT-PCR) and the product of the proband was amplified with nested primers and sequenced . RESULTS A homozygous G to A transition at the 5 end of intron 7 was detected in the patient . In RT-PCR

Figure 2. NCBI corpus annotation software. Each annotator selects a PubMed ID from the current working set, and is directed to this screen. Annotation categories are: Specific Disease (highlighted in yellow), Disease Class (green), Composite Mention (blue), or Modifier (purple). To annotate a disease mention in text, annotators highlight the phrase and click on the appropriate label on top of the editor screen. To delete a disease mention, annotators highlight the phrase and click on the Clear label on top of the editor. Annotators can retrieve the last saved version of their annotations for each particular document by clicking on “Last Saved” button. Annotators save their work by clicking on Submit button at the bottom of editor screen.

cross-validation scheme, where all sentences from the same PubMed abstract were assigned to the same split. The learning was performed on 9-folds and then, the PubMed abstracts assigned to the 10th fold were annotated for disease mentions on a sentence-by-sentence basis.

Annotation Phase I consisted of each pre-annotated abstract in the corpus being read and reviewed by two annotators working independently. Annotators could agree with the pre-annotation, remove it, or adjust its text span. Annotators could also add new annotations. After this initial round of annotations, a summary document was created highlighting the agreement and differences between two annotators in the annotations they produced for each abstract. This constituted the end of phase I. The pair of annotators working on the same set at this stage was given the summary document and their own annotations of Phase I.

In annotation Phase II, each annotator examined and edited his or her own annotations by reviewing the different annotations reported in the Phase I summary document. This resulted in a new set of annotations. After this round, a second summary document highlighting the agreement and differences between two annotators was created for each pair of annotators to review.

After phase II, each pair of annotators organized meetings where they reviewed, discussed and resolved their differences. After these meetings, a reconciled set of annotations was produced for each PubMed abstract. The final stage of the annotation process consisted of the first author going over all annotated segments and ensuring that annotations were consistent both in category and in text span across different abstracts and different annotation sets. For example if the phrase “classical galactosemia” was annotated in one abstract as a Specific Disease mention, all occurrences of that phrase throughout the corpus should receive consistent annotation. Identified hard cases were discussed at a meeting where all annotators were present and a final decision was made to reconcile differences. The final corpus is available at: http://www.ncbi.nlm.nih.gov/CBBresearch/Fellows/Dogan/disease.html

2.4 Annotation software

Annotation was done using a web interface (the prototype of PubTator (Wei et al., 2012)), as shown in Figure 2. Each annotator was able to log into the system and work independently. The system allowed flexibility to make annotations in the defined categories, modify annotations, correct the text span, delete as well as go back and review the process as often as needed. At the end of each annotation phase, annotators saved their work, and the annotation results were compared to find agreement and consistency among annotations.

2.5 Annotation evaluation metrics

We measured the annotators' agreement at phase I and II of the annotation process. One way to measure the agreement between two annotators is to measure their observed agreement on the sample of annotated items, as specified in Equation (1).

Agreement statistics are measured for each annotator pair, for each shared annotation set. Then, for each annotator pair the average agreement statistic is computed over all annotation sets shared between the pair of annotators. The final agreement statistic reflects the average and standard deviation computed over all annotator pairs. This is repeated for both phases.

Agreement between two annotators is measured on two levels: one, both annotators tag the same exact phrase based on character indices as a disease mention, and two, both annotators tag the same exact phrase based on character indices as a disease mention of the same category.

2.6 Application of the NCBI corpus

To compare the two disease corpora with regard to their intended primary use in training and testing machine learning algorithms, we performed a 10-fold cross validation experiment with BANNER (Leaman et al, 2009). We evaluated BANNER performance and compared Precision, Recall and F-score values for BANNER when trained and tested on AZDC corpus and the NCBI disease name corpus, respectively. In these experiments, disease mentions of all categories were included and are discussed in the Results section.

