UNIVERSITY OF CALIFORNIA, SAN DIEGO

Learning from the Catalog of GWAS to Extract Population Characteristics

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

in

Computer Science

by

Kashyap Ravi Tumkur

Committee in charge:

Professor Chun-Nan Hsu, Chair
Professor Sanjoy Dasgupta
Professor Lawrence K. Saul

2015

The Thesis of Kashyap Ravi Tumkur is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2015

DEDICATION

To my family.

EPIGRAPH

I went to the librarian and asked for a book about stars... and the answer was stunning. It was that the Sun was a star but really close. The stars were suns, but so far away they were just little points of light. The scale of the universe suddenly opened up to me. It was a kind of religious experience. There was a magnificence to it, a grandeur, a scale which has never left me. Never ever left me.

Carl Sagan

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Chun-Nan Hsu for his support as the chair of my committee. His guidance has proved to be invaluable throughout this work.

I would also like to acknowledge my partners in research, Shitij Bhargava, Suvir Jain, Gordon Lin, and Tsung-Ting Kuo, whose input and contributions have made this work possible.

Furthermore, I would like to thank Prof. Sanjoy Dasgupta and Prof. Lawrence K. Saul for graciously giving me their time in reviewing this thesis.

This work was supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under award number U01HG006894 and I am grateful to Lucia Hindorff, the program official at NHGRI for her support.

Material in part from the Introduction and Chapters 1, 3, and 4 is currently being prepared for submission for publication. The thesis author was the primary investigator and author of this material.

VITA

2013	Bachelor of Engineering, National Institute of Engineering
2013–2014	Teaching Assistant, Department of Computer Science and Engineering University of California, San Diego
2014–2015	Graduate Student Researcher, Department of Biomedical Informatics University of California, San Diego
2015	Master of Science, University of California, San Diego

ABSTRACT OF THE THESIS

Learning from the Catalog of GWAS to Extract Population Characteristics

by

Kashyap Ravi Tumkur

Master of Science in Computer Science

University of California, San Diego, 2015

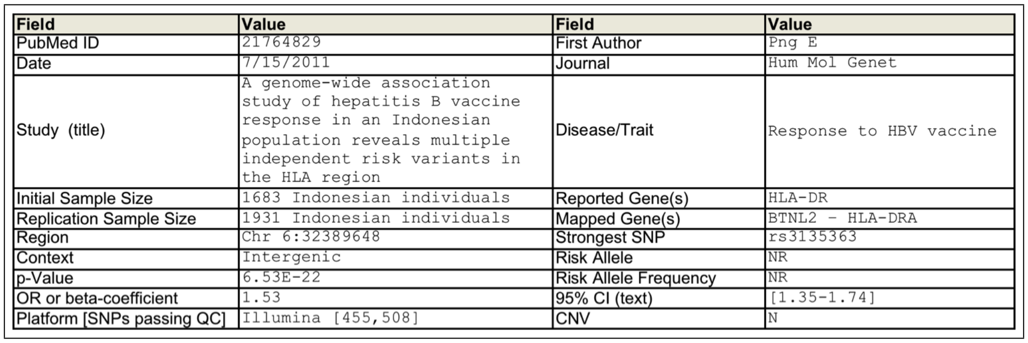
Professor Chun-Nan Hsu, Chair

The Genome-Wide Association Study (GWAS) Catalog is a manually curated, literature-derived collection of all GWAS. This thesis describes a general approach to using this curated data as training examples to extract the characteristics of population samples in GWAS, i.e., the experimental stage, ethnicity groups of the individuals in the populations involved, and the numeric sizes of the sample population pools. As using curated data in Machine Learning for Natural Language Processing is challenging due to the lack of annotations, we formulate the problem as cost-sensitive learning from noisy labels, where the cost is estimated by a committee that considers both curated data and the text. We evaluate this approach on the two distinct problems of extracting sample characteristics as relations of the form $\langle\texttt{stage, ethnicity}\rangle$ and $\langle\texttt{stage, ethnicity, size}\rangle$. We obtain macro F1 scores greater than 0.8 and 0.7 for the two tasks respectively, outperforming similar but cost-insensitive techniques.

