UNIVERSITY OF CALIFORNIA, SAN DIEGO

Using Text Mining to Accelerate Automatic Curation of Biomedical Databases

A thesis submitted in partial satisfaction of the
requirements for the degree of Masters of Science

in

Computer Science

by

Suvir Jain

Committee in charge:

Professor Chun-Nan Hsu, Chair
Professor Kamalika Chaudhuri
Professor Lawrence Saul

2015

The thesis of Suvir Jain is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2015

DEDICATION

To my family for all their love and support.

EPIGRAPH

The season of failure is the best

time for sowing the seeds of success.

—Paramahansa Yogananda

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Prof. Chun-Nan Hsu. One of my primary goals of undertaking the MS program at UCSD was to work on interesting and challenging research problems related to AI and NLP. Under Prof. Hsu's guidance, I worked on exciting projects and continously pushed the boundaries of my knowledge. His calm demeanor, enthusiasm and patient approach were the perfect ingredients for an environment that promoted both learning and progress. It would be difficult to imagine a better mentor for my MS thesis.

I would like to sincerely thank the other members of our research group: Shitij Bhargava, Kashyap Tumkur, Gordon Lin and Tsung-Ting (Tim) Kuo.

My sincere thanks to other members of my MS committee : Prof. Lawrence Saul and Prof. Kamalika Chaudhuri, for taking out valuable time from their busy schedules to review my thesis.

This work was supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under award number U01HG006894.

Parts of chapters 3, 4, 5 and 6 are currently being prepared for submission in publication of the material. The thesis author was the primary investigator and author of this material except chapter 3. Prof. Chun-Nan Hsu was the primary author of chapter 3.

ABSTRACT OF THE THESIS

Using Text Mining to Accelerate Automatic Curation of Biomedical Databases

by

Suvir Jain

Masters of Science in Computer Science

University of California, San Diego, 2015

Professor Chun-Nan Hsu, Chair

Numerous publicly available biomedical databases derive data by curating from literatures. However, using curated data in Machine Learning is challenging, because the exact mentions and locations in the text are lacking. This thesis describes a general approach to use curated data as training examples for information extraction. The idea is to formulate the problem as cost-sensitive learning from noisy labels, where the cost is estimated by a committee of classifiers that consider both curated data and the text.

Chapter 1

Introduction

1.1 Problem Description

Scientific literature is being produced at an exponential pace. Among various articles published in journals, conferences and workshops, often only they key results of a publication are of interest to other researchers in the same or related fields. To meet this need, there exist structured databases of knowledge that are prepared by domain experts. The standard method of preparing these structured repositories of knowledge involves annotators and curators. They are tasked with reading the entire publication to extract the entities of interest. Over the past few years, text mining has become a useful approach to partially or completely automating this process of extracting entities of interest.

The science behind this use of text mining draws upon the well studied natural language processing (NLP) problem of Name Entity Recognition (NER). Biomedical text mining, the domain of text used in this thesis, has been a promising area for the application of text mining from scientific literature to create and update structured databases of knowledge [MS99]. In theory, well studied NLP methods should be adequate to reduce human effort significantly. However, in practice, text mining often

falls short and manual curation continues to be the standard practice today [WDC+09, DWR+13, ABB+08, HWvM+10].

Other approaches have been considered as well. One such approach involves the use of highly standardized templates [Mon05]. Arguably, a template is an option that constrains the types of knowledge that can be extracted. Updating templates and promoting adoption of standardized templates are other pitfalls. Some argue that crowdsourcing can leverage the power of crowds to perform better than cutting-edge NLP techniques [BDK+14, GS13, SOJN08].

All the approaches have merits and deficiencies. However, if one considers scale of growth and the heterogenous nature of free-text, there is considerable potential for automatic or semi-automatic [BCF+07] NLP-based approaches. Only by nearly eliminating the human element from the process, can we aim to achieve a fully scalable solution to this problem.

1.2 Annotation Vs Curation

Annotation and curation are fundamentally different processes.

An annotation is a label applied to a span of text. Hence, an annotation appears verbatim in the source text. Annotated databases are rather labor-intensive albeit ideal for text mining applications.

On the other hand, curated databases are prepared by domain experts who use a common terminology to describe entities in text. For example, consider the Catalog of Genome Wide Association Studies (GWAS) [WMM+14, HSJ+09], an online database developed by the National Human Genome Research Institute (NHGRI). Curators of the Catalog of Genome Wide Association Study (GWAS) use terms present in the Experiment Factor Ontology (EFO) as curation labels.

Curated databases require in-depth domain knowledge to produce but the result often requires post-processing before being useful for text mining. Creation of annotated databases is a labor-intensive task but the result is more amenable to application of text mining.

1.3 Curated Databases

A large number of biomedical databases are available in the public domain. Many of them contain data derived directly from the published literature either by curation by teams of experts or submitted by authors or researchers. A survey estimated that in 2013, of a total of 290 papers on biomedical databases that were published that also provided open URL links to access the data. Among these 290 databases, 77.59% of them collected data from the literature and contained citations as supportive information [KL14].

1.4 Challenges in Learning from Curated Databases

Curated databases cannot be readily used as training examples because the databases provide no information of where and how the data were derived from the text. The upper panel of Figure 1.1 shows an example entry in the Catalog of GWAS. Each entry represents an observed association reported in an article, specifying that an association between a genetic variant, given in the data field `Strongest SNP`, and a phenotype, given in `Disease/Trait`, was observed from this study from an initial stage sample, given in `Initial Sample Size`. The entry also specifies that the observation was validated with a replication sample, given in `Replication Sample Size`. Other data fields include information of where the genetic variant resides in the genome and statistical strength of the observation.