To compare the effect of improvement in disease name recognition, the different disease category annotations present in the NCBI corpus were

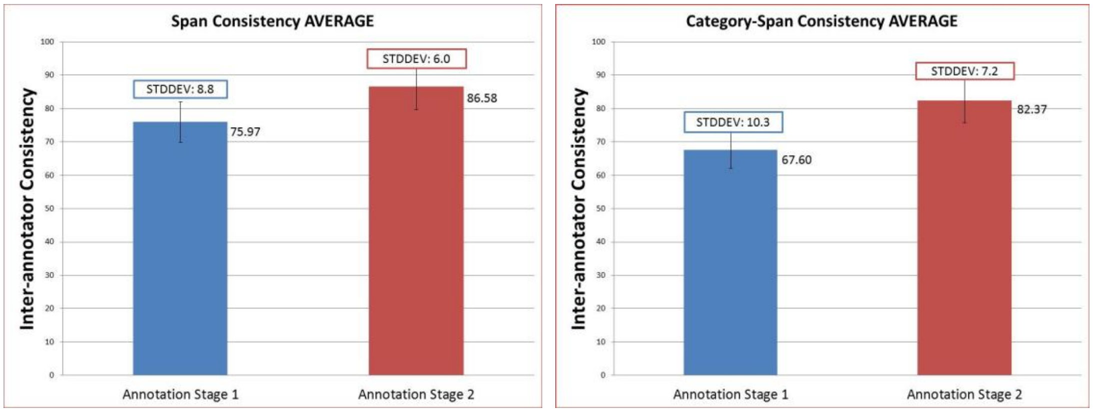


Figure 3 Inter-annotator annotation consistency measured at the span and span-category level

Table 3 The annotation results and corpus characteristics

Characteristics of the corpus	*NCBI corpus*	*AZDC*
Annotators	12	1
Annotated sentences in citation	ALL	Selected
PubMed Citations	793	793
Sentences	6,651	2,784
Sentences with disease annotations	3,752	1,757
Total disease mentions	6,900	3,228
Specific Disease	3,924	-
Disease Class	1029	-
Modifier	1,774	-
Composite Mention	173	-

(1)

$$Consistency = 100 * 2 * \frac{Agreement}{Annotations1 + \mathrm{Annotations2}}$$

flattened into only one single category. This made the NCBI corpus compatible with the AZDC corpus.

3 Results and Discussion

3.1 Results of Inter-Annotator Agreement

Figure 3 shows the inter-annotator agreement results after Phase I and Phase II of the annotations. These statistics show a good agreement between annotators, especially after phase II of annotations. In particular, both span-consistency measure and span-category consistency measure is above 80% after phase II. These values show that our corpus reflects a high quality of annotations and that our two-stage annotation steps are effective in improving corpus consistency.

3.2 Agreement between automatic pre-annotation and final annotation results

In our previous work (Neveol et al, 2009) we have shown that automatic pre-annotation is found helpful by most annotators in assisting large-scale annotation projects with regard to speeding up the annotation time and improving annotation consistency while maintaining the high quality of the final annotations. Thus, we again used pre-annotation in this work. To demonstrate that human annotators were not biased towards the computer-generated pre-annotation, we compared the final annotation with the pre-annotation results. There are a total of 3295 pre-annotated disease mentions: 1750 were found also in the final corpus while the remaining 1545 were either modified or deleted. Furthermore, the final corpus consists of additional 3605 new annotations. Overall, the agreement between pre-annotation and final annotation results is only 35%.

3.3 Statistics of the NCBI disease corpus

After two rounds of annotation, several annotator meetings and resolving of inconsistencies, the NCBI corpus contains 793 fully annotated PubMed citations for disease mentions which are divided into these categories: Specific Disease, Disease Class, Composite Mention and Modifier. As shown in Table 3, the NCBI corpus contains more than 6K sentences, of which more than half contain disease mentions. There are 2,161 unique disease mentions total, which can be divided into these categories: 1,349 unique Specific Disease mentions, 608 unique Disease Class mentions, 121 unique Composite Disease mentions, and 356 unique Modifier disease mentions. The NCBI disease name corpus is available for download and can be used for development of disease name recognition tools, identification of Composite Disease Mentions, Disease Class or Modifier disease mention in biomedical text.

3.4 Characteristics of the NCBI corpus

This annotation task was initially undertaken for purposes of creating a larger, broader and more complete corpus for disease name recognition in biomedical literature.