Introduction

0.1 Background

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 successfully produced the discovery (of an association in a population) and replication (validation of the discovered association in an independent cohort) of many new disease loci. Such discovered genetic associations have led to development of better strategies to diagnose, treat and prevent diseases. As 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 (genes that are individually responsible for multiple, seemingly unrelated phenotypic traits). Such a database has been created and maintained online by the National Human Genome Research Institute (NHGRI), called A Catalog of Published Genome-Wide Association Studies (Catalog of GWAS) [49]. The catalog has led to interesting characterization of previous results in GWAS [20] and NHGRI has been continuing to update and curate the catalog regularly by a team of expert curators, who enter study-level data into specific fields in the database.



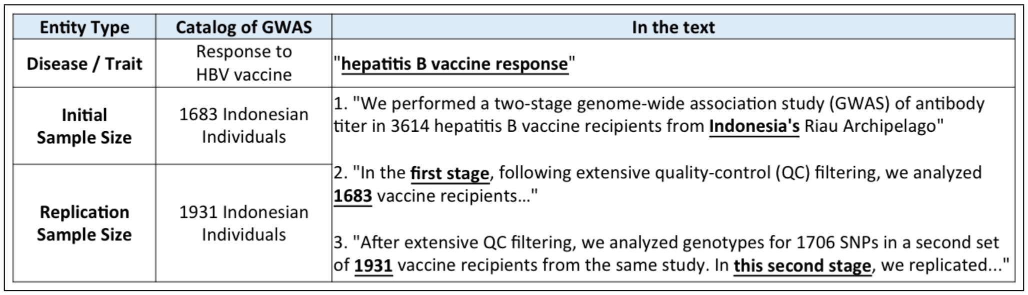
Field	Value	Field	Value
PubMed ID	21764829	First Author	Png E
Date	7/15/2011	Journal	Hum Mol Genet
Study (title)	A genome-wide association study of hepatitis B vaccine response in an Indonesian population reveals multiple independent risk variants in the HLA region	Disease/Trait	Response to HBV vaccine
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

Figure 1. Example of an entry in the Catalog of GWAS.

0.2 The Catalog of GWAS

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 regularly updated by systematically selecting research articles reporting large-scale GWAS. On a weekly basis, epidemiologists from NHGRI's Office of Population Genomics manually curate study-level fields of information from published GWAS and add them to the catalog. As of May 21, 2015, the Catalog of GWAS has been inserted with approximately 29,000 entries extracted from nearly 2,200 distinct articles.

Figure 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 the `Disease/Trait` field, 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`. These latter two data fields describe the characteristics of the population samples used in the GWAS, and are the focus of this thesis. Other data fields also include information of where the genetic variant resides in the genome and statistical strength of the observation.



Entity Type	Catalog 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" 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..."
Replication Sample Size	1931 Indonesian Individuals	

Figure 2. Example of curated data in the Catalog of GWAS entry matched to passages in text of the source article [37].

0.3 The Problem

Our goal is to automate the curation of these study-level data fields from GWAS articles, using a Machine Learning approach to Natural Language Processing. In this thesis, we focus on the characteristics of the sample populations used in the experiments, i.e., the experimental stage ("initial" or "replication"), the ethnicity groups of the individuals involved, and the size of the sample population pool. Collaborative work has been performed for extracting the other data fields towards the larger goal of extracting all the information recorded by the Catalog of GWAS.

Note that the curated data from the catalog cannot readily be used as training examples because it provides no annotations, i.e., information of where and how the data were derived from the text. Instead, it is necessary to automatically match the curated data to potentially corresponding segments of text, or mentions, and use these as training instances for the machine learning approach.

Figure 2 shows the matched 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 2, 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 searching, categorization, summarization, and data integration, curators usually adopt a standardized terminology different from that used in the text. Also, humanly 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 the location from where the data should be extracted. In summary, curated data are useful but imperfect.