Field	Value	Field	Value
PubMed ID	21764829	First Author	Png E
Date	7/15/2011	Journal	Hum Mol Genet
Study (title)	A genome-wide association study of hepatitis B vaccine response in an Indonesian population reveals multiple independent risk variants in the HLA region	Disease/Trait	Response to HBV vaccine
Initial Sample Size	1683 Indonesian individuals	Reported Gene(s)	HLA-DR
Replication Sample Size	1931 Indonesian individuals	Mapped Gene(s)	BTNL2 - HLA-DRA
Region	Chr 6:32389648	Strongest SNP	rs3135363
Context	Intergenic	Risk Allele	NR
p-Value	6.53E-22	Risk Allele Frequency	NR
OR or beta-coefficient	1.53	95% CI (text)	[1.35-1.74]
Platform [SNPs passing QC]	Illumina [455,508]	CNV	N

Entity Type	Catelog of GWAS	In the text
Disease / Trait	Response to HBV vaccine	"**hepatitis B vaccine response**"
Initial Sample Size	1683 Indonesian Individuals	1. "We performed a two-stage genome-wide association study (GWAS) of antibody titer in 3614 hepatitis B vaccine recipients from **Indonesia's** Riau Archipelago"
Replication Sample Size	1931 Indonesian Individuals	2. "In the **first stage**, following extensive quality-control (QC) filtering, we analyzed **1683** vaccine recipients…" 3. "After extensive QC filtering, we analyzed genotypes for 1706 SNPs in a second set of **1931** vaccine recipients from the same study. In **this second stage**, we replicated..."

Figure 1.1: Example of an entry in the Catalog of GWAS (upper panel) and after matching to the curated data in the text of the source paper [PTO+11].

The Catalog of GWAS is created and regularly updated by systematically selecting research articles reporting large-scale GWAS and then manually extracting study-level fields of information. The lower panel of Figure 1.1 shows the matching result of the entry in the Catalog of GWAS with the actual passages in the text of the article for three data fields. This example illustrates why curated data can be both useful and not useful as training examples. They are useful because matching the data to the text will create training examples. They are not useful because the matching is not trivial. As shown in Figure 1.1, matching between the data and text requires background knowledge. In fact, curated data rarely provide verbatim copies of what mentioned in the source article. For the purpose of easy-to-search, categorization, summarization, and data integration, curators usually adopt to a standardized terminology different from the text. Also, human curated data inevitably contain typos and inconsistencies in following standards. Even when an exact match with curated data is found, the passage might be about a review of previous results but not where the data should be extracted. To sum up, curated data are

useful but imperfect.

1.5 Cost-sensitive Learning to the Rescue

Machine learning has shown its potential in NLP and has been widely applied in commercial applications. Machine learning algorithms have often won in international challenges on biomedical text mining [KMS+08, ZDFYC07, SDF12, ZPZ+13]. However, supervised statistical learning algorithms require large training examples, which may need an effort no less than creating a manually curated database.

This thesis presents a general approach to using curated data from existing biomedical databases as training examples of NLP. The key idea is to estimate the reliability of the training examples from a *committee* of computer programs, then use a *cost-sensitive* learning algorithm to learn from training examples weighted by the estimated reliability. In Machine Learning, this is known as an approach to agnostic learning from data with noisy labels [LT14, NDRT13, SPI08, FV14, Ser03, KK09, Bou09] and has been intensively studied but, to the best of our knowledge, never been applied to the problem of learning from curated data. The work in this thesis applies this approach to the problem of extracting the target disease/traits of a study from biomedical literature.

The reminder of this thesis is organized as follows. Chapter 2 reviews related work and provides more detailed background. Chapter 3 presents a general framework of the approach to using curated data as training examples. Chapter 4 provides details about the data set and the implementation of the proposed approach used to extract study targets of a GWAS study. Chapters 5 present the results of the various approaches used. Lastly, chapter 6 analyzes the results, summarizes conclusions, and describes potential future work.

Chapter 2

Background and Related Work

2.1 Cost-sensitive Learning

In a typical learning problem, the main goal of the learning step is to minimize the number of incorrect predictions made by the learned model. The "cost" of any mis-prediction is equivalent to the cost of any other mis-prediction. Certain applications require a more nuanced approach to quantifying cost of mis-predictions.

Fraud detection is a good example of one such application. The financial repercussions of a false negative classification i.e. a missed fraud are quite high. Similarly, false positive classification of a fraud detection can potentially result in a customer service nightmare. However, the consequences are not as dire as that of aa missed fraud. In other words, the cost of a false negative is lower than that of a false positive.

In our application domain, we apply cost sensitive learning to identifying the target disease (or trait) of a GWAS study. A GWAS study contains several mentions of various diseases and associated traits. When identifying the target, the cost of missing out a potential disease mention is greater than the cost of incorrectly picking a disease mention as the true target of the study. Hence, the problem lends itself suitably to a cost

sensitive approach.

For more background on cost-sensitive learning, [Elk01] and [ZE01] provide a review of the theoretical foundations. They describe in detail how the cost of misclassification plays an important role in cost-sensitive learning algorithms. [CS98] provide more background on the motivating example of fraud detection described above.

The work described in this thesis applies cost sensitive learning in a problem domain with noisy labels. This problem is also called "Agnostic Active Learning" and [BBL06] states and analyzes one of first algorithms proposed for such a problem. [DMH07] also state an algorithm for this problem.

2.2 Recognizing Disease Mentions

The task of recognizing disease names in free text can be formalized as a *Named entity recognition (NER)* task. Since diseases are often mentioned in biomedical texts, *Biomedical named entity recognition (BNER)* is a more apt formalism.

There are a lot of BNER systems focusing specifically on gene/protein mention recognition such as [THWL09]. [BDS$^+$08] describe a system that extracts disease names using features generally used for gene/protein mention recognition. [AS08] describe a system that can identify drug name mentions.

[JJRL$^+$08] describe some solutions for disease named entity recognition. [LG$^+$08] describe BANNER, a publicly available system specifically for extracting biomedical entities. [CF$^+$10] describes a system based on a feature set tuned specifically for recognizing disease names.

2.3 Entity Normalization

A challenging next step after BNER is to normalize entity mentions in terms of entities from a standardized knowledge source. [KLH11], [Coh05], [HPL+08] and [WTH09] describe systems that normalize gene name mentions. [MLW+08] provide an overview of methods used for the gene normalization task in BioCreative II.

In the context of normalizing disease names, [LIDL13] describe a system that uses a pairwise ranking algorithm to rank pairs of disease mentions and normalized disease terms. [IDL12] describe an inference method for disease name normalization. [KSA+13] describe a rule-based system for disease normalization.