The NCBI corpus addresses the inconsistencies of missed annotations by using a pool of experts for annotation and creating the annotation environment of multiple discussions and multiple rounds of annotation. The NCBI corpus addresses the problem of recognition of abbreviated disease mentions by delivering annotations for all sentences in the PubMed abstract. Processing all sentences in a document allows for recognition of an abbreviated form of a disease name. An abbreviated term could be tagged for later occurrences within the same document, if an abbreviation definition is recognized in one of the preceding sentences.

NCBI corpus provides a richer level of annotations characterized by four different categories of disease mentions: Specific Disease, Disease Class,

Table 4 NCBI corpus as training, development and testing sets for disease name recognition

Corpus Characteristics	*Training set*	*Development set*	*Test set*
PubMed Citations	593	100	100
Total disease mentions	5148	791	961
Specific Disease	2959	409	556
Disease Class	781	127	121
Modifier	1292	218	264
Composite Mention	116	37	20

Table 5 BANNER evaluation results on AZDC (original) corpus and on the NCBI corpus.

CRF-order	*Corpus*	*Precision*	*Recall*	*F-score*
1	AZDC	0.788	0.743	0.764
1	NCBI	**0.859**	**0.824**	**0.840**
2	AZDC	0.804	0.752	0.776
2	NCBI	**0.857**	**0.820**	**0.838**

Composite Mention and Modifier. Specific Disease mentions could be linked to one specific definition without further categorization, allowing for future normalization tasks. Composite Disease Mentions identify intricate lexical strings that express two or more disease mentions, allowing for future natural language processing tasks to look at them more closely. Modifier disease mentions identify non-noun phrase mentions, again useful for other text mining tasks.

Finally, the corpus can be downloaded and used for development and testing for disease name recognition and other tasks. To facilitate future work, we have divided the corpus into training, development and testing sets as shown in Table 4.

3.5 The NCBI corpus as training data for disease mention recognition

We replicated the BANNER experiments by comparing their cross-validation results on the original corpus (AZDC) and on the NCBI corpus. Our results reveal that BANNER achieves significantly better performance on the NCBI corpus: a 10% increase in F-score from 0.764 to 0.840. Table 5 shows detailed results for BANNER processing in precision, recall and F-score, for both corpora.

In addition, we performed BANNER experiments on the newly divided NCBI corpus with the following results: BANNER achieves an F-score of 0.845 on a 10 fold cross-validation experiment on the NCBI training set, an F-score of 0.819 when tested on the NCBI development set, after trained on the NCBI training set, and an F-score of 0.818 when tested on NCBI test set, after trained on NCBI training set.

3.6 Limitations of this work

The NCBI corpus was annotated manually, thus the tags assigned were judgment calls by human annotators. Annotation guidelines were established prior to the annotation process and they were refined during the annotation process, however grey areas still remained for which no explicit rules were formulated. In particular, inclusion of qualitative terms as part of the disease mention is a matter of further investigation as illustrated by the following example:

- Acute meningococcal pericarditis – Constitutes a disease mention and, exists as a separate concept in UMLS, however
- Acute Neisseria infection – May or may not include the descriptive adjective.

Similarly:

- Classical galactosemia – Includes the descriptive adjective, because it corresponds to a particular form of the disease.
- Inherited spinocerebellar ataxia – May or may not include the descriptive adjective.

Names containing conjunctions are difficult to tag. Although it might seem excessive to require a named entity recognizer to identify the whole expression for cases such as:

- Adenomatous polyps of the colon and rectum,
- Fibroepithelial or epithelial hyperplasias,
- Stage II or stage III colorectal cancer,

The NCBI disease name corpus rectifies this situation by annotating them as Composite Mention disease name category, thus, allowing for future NLP application to develop more precise methods in identifying these expressions.

Moreover, sentences which contained nested disease names require further attention, as the current annotation rule of annotating only contiguous phrases cannot select the outer mentions.

Finally, our current annotation guideline requires that only one of the four categories be assigned to each disease mention. This is not ideal because a disease mention may actually fit more than one category. For instance, a mention can be tagged as both "Modifier" and "Disease Class". In practice, for obtaining consistent annotations, the priority was given in the order of "Modifier", "Composite Mention", "Disease Class", and "Specific Disease" when more than one category deems appropriate. This aspect should be addressed at future work.