0.4 Our Approach

This thesis presents and implements a general approach to using curated data from existing biomedical databases as training examples for 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 [28, 36, 43, 12, 41, 23, 7] and has been intensively studied but, to the best of our knowledge, never been applied to the problem of learning from curated data.

We implement the approach and apply it to two problems of information extraction from the biomedical literature. Task 1 is to extract pairs of stage (initial or replication) and ethnicity background of the study samples from the GWAS articles using data from the Catalog of GWAS as the training examples, and can perform as well as 0.83 in macro F1 score for the extraction task. Task 2 is to extract triplets of the stage ("initial" or "replication"), ethnicity background and sample size from the GWAS articles, also using the curated data from the Catalog of GWAS as the training examples, and performs as well as 0.73 in macro F1 score.

0.5 Applicability

The applicability of this approach is not limited to the Catalog of GWAS. A large number of biomedical databases are available in the public domain, and many contain data derived directly from published literature either through manual curation by teams of experts or structured information submitted by authors or researchers. A survey estimated that, in 2013, a total of 290 papers on biomedical databases were published that also provided open URL links to access the data. Among these 290 databases, 77.59% of them collected data from scientific literature and contained citations as supportive information [25]. Text mining using these curated databases as training data has gained interest in recent years, and we intend for our approach to be generalizable across these databases as well.

0.6 Related Work

Natural Language Processing (NLP) and Text Mining from scientific literature has been considered promising for creating and updating structured databases of biomedical knowledge [31], and as such, there has been significant work in the field that is related to our goal of developing a general approach applicable to different entity types and tasks. The approach is related to learning from data with noisy labels and learning from crowds [39, 43, 50], where crowd inputs are considered noisy, and to the general problem of relation extraction, i.e., identifying relationships between entities that are mentioned in the text, and work from which we have drawn inspiration and ideas are described in the following subsections.

0.6.1 Text Mining from Scientific Literature

Text mining from scientific literature has been considered promising for creating and updating structured databases of biomedical knowledge [31], but it often falls short and currently, manual curation by experts is still the standard practice for these tasks [51, 11, 1, 18]. Some even argue that no text mining or Natural Language Processing (NLP) is necessary when researchers report results following a standardized template [35]. Others argue that crowdsourcing may yield better performance than state-of-the-art NLP solutions [9, 13, 45]. However, given that scientific publications are still written in free text and their number is growing geometrically, a scalable and sustainable approach still requires automatic or semi-automatic approaches [5]. Machine Learning (ML) has exhibited its potential in NLP and been widely applied in commercial applications. ML algorithms have also often achieved success in international challenges on biomedical text mining [26, 55, 44, 54]. However, supervised statistical learning algorithms require large sets of training examples, which may require an effort no less than creating a manually curated database. In this thesis, we explore the idea of training cost-sensitive learners to extract information from free text, by generating noisy labels through automatic and weak annotation of the text from curated data.

0.6.2 Learning from Noisy Labels

In Machine Learning, the problem of agnostic learning from data with noisy labels has been intensively studied [28, 36, 43, 12, 41, 23, 7]. Natarajan et al. [36] provide guarantees for risk minimization under random label noise. Also, [7] deals with weakly supervised models for learning from uncertain data, while [41, 23] apply boosting to learn from noise data by relabeling. Liu et al. [28] propose a learning method to deal with random classification noise by importance re-weighting, which assigns cost to noisy-labeled training examples. Finally, Sheng et al. [43] propose a repeated-labeling strategies of increasing complexity.

However, none of the previously proposed methodologies have been applied to the problem of automatic data curation. In this thesis, we focus on developing a novel agonostic learning framework to specifically deal with the problem of automatic data curation from the Catalog of GWAS.

0.6.3 Relation Extraction

The problem of relation extraction has received much focus in recent years, with several methodologies arising for application to various problems [4]. Relation extraction has been applied to biomedical literature, targeted to problems such as protein-protein interactions using straightforward explorations of all possible relations in text, grouping strategies and graph-based methods [27, 33]. However, real-world applications of learning to extract entities and relations from curated databases share the problem of a lack of annotations, leading to techniques for distant or minimal supervision from curated databases [34, 8, 27, 40], as well as crowdsourced approaches to obtaining annotated sets for learning [3]. The feature space for biomedical named entity extraction using Conditional Random Fields (CRFs), as in our approach, has been explored in [42, 17], and for that for relation extraction using Support Vector Machines (SVMs) has also been explored in [16, 22, 47].