Chapter 3

Cost-sensitive Committee Learning Framework

This section describes a general framework of the approach to using curated data as training examples. The approach aims to be general enough to be applicable for extracting all kinds of entities in free-text. The work described in this thesis specifically applies this approach to extracting target disease/traits in a study. Colleagues of the author have used the same framework to extract other entities such as ethnicity of a sample, size of a sample etc. from a GWA study.

Figure 3.1 shows the five components and the workflow of the whole learning approach. The input is a large corpus of research articles for training. For each article, **Step (A)** identifies the passages that may contain the information to be extracted in the text. The identification of passages should be inclusive in the sense that any suspect passages will be extracted and no passage is missed.

Step (B) pairs each passage with a piece of matched curated data and creates a feature vector for the pair as the input to the committee classifiers. For example, we pair passage 2 in Figure 1.1 (lower panel) to data item "`1683 Indonesian Individuals`"

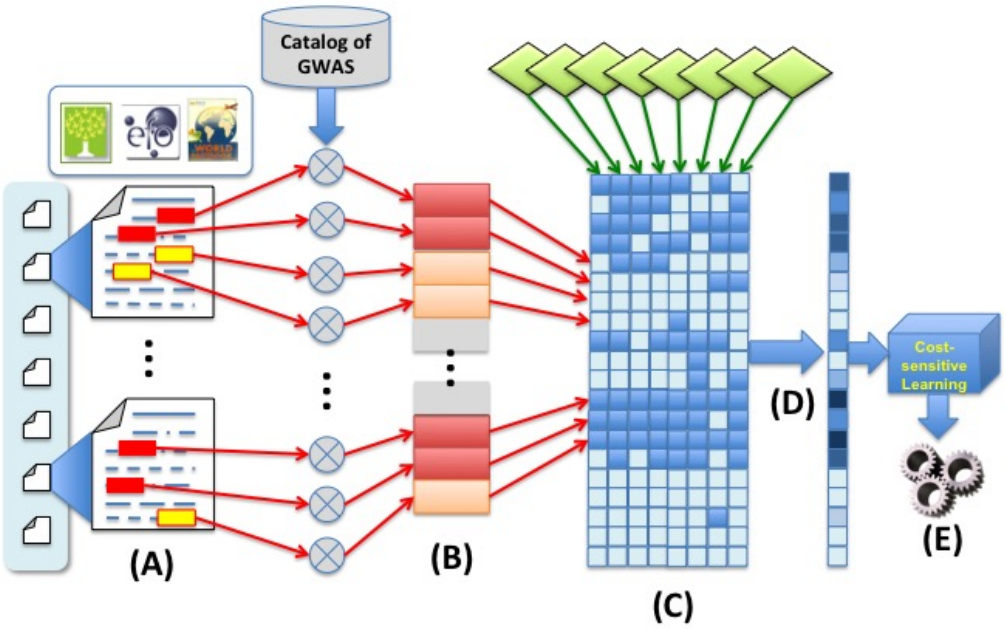

Figure 3.1: System architecture summarizing the steps in the machine learning training process.

from the Catalog of GWAS, because passage 2 is likely where the data item was derived. Again, the matching should be inclusive to contain all potential pairs. Note that although the features are created from one passage, the feature creator may take whatever context in the article where the passage is extracted to create the features. In this way, we can provide the learner to learn from a wide variety of free-text expressions.

Step (C) then sends the feature vectors to a committee of classifiers (diamonds on top of Figure 3.1). Each classifier classifies each pair into positive, if the passage is deemed to contain the information given in the curated data, or negative otherwise. The classifiers can be as "weak" as simple decision rules, like "whether the passage contains a substring that exactly matches the curated data." Therefore, each committee member classifier provides noisy positive-negative labels of the passages extracted from the text. Combining the classification results of all committee members for all extracted passages creates a a large matrix of yes-or-no votes, where each element (i, j) containing the vote from classifier i for candidate passage j.

Step (D) estimates from the matrix the probability that candidate passage j is truly positive by a label estimator that applies an Expectation-Maximization (EM) algorithm to compute maximum likelihood estimation of the probabilities, which can then be treated as the weight, or the reliability of a candidate training example. A similar approach was used in the BioCreative III gene normalization task [ARA+11] to create a silver standard. The EM algorithm works as follows:

1. **Input**: matrix M of committee (column)-passages (row), where each element in the matrix is either positive ($= 1$) or negative ($= 0$);

2. Let p_i be the probability that the i-th passage should be positive, e_j be the error rate of the j-th committee classifier; Let $t = 0$;

3. Initialize $e_j(0) = 0$ for all j;

4. Update $p_i(t) = \frac{\sum(1-e_j(t-1))M_{ij}+k}{J+K}$, where J is the number of the committee, k/K is the Laplace prior;

5. Update $e_j(t) = \frac{\sum p_i(t)M_{ij}+k'}{I+K'}$, where I is the number of the passages, k'/K' is the Laplace prior;

6. $t = t+1$ and repeat update steps until convergent;

7. **Output**: $\hat{p_i}$ and $\hat{e_j}$ be the final values.

With the estimated probability of each candidate passage, we can assign it a *cost*, and train a cost-sensitive learner [YXR14, CZL09, LWWL11] using the candidate passages as the cost-weighted training examples to learn to select correct passages that contain the desired information as **Step (E)**. The cost that we use here is derived according to Lemma 1 in [LT14], where the problem of classification with noisy labels is solved by importance re-weighting. They show that an error bound can be achieved if the misclassification cost of a training example (x,y) is set to $p(y|x)/p_\rho(y|x)$, where ρ denotes sampling from a noise perturbed distribution. Though neither $p(y|x)$ nor $p_\rho(y|x)$ are known, we can approximate $p_i(y =\text{“+”}|x)$ by $\hat{p_i}$ and $p_\rho(y =\text{“+”}|x)$ by $p(\hat{p}(y =\text{“+”}|x) > 0.5)$ for a training example estimated as positive and analogously for a negative one. That is, let $y_i = \text{round}(\hat{p_i})$. If $y_i = 1$ then $c_i = \frac{\hat{p_i}}{\sum_i y_i/I}$, else $c_i = \frac{1-\hat{p_i}}{1-\sum_i y_i/I}$.