4 Conclusions

We have described the NCBI disease name corpus of tagged disease mentions in 793 PubMed titles and abstracts. The corpus was designed to capture disease mentions in the most common sense of the word, and is particularly relevant for biomedical information retrieval tasks that involve diseases. Annotations were performed for all sentences in a document, facilitating the future applications of complex information retrieval tasks connecting diseases to treatments, causes or other types of information. Annotation guidelines were designed with the goal of allowing flexible matching to UMLS concepts, while retaining true meaning of the tagged concept. A more detailed definition on what constitutes a disease name, accompanied with additional annotation rules, could help resolve some existing inconsistencies. The current corpus is reviewed several times by several annotators and describes a refined scale of annotation categories. It allows the separate definition and annotation of Composite mentions, Modifiers and distinguishes between Disease Class mentions versus Specific Diseases. The corpus is available for download[1].

Acknowledgments

Funding: This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

We sincerely thank Robert Leaman and Graciela Gonzalez for their help with BANNER, and the whole team of 12 annotators for their time and expertise on annotation of this corpus.

References

Aronson, A., Lang, F. 2010. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3): 229-236.

Burgun, A., Bodenreider, O. 2008. Accessing and integrating data and knowledge for biomedical research. *Yearb Med Inform*, 91-101.

Campos, D., Matos, S., Lewin, I., Oliveira, J., Rebholz-Schuhmann, D. 2012. Harmonisation of gene/protein annotations: towards a gold standard MEDLINE. *Bioinformatics*, 1;28(9):1253-61

Chapman, W.W., Savova, G.K., Zheng, J., Tharp, M., Crowley, R. 2012. Anaphoric reference in clinical reports: Characteristics of an annotated corpus. *J Biomed Inform*

Chowdhury, F.M., Lavelli, A. 2010. Disease mention recognition with specific features. *BioNLP*, 91-98.

Grouin, C., Rosset. S., Zweigenbaum, P., Fort, K., Galibert, O., Quintard, L. 2011. Proposal for an extension of traditional named entities: From guidelines to evalua-tion, an overview. *5th law workshop*, 92-100.

Islamaj Dogan, R., Murray, G. C., Neveol, A., Lu, Z. 2009. Understanding PubMed user search behavior through log analysis. *Database* (Oxford): bap018.

Jimeno,A., Jimnez-Ruiz, E., Lee, V., Gaudan, S., Berlanga,R., Reholz-Schuhmann, D.2008. Assessment of disease named entity recognition on a corpus of anno-tated sentences. *BMC Bioinformatics*, 9(S-3).

Leaman, R., Miller, C., Gonzalez, G. 2009. Enabling Recognition of Diseases in Biomedical Text with Ma-chine Learning: Corpus and Benchmark. *Symposium on Languages in Biology and Medicine*, 82-89.

Neveol, A., Li, J., Lu, Z. 2012. Linking Multiple Disease-related resources through UMLS. *ACM International Health Informatics*.

Neveol, A., Islamaj Dogan, R., Lu, Z. 2011. Semi-automatic semantic annotation of PubMed Queries: a study on quality, efficiency, satisfaction. *J Biomed Inform*, 44(2):310-8.

Rinaldi, F., Kaljurand, K., Sætre, R. 2011. Terminological resources for text mining over biomedical scientific literature. *Artificial intelligence in medicine* 52(2)

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

Tanabe, L., Xie, N., Thom, L., Matten, W., Wilbur, W.J. 2005. GENETAG: a tagged corpus for gene /protein named entity recognition. *BMC Bioinformatics*, 6:S3.

Thompson, P., Iqbal, S.A., McNaught, J., Ananiadou, S. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10:349.

Warren A., Kibbe J.D.O., Wolf W.A., Smith M.E., Zhu L., Lin S., Chisholm R., Disease Ontology. 2006

Wei C., Kao, H., Lu, Z., 2012. PubTator: A PubMed-like interactive curation system for document triage and literature Curation. *In proceedings of BioCreative* workshop, 145-150.

Yeganova, L., Comeau, D.C., Wilbur, W.J. 2011. Machine learning with naturally labeled data for identifying abbreviation definitions. *BMC Bioinformatics*. S3:S6

Yeh, A., Morgan, A., Colosime, M., Hirschman, L. 2005. BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinformatics*, 6(Suppl 1):S2

[1] http://www.ncbi.nlm.nih.gov/CBBresearch/Fellows/Dogan/disease.html