Our problem differs from the above applications in that the sample size (the number of individuals in a population), as a numeric entity, cannot belong to a well-defined vocabulary and, unlike text, is often found in a variety of contexts such as tables and figures in the literature. Further, most approaches to relation extraction assume that the entities exhibiting the relation are present in a single sentence; this is also often not the case for our problem.

0.7 Layout of Thesis

The remainder of this thesis is organized as follows: Chapter 1 presents the general framework of the approach to using curated data as training examples. Chapter 2 describes the data preparation and preprocessing steps required to analyze the data, and Chapters 3 and 4 report the implementations of the approach and the results for the two information extraction problems described above. Finally, Chapter 5 summarizes the results, describes the overall conclusions, and explores future work.

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

Chapter 1

Cost-Sensitive Learning

This chapter describes the general cost-sensitive learning approach used to learn from the curated data. This general approach is developed in conjunction with collaborators and is commonly applied to the task of extracting fields from the Catalog of GWAS.

In this thesis, this approach is implemented and evaluated on the two distinct information extraction tasks of extracting the ethnicity groups of genome-wide association studies' populations, and of extracting relations between the ethnicity groups and the size of the populations involved. Collaborative work not mentioned here include the application of this approach to the tasks of extracting other fields from the Catalog of GWAS, such as diseases and traits that correspond to each study. The combined work leads up to the overarching goal of learning to extract all relevant fields from GWAS articles.

The next section describes this common framework, and the following section describes the extensions made to the framework for the problems described in this thesis.

1.1 Cost-Sensitive Learning Framework

Figure 1.1 shows the five components and the workflow of the overall learning approach. The input is a large corpus of research articles for training. For each article,