We note that this cost-sensitive classifier may use a completely different set of features to characterize a passage.

After all of the learning described above completes, to extract desired data from a given new article, we apply the same **Step (A)** to extract passages and send them to the cost-sensitive classifier to extract data from positive passages.

The content in this chapter in part is currently being prepared for submission for publication of the material. The thesis author was the primary investigator of this material. Prof. Chun-Nan Hsu was the primary author of this chapter.

Chapter 4

Applying the Framework to Extract Study Targets

4.1 Data Set Used

We gather the articles curated and included in the Catalog of GWAS. A genome-wide association study (GWAS) is an approach to detecting genetic variations associated with particular diseases or traits by scanning markers across the genomes of a large-scale sample of subjects in a high-throughput manner. In less than a decade, GWAS studies have been successfully producing discovery and replication of many new disease loci. Discovered genetic associations have led to development of better strategies to diagnose, treat and prevent diseases. The number of GWAS is growing rapidly. There is a need for a database that allows researchers to easily query and search for previous results. A well-curated database also provides a resource for overview investigations and summarization of associated genetic sites and may help suggest pleiotropic genes. Such a database has been created and maintained by the National Human Genome Research Institute (NHGRI), called "A Catalog of Published Genome-Wide Association Studies"

(Catalog of GWAS). The catalog has led to interesting characterization of previous results in GWAS and NHGRI has been continuing to update and curate the catalog regularly by a team of expert curators.

The Catalog of GWAS was first released on November 25, 2008 with 5,120 entries available for search. Since then a large number of new GWAS articles were published and the Catalog has been updated regularly. On a weekly basis, epidemiologists from NHGRI's Office of Population Genomics manually curate information from published GWAS and add them to the catalog. As of May 2015, the catalog of GWAS contains totally 28,870 entries and 1290 unique curated disease/traits.

Work described in this thesis involved 1,272 publications for which a curated target disease or trait is available. Among them, 307 papers have full text available in the NXML format [NCfBI15] and are used as test data. The rest 965 papers serve as training data. Their PDF versions are all transcribed into a XML format using the method proposed in [Hsu14, CPV13].

The curated catalog was provided to us by NHGRI in the form a spreadsheet [EFO14]. The spreadsheet contained the following information.

1. *Curated study target*: Disease or trait chosen by curators as the study target of the publication. This is the study target that all the work in this thesis is trying to predict.

   This may not necessarily match the terms as they appear in the text. To account for this difference, we had to make some changes in the evaluation phase. For the hold-out 307 articles used to evaluate the results, we manually augmented these terms with the corresponding terms that actually appear in the articles to serve as our gold standard.

2. *EFO term*: The Experimental Factor Ontology (EFO) [MHA+10] is an open

access ontology of experimental variables. These variables include disease and their associated traits. This column lists the EFO term chosen by curators as the normalized form of the study target. The study target and EFO term can be identical in some cases but quite different in others. For example, schizophrenia appears both as a study target and the corresponding EFO term. In another example, the normalized EFO term for the study target "Male-pattern baldness" is "androgenetic alopecia".

3. *EFO parent term*: A less granular term from the EFO for the same study target. This term is often very general and would not appear in the text itself. For example, the parent term corresponding to the EFO term "waist-hip ratio" is "body measurement".

4. *PubMed ID*: The identification number assigned to the GWAS publication in PubMed.

5. *Author name*: Name of the primary author of the GWAS publication.

6. *Date of publication*: Date on which the journal was published.

7. *Journal name*: Name of the journal in which the publication was published.

For example, the GWAS publication with PubMed ID 18463370 [MMB+08] was published in the National English Journal of Medicine on 9th May 2008. The primary author was "Maris JM". The curated study target was "Neuroblastoma". The corresponding EFO term and EFO parent term were "neuroblastoma" and "cancer", respectively.

4.2 Extracting Study Targets

The problem of identify target disease or traits of a GWAS study is different from the well-studied problem of disease mention tagging and normalization [LIDL13, LL14, DLL14] in that not all mentions but only the study targets need to be identified and that GWAS study targets include not only diseases but traits like eye colors, response to treatments, sleeping habits, and other phenotypes.

Step (A): Passage Extractor. In this step, we identify all mentions of any disease or trait using exact string matching approach, which is based on a dictionary of all diseases and traits from the search menu of the Web query interface of the Catalog of GWAS [1].

After string matching, we extracted 117,384 mentions in the training and 72,914 in the test data. Note that these numbers are the total mentions of any disease or trait in all articles in the data set (*i.e.*, a paper usually contains multiple mentions).

Step (B): Feature Creator. The following features are generated for each training sample:

- *Token-based features*: Character-level n-grams of the mention.
- *Context-based features*: Word-level and character-level n-grams for up to 10 words before and after the mention.
- *Position-based features*: The location of each mention can be indicated using positional tags (e.g., ⟨article-title⟩ and ⟨abstract⟩) in the converted XML papers; therefore, whether a mention locates within positional tags are extracted as binary features. These tags, however, are not always available from the PDF transcribed XML versions.

For each mention, token-based and context-based features are represented as

[1] https://www.genome.gov/26525384

normalized TF-IDF vectors. Together with position-based features, each mention has approximately 120,000 features.

Step (C): Committee of Classifiers. To build the committee matrix, we design five rule-based binary classifiers as follows.

- *Title or Abstract*: Whether disease / trait mention occurs in the title or abstract of the paper; this is a simple yet strong indicator of a disease mentioned being the actual target of the paper.

- *Exact match*: Whether disease / trait mention exactly matches the target given by human curator.

- *Sub-string match*: Whether disease / trait mention partially matches the target given by the human curator (*e.g.*, a mention of "Diabetes" would be classified as positive, if the human curator determines the disease as "Type-2 Diabetes").

- *Synonym*: Whether disease / trait mention is an exact or partial match of a synonym of the target determined by the human curator. The synonyms are collected from UMLS [Bod04]; for a given disease or trait mention, all UMLS concepts that shared the same `CONCEPT-ID` are considered to be synonymous. To reduce noise, we only keep the synonyms which are in English and are preferred terms (*i.e.*, the `IS-PREF` flag set to `Y` in UMLS).

- *Compound token*: Whether the mention has multiple tokens separated by a space or hyphen (*e.g.*, "Parkinson's disease" would be classified as positive because it consists of compound tokens separated by a space).

It should be noted that although some rule-based classifier is extremely weak (*e.g.*, compound token), our idea is to show that multiple weak classifiers can actually contribute to a strong committee and make accurate predictions.

Step (D): Label Estimator. In this step, we apply the EM method described in Section 3 to label each pair of passage and curated disease or trait with an estimated confidence (*i.e.*, the conditional probability given the pair is positive).

Step (E): Cost-sensitive Learner. In this step, we utilize the estimated confidence generated by the Label Estimator to assign the cost to train a cost-sensitive variant of Support Vector Machine (SVM).

Post-processing. Each paper may contains multiple mentions of various disease and traits. For example, a paper may contain 10 times of "Diabetes", and 30 times of "Hypertension". However, our cost-sensitive classifier may predict only part of them to be positive. This is reasonable because even though two sentences mention the same disease / trait, it is not always the case that both are stating that the disease / trait is the study target. We consider the following two scores for the post-processing, inspired by TF and IDF, respectively:

1. $P_{TF} = \frac{V_{pi}}{V_{pi}+V_{ni}}$, where V_{ni} is the number of negative votes assigned to the i-th candidate. In our example, $P_{TF}(\text{“}Diabetes\text{”}) = 8/10 = 0.8$, and $P_{TF}(\text{“}Hypertension\text{”}) = 12/30 = 0.4$.

2. $P_{IDF} = \frac{V_{pi}}{\sum V_{pi}}$, where V_{pi} is the number of positive votes assigned to the i-th candidate. In our example, $P_{IDF}(\text{“}Diabetes\text{”}) = 8/(8+12) = 0.4$, and $P_{IDF}(\text{“}Hypertension\text{”}) = 12/(8+12) = 0.6$.

To combine P_{TF} and P_{IDF}, we apply two mean computation, namely arithmetic and harmonic, to calculate the final scores and determine our predicted disease / traits. The harmonic mean better represents the mean value of these two metrics. That is because the P_{TF} and P_{IDF} values are often quite small and include outlying values. Harmonic mean is a more sensitive measure in such cases. In our example, the arithmetic mean of "Diabetes" is 0.6, while that of "Hypertension" is 0.5; therefore we predict the disease of

the paper as “Diabetes” using the arithmetic mean method. The results using harmonic mean method can be computed in the similar way.

4.3 EFO Term Identification

The catalog of GWAS also provides a normalized EFO term corresponding to each curated study target, as described in Section 4.1. For each of the two study targets identified by the “harmonic mean” approach in the previous section, we used the string matching algorithms to predict the most likely corresponding EFO term. If any one of these two predictions matched that in the curated gold standard, it was considered a positive result.

We experimented with two string matching techniques as follows.

1. **Edit distance** Levenshtein distance was computed between the two possible study targets and the complete list of all study targets in the GWAS catalog. The EFO term corresponding to the most similar study target was chosen as the possibly correct EFO term.

2. **Gestalt pattern matching approach** The methodology followed was identical to the edit distance approach. The gestalt pattern matching or Ratcliff/Obershelp pattern-matching algorithm ([RM88]) approach tries to recursively look for the longest contiguous matching subsequences. In this approach, it is possible to discover similar strings that do not have the minimal edit distance.

In our experiments, the Gestalt pattern matching approach performed better than the edit distance approach. Results for both these approaches are presented in the results section.

The content in this chapter in part is currently being prepared for submission for

publication of the material. The thesis author was the primary investigator and author of this material.

Chapter 5

Results

In our experiment, we apply our method to predict the top-2 (either using arithmetic or harmonic mean method) target diseases or traits for each article. The reason of choose top-2 is that GWAS publications may focus on one or more diseases / traits. The Catalog of GWAS, however, provides only one target disease / trait for each article. To evaluate our results, we use accuracy based on the top-2 predictions. That is, if either of these two predicted disease / traits matches the human curator's annotation, it is considered a true positive for computing accuracy.

We compare our proposed cost-sensitive learner with cost-insensitive learner. Also, we attempted the following additional alternatives to improve the cost-sensitive learner:

- *Principle Component Analysis (PCA)*: We use PCA method to reduce the dimension of the feature vector to 10,000. We tried various settings of the dimensions and chose the best one.

- *Sparse Random Projection (SRP)*: Similar to PCA, we use another dimension reduction method, sparse random projection, to reduce the dimension of feature vector to 5,000.

- *BIOADI*: We identify and normalize the abbreviations in the input text using the BIOADI system [KLLH09] to pre-process data in an attempt to minimize the chance of missing an abbreviated disease or trait mention.

- *Conditional Random Field (CRF)*: In order to deal with new diseases and traits that do not appear in training dictionary, we also tried to apply CRF in the Passage Extractor step of our method. The design of the features for the CRF is based on the method described in [CF$^+$10]; we use a mixture of general linguistic, orthographic, contextual, syntactic dependency, and dictionary lookup features. By using this CRF model, we discover 59,648 mentions in test data.

We split the available articles into training and test sets as described in Section 4.1 to test all methods.

Table 5.1 shows the performance results, which show that the cost-sensitive learner outperforms the cost insensitive learner, and that harmonic averaging outperforms arithmetic averaging. However, additional alternatives to reduce dimensionality (PCA and SRP) and improve passage extraction (BIOADI and CRF) fail to improve the result of the cost-sensitive learner.

Table 5.2 shows the results for the accuracy of the corresponding EFO term identified. For each of the two study targets identified by the "harmonic mean" approach in Table 5.1, we used the string matching approach described in Section 4.3 to come up with the most likely corresponding EFO term. If any of these predictions matched the EFO term in the curated gold standard (as described in Section 4.1) , it was considered a correct match. The Gestalt pattern matching approach performed better than the Edit Distance approach. Results are presented for both these approaches.

Note that if the study target was identified incorrectly, the EFO term would also be incorrect. Keeping this in mind, these numbers were normalized to reflect the accuracy with respect to those papers in which the study target was correctly identified.

For example, the Gestalt pattern matching approach could identify the correct EFO term in 77.55% of papers that were correctly identified as positives by the cost-insensitive method using Harmonic mean ranking.

Though these simple edit distance metrics can reach near 80% of accuracies for all alternative methods, ideally all positives should match to their correct corresponding EFO terms and there is plenty of room for improvement.

Table 5.1: Accuracy of identifying target disease / trait mention of a GWAS study

Method	Arithmetic	Harmonic
Cost-Insensitive	68.65%	79.62%
Cost-Sensitive	78.05%	**87.46%**
PCA	76.54%	82.73%
SRP	76.22%	84.03%
BIOADI	75.57%	87.29%
CRF	65.79%	75.24%

Table 5.2: Accuracy of EFO term identification

Method	Edit Distance	Gestalt pattern matching
Cost-Insensitive	75.59%	77.55%
Cost-Sensitive	75.26%	78.13%
PCA	78.34%	**81.88**%
SRP	77.51%	77.51%
BIOADI	75.37%	78.73%
CRF	78.35%	80.95%

The content in this chapter in part is currently being prepared for submission for publication of the material. The thesis author was the primary investigator and author of this material.

Chapter 6

Discussion of Results and Conclusion

A large number of curated biomedical databases available in the public domain provides an unprecedented opportunity to train NLP systems to comprehend biomedical publications. This thesis presents an approach to take advantage of this opportunity. The approach applied methods from learning from noisy-label and committee classifiers to assign costs to train cost-sensitive classifiers. This was tested on the problem of extracting target disease/trait in a GWAS publication.

6.1 Discussion of Results

6.1.1 Number of Suggestions Considered

Our proposed method returns up to 2 suggestions because there are a few studies with 2 target disease or traits. We experimented with more than 2 suggestions as well. If we consider five suggestions, the accuracy of at least one of them being correct is higher than 93%. However, we don't think that is a reasonable approach and decided to just consider 2 suggestions. If we return a single suggestion, the accuracy drops by a few percentage points but stays above 82%.

Mentions	Curated study target	Prediction by system	Comments
Schizophrenia is an often devastating neuropsychiatric illness. Understanding the genetic variation affecting response to antipsychotics is important to develop novel diagnostic tests to match individual schizophrenic patients to the most effective and safe medication.	Response to Antipsychotics	Schizophrenia	Pharmacogenomic study
With respect to aging and longevity traits, 149 deaths occurred at a mean age at death of 83 years (range 46 to 99 years) and 713 participants achieved age 65 years or greater.	Aging traits	Age at death	Varying concept granularity
Reports of pedigree studies or twin studies have shown that genetic factors are important in determining serum total immunoglobulin and specific antibody levels in human, with genetic heritability for IgM ranging from 45% to 55%.	IgM levels	Immunoglobulin	Varying concept granularity
Multiple genetic loci associated with obesity or body mass index (BMI) have been identified through genome-wide association studies conducted predominantly in populations of European ancestry.	Body mass index	Obesity	Terminology Dichotomy
Osteoporosis is a common highly heritable skeletal disease characterized by reduced bone mineral density (BMD) and deteriorated bone microstructure, resulting in an increased risk of fracture.	Bone Mineral Density	Osteoporosis	Terminology Dichotomy
However, despite numerous loci and candidate genes linked and associated with atopy-related traits, very few have been associated consistently with total IgE. This study describes the first large-scale, genome-wide scan on total IgE.	Ige levels	Atopy	Terminology Dichotomy
Our analyses identified eight loci demonstrating genome-wide significant association with systolic or diastolic blood pressure, with each locus also providing substantial evidence for association with hypertension.	Diastolic blood pressure	Hypertension	Terminology Dichotomy
Our examination of 22 different common traits in nearly 10,000 participants revealed associations among several single-nucleotide polymorphisms (SNPs, a type of common DNA sequence variation) and freckling, hair curl, asparagus anosmia (the inability to detect certain urinary metabolites produced after eating asparagus), and photic sneeze reflex (the tendency to sneeze when entering bright light).	Common Traits	Sneezing	Several target entities

Figure 6.1: Example of sentences in free text from which the system extracts study targets.

6.1.2 Challenges and Issues

Figure 6.1 shows some examples illustrating the issues that make this problem difficult. These issues are described in more detail below.

- *Pharmacogenomic studies* Papers describing pharmacogenomic studies are another challenging example. For example, one such study's target was "Response to citalopram treatment". Note that the string "Response to citalopram treatment" does not contain any disease or trait mentions. The system discovered various possible study targets such as "schizophrenia", "major depressive disorder", "type 2 diabetes" *etc.* In absence of a more accurate curated label, it is difficult to use a machine learning approach that can handle such examples.

Figure 6.1 provides more examples of sample sentences occurring in the text of such studies.

- *Varying concept granularity* In some cases, the disease or trait mentions in the paper are at a different conceptual granularity from the curator's label. For example, in one of the papers, the system correctly identifies "sleepiness" and "insomnia" as the two study targets. However, the curator labeled "Sleep-related phenotypes" as the study target. A system that could deterministically map traits of varying granularity to a single concept would have helped in such cases.

  Figure 6.1 provides more examples of sample sentences occurring in the text of such studies.

- *Terminology dichotomy* In its current form, the system deals in an identical manner with disease and trait mentions. The curators of GWAS catalog also only label a single column "Disease / trait". However, semantically the two are quite different. Results could have been further improved in the presence of a system that can normalize diseases and their associated traits. For example, "obesity" is associated with the traits "body mass index" and "waist-hip ratio". Studies aim at obesity may measure the traits of the subjects. In some cases, the system correctly discovers "obesity" as the target disease but human curators might have labeled "body mass index" as the target.

  Figure 6.1 provides more examples of sample sentences occurring in the text of such studies.

- *Studies with several target entities to extract.* For example, one of the papers studied 22 associated traits. In such cases, curators choose a broad target disease / traits such "Common traits", "Quantitative traits" or "Selected biomarker traits".

Without more specific curation labels, it is difficult to train a machine learning system from these curated data.

Figure 6.1 provides more examples of sample sentences occurring in the text of such studies.

6.2 Better Curation Guidelines

One of the end goals for text mining would include automated systems that could some day replace human curators. In fact, with the accelerating pace of GWAS study publications, such systems might soon be a necessity. However, during the course of this work, we realized that curation standards need to be revised in order to generate reliable labels that can be used to train machine learning systems. An ideal set of curation guidelines would have to better handle many of the issues discussed in the points above -

- Vague labels
- Labels of different granularity
- Treating diseases and traits separately
- Possibly considering different standard for different types of GWA studies.

Nevertheless, the results show that our approach is effective and outperforms alternative approaches by reaching a F1 score greater than 0.8. We will continue to investigate if it is possible to define standard passage extractors and weak learners applicable to extract commonly interested biomedical entities, attributes and relations to enable rapid development and portability between domains for biomedical literature mining.

6.3 Conclusion and Future Work

Efforts to learn from curated databases often do not perform as well as expected due to the difficulty in learning from curated labels. The work presented in this thesis has shown that it is possible to learn quite accurately from such curated databases. Curated labels are a noisy source of labels for any supervised learner. However, using a committee of simple learners to estimate the true label mitigates this shortcoming to a great extent.

Future work in this direction could potentially involve semi-supervised methods. Semi-supervised methods hold the promise to significantly accelerate and partially automate the curation of databases. Text mining methods that learn both from a curated subset of the database and the uncurated remainder, would be promising avenues for future research endeavors.

The content in this chapter in part is currently being prepared for submission for publication of the material. The thesis author was the primary investigator and author of this material.

Bibliography

[ABB+08] Russ B. Altman, Casey M. Bergman, Judith Blake, Christian Blaschke, Aaron Cohen, Frank Gannon, Les Grivell, Udo Hahn, William Hersh, Lynette Hirschman, Lars Juhl J. Jensen, Martin Krallinger, Barend Mons, Seán I. O'Donoghue, Manuel C. Peitsch, Dietrich Rebholz-Schuhmann, Hagit Shatkay, and Alfonso Valencia. Text mining for biology–the way forward: opinions from leading scientists. *Genome biology*, 9 Suppl 2(Suppl 2):S7+, 2008.

[ARA+11] Cecilia Arighi, Phoebe Roberts, Shashank Agarwal, Sanmitra Bhattacharya, Gianni Cesareni, Andrew C. Aryamontri, Simon Clematide, Pascale Gaudet, Michelle Giglio, Ian Harrow, Eva Huala, Martin Krallinger, Ulf Leser, Donghui Li, Feifan Liu, Zhiyong Lu, Lois Maltais, Naoaki Okazaki, Livia Perfetto, Fabio Rinaldi, Rune Saetre, David Salgado, Padmini Srinivasan, Philippe Thomas, Luca Toldo, Lynette Hirschman, and Cathy Wu. BioCreative III interactive task: an overview. *BMC Bioinformatics*, 12(Suppl 8):S4+, 2011.

[AS08] Pankaj Agarwal and David B Searls. Literature mining in support of drug discovery. *Briefings in bioinformatics*, 9(6):479–492, 2008.

[BBL06] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72. ACM, 2006.

[BCF+07] William A. Baumgartner, K. Bretonnel Cohen, Lynne M. Fox, George Acquaah-Mensah, and Lawrence Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics (Oxford, England)*, 23(13):i41–48, July 2007.

[BDK+14] John D. Burger, Emily Doughty, Ritu Khare, Chih-Hsuan H. Wei, Rajashree Mishra, John Aberdeen, David Tresner-Kirsch, Ben Wellner, Maricel G. Kann, Zhiyong Lu, and Lynette Hirschman. Hybrid curation of gene-mutation relations combining automated extraction and crowdsourc-

ing. *Database : the journal of biological databases and curation*, 2014, 2014.

[BDS+08] Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics*, 9(1):207, 2008.

[Bod04] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[Bou09] Charles Bouveyron. Weakly-Supervised Classification with Mixture Models for Cervical Cancer Detection. In ., pages 1021–1028. ., 2009.

[CF+10] Mahbub Chowdhury, Md Faisal, et al. Disease mention recognition with specific features. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 83–90. Association for Computational Linguistics, 2010.

[Coh05] Aaron M Cohen. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings of the acl-ismb workshop on linking biological literature, ontologies and databases: Mining biological semantics*, pages 17–24. Association for Computational Linguistics, 2005.

[CPV13] Alexandru Constantin, Steve Pettifer, and Andrei Voronkov. PDFX: fully-automated PDF-to-XML conversion of scientific literature. In *Proceedings of the 2013 ACM symposium on Document engineering*, DocEng '13, pages 177–180, New York, NY, USA, 2013. ACM.

[CS98] Philip K Chan and Salvatore J Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *KDD*, volume 1998, pages 164–168, 1998.

[CZL09] Xiao Chang, Qinghua Zheng, and Peng Lin. Cost-sensitive Supported Vector Learning to Rank Imbalanced Data Set. In *Proceedings of the Intelligent Computing 5th International Conference on Emerging Intelligent Computing Technology and Applications*, ICIC'09, pages 305–314, Berlin, Heidelberg, 2009. Springer-Verlag.

[DLL14] Rezarta Islamaj I. Doğan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, February 2014.

[DMH07] Sanjoy Dasgupta, Claire Monteleoni, and Daniel J Hsu. A general agnostic active learning algorithm. In *Advances in neural information processing systems*, pages 353–360, 2007.

[DWR+13] Allan P. Davis, Thomas C. Wiegers, Phoebe M. Roberts, Benjamin L. King, Jean M. Lay, Kelley Lennon-Hopkins, Daniela Sciaky, Robin Johnson, Heather Keating, Nigel Greene, Robert Hernandez, Kevin J. McConnell, Ahmed E. Enayetallah, and Carolyn J. Mattingly. A CTDVPfizer collaboration: manual curation of 88,000 scientific articles text mined for drugVdisease and drugVphenotype interactions. *Database*, 2013:bat080+, January 2013.

[EFO14] Gwas to efo mappings, http://www.ebi.ac.uk/fgpt/gwas/ontology/gwas-efo-mappings201405.xlsx, May 2014.

[Elk01] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Citeseer, 2001.

[FV14] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *Neural Networks and Learning Systems, IEEE Transactions on*, 25(5):845–869, 2014.

[GS13] Benjamin M. Good and Andrew I. Su. Crowdsourcing for bioinformatics. *Bioinformatics*, 29(16):1925–1933, August 2013.

[HPL+08] Jörg Hakenberg, Conrad Plake, Robert Leaman, Michael Schroeder, and Graciela Gonzalez. Inter-species normalization of gene mentions with gnat. *Bioinformatics*, 24(16):i126–i132, 2008.

[HSJ+09] Lucia A. Hindorff, Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, June 2009.

[Hsu14] Chun-Nan Hsu. Accelerating curation of the catalog of gwas by automatic text mining. *Poster presentation in The 64th Annual Meeting of The American Society of Human Genetics (ASHG 2014)*, 2014.

[HWvM+10] Kristina Hettne, Antony Williams, Erik van Mulligen, Jos Kleinjans, Valery Tkachenko, and Jan Kors. Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining. *Journal of Cheminformatics*, 2(1):3+, 2010.

[IDL12] Rezarta Islamaj Dogan and Zhiyong Lu. An inference method for disease name normalization. In *2012 AAAI Fall Symposium Series*, 2012.

[JJRL+08] Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC bioinformatics*, 9(Suppl 3):S3, 2008.

[KK09] Adam Kalai and Varun Kanade. Potential-Based Agnostic Boosting. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 880–888. ., 2009.

[KL14] Yong Zher Koh and Maurice HT Ling. Catalog of biological and biomedical databases published in 2013. *Computational and Mathematical Biology*, 2014.

[KLH11] Cheng-Ju Kuo, Maurice HT Ling, and Chun-Nan Hsu. Soft tagging of overlapping high confidence gene mention variants for cross-species full-text gene normalization. *BMC bioinformatics*, 12(Suppl 8):S6, 2011.

[KLLH09] Cheng-Ju Kuo, Maurice HT Ling, Kuan-Ting Lin, and Chun-Nan Hsu. Bioadi: a machine learning approach to identifying abbreviations and definitions in biological literature. *BMC bioinformatics*, 10(Suppl 15):S7, 2009.

[KMS+08] Martin Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, and Alfonso Valencia. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome biology*, 9 Suppl 2(Suppl 2):1–9, 2008.

[KSA+13] Ning Kang, Bharat Singh, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 20(5):876–881, 2013.

[LG+08] Robert Leaman, Graciela Gonzalez, et al. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 13, pages 652–663. Citeseer, 2008.

[LIDL13] Robert Leaman, Rezarta Islamaj Dogan, and Zhiyong Lu. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics (Oxford, England)*, 29(22):2909–2917, November 2013.

[LL14] Robert Leaman and Zhiyong Lu. *Proceedings of BioNLP 2014*, chapter Automated Disease Normalization with Low Rank Approximations, pages 24–28. Association for Computational Linguistics, 2014.

[LT14] Tongliang Liu and Dacheng Tao. Classification with Noisy Labels by Importance Reweighting, November 2014.

[LWWL11] Hung-Yi Lo, Ju-Chiang Wang, Hsin-Min Wang, and Shou-De Lin. Cost-Sensitive Multi-Label Learning for Audio Tag Annotation and Retrieval. *Multimedia, IEEE Transactions on*, 13(3):518–529, June 2011.

[MHA+10] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010.

[MLW+08] Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, et al. Overview of biocreative ii gene normalization. *Genome biology*, 9(Suppl 2):S3, 2008.

[MMB+08] John M Maris, Yael P Mosse, Jonathan P Bradfield, Cuiping Hou, Stefano Monni, Richard H Scott, Shahab Asgharzadeh, Edward F Attiyeh, Sharon J Diskin, Marci Laudenslager, et al. Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *New England Journal of Medicine*, 358(24):2585–2593, 2008.

[Mon05] Barend Mons. Which gene did you mean? *BMC bioinformatics*, 6(1), June 2005.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition, June 1999.

[NCfBI15] U.S. National Library of Medicine National Center for Biotechnology Information. Pubmed central open access subset, April 2015.

[NDRT13] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1196–1204. Curran Associates, Inc., 2013.

[PTO+11] Eileen Png, Anbupalam Thalamuthu, Rick T. H. Ong, Harm Snippe, Greet J. Boland, and Mark Seielstad. A genome-wide association study of hepatitis B vaccine response in an Indonesian population reveals multiple independent risk variants in the HLA region. *Human Molecular Genetics*, 20(19):3893–3898, October 2011.

[RM88] John W Ratcliff and David E Metzener. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46, 1988.

[SDF12] MatthewS Simpson and Dina Demner-Fushman. Biomedical Text Mining: A Survey of Recent Progress. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 465–517. Springer US, 2012.

[Ser03] Rocco A. Servedio. Smooth Boosting and Learning with Malicious Noise. *J. Mach. Learn. Res.*, 4:633–648, December 2003.

[SOJN08] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[SPI08] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 614–622, New York, NY, USA, 2008. ACM.

[THWL09] Manabu Torii, Zhangzhi Hu, Cathy H Wu, and Hongfang Liu. Biotagger-gm: a gene/protein name recognition system. *Journal of the American Medical Informatics Association*, 16(2):247–255, 2009.

[WDC+09] Thomas C. Wiegers, Allan Peter P. Davis, K. Bretonnel Cohen, Lynette Hirschman, and Carolyn J. Mattingly. Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). *BMC bioinformatics*, 10(1):326+, 2009.

[WMM+14] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(Database issue):D1001–D1006, January 2014.

[WTH09] Joachim Wermter, Katrin Tomanek, and Udo Hahn. High-performance gene name normalization with geno. *Bioinformatics*, 25(6):815–821, 2009.

[YXR14] Ming-Feng Tsai Yu-Xun Ruan, Hsuan-Tien Lin. Improving ranking performance with cost-sensitive ordinal classification via regression. *Information Retrieval*, 2014.

[ZDFYC07] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358–375, September 2007.

[ZE01] Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–213. ACM, 2001.

[ZPZ+13] Fei Zhu, Preecha Patumcharoenpol, Cheng Zhang, Yang Yang, Jonathan Chan, Asawin Meechai, Wanwipa Vongsangnak, and Bairong Shen. Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, 46(2):200–211, April 2013.