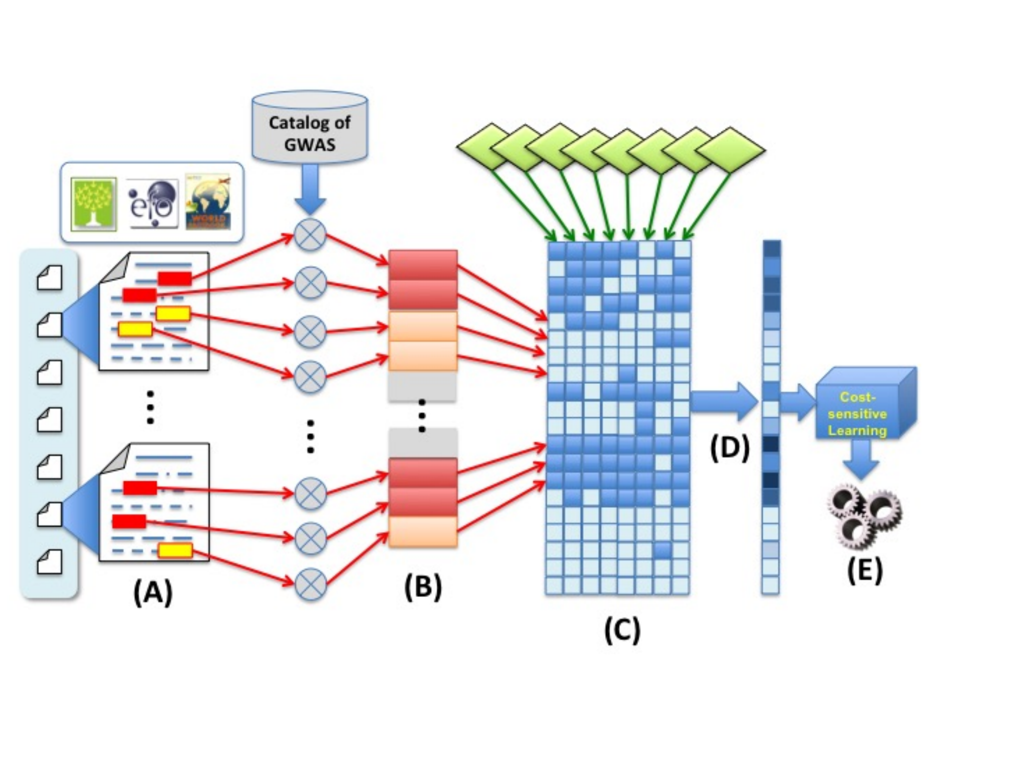


Figure 1.1. System architecture summarizing the steps in the machine learning process.

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 candidate passages will be extracted and no relevant passage is omitted.

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, Passage 2 in Figure 2 is paired with data item `"1683 Indonesian Individuals"` from the Catalog of GWAS, because Passage 2 is possibly the location from where the data item was derived. Again, the matching should be inclusive to contain all such 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 at the top of Figure 1.1). Each classifier classifies each pair as "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, such as "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 the committee members for all the extracted passages creates a a large matrix of yes/no votes, where each element (i, j) containing the vote from classifier i for candidate passage j.

Step (D) estimates from this 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 [2] to create a silver standard. The EM algorithm works as follows:

1. **Input**: matrix M of committees (columns)-passages (rows), 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, and 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, and k/K 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, and k'/K' the Laplace prior;

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

7. **Output**: $\hat{p}_i$ and $\hat{e_j}$, the final values.

With the estimated probability of each candidate passage, we can assign it a *cost*, and train a cost-sensitive learner [52, 10, 29] 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 used here is derived according to Lemma 1 in [28], where the problem of classification with noisy labels is solved by importance reweighting. 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 steps described above complete, 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.

1.2 Extensions to the Framework

The framework described above targets the general problem of classifying instances of entities and relations as positive or negative. However, this is not always sufficient. In extracting ethnicity groups as entities or pairs of ethnicity groups and sample sizes as relations, we would also like to know whether the entities and relations belong to the “initial” or “replication” experimental stages of the GWAS.

This can be formulated as another instance of a classification problem, and the cost-sensitive learning approach can be extended to handle this in a straightforward manner: **Steps (C)** through **(E)** are extended to be two similar cost-sensitive learning stages, each following the general approach described above, with the output of one being the input of the other. More precisely, Stage 1 classifies an input entity or relation as positive or negative, and the instances labeled positive are then forward to the next stage. Stage 2 then classifies these instances into either the initial or replication experimental stages.

This extended framework is used in common for the two tasks described in the following chapters.

Material in part from Chapter 1 is currently being prepared for submission for publication. The thesis author was the primary investigator and author of this material.

Chapter 2

Data Preprocessing

To successfully extract the characteristics of the sample populations in an article, it is necessary to obtain the text data of the article and prepare it for processing and learning. This chapter describes the preprocessing required to extract text from XML, the original format of the articles.

During preprocessing, it is also possible to remove elements of the text that would degrade the performance of the system, and these are also described in the following sections.

2.1 XML to Text Transcription

Articles are originally in XML format, obtained either as publicly-available NXMLs from Pubmed Central, a free full-text archive of biomedical and life sciences literature, or XMLs transcribed from PDFs through our in-house PDF transcription engine.

The full text of the articles is obtained by traversing these XML files and appending all the text elements within XML tags, thus removing all XML formatting while retaining all text content.

2.1.1 XML Tag-based Removal

When traversing the XML files, it is possible to remove the text within certain tags entirely, when it is known that the content of these tags are irrelevant to the task of extraction of population characteristics. These tags fall broadly into the following categories:

- *Article metadata:* This includes tags that describe characteristics surrounding the nature of publication of the study, and not the content itself, such as `article-id`, `journal-meta`, `copyright-statement`, `author-notes`.
- *Formatting tags:* This includes various tags that are necessary for specifying the design and formatting of the article, and are also irrelevant to the textual content, such as `fpage`, `lpage`.
- *Irrelevant tags:* This includes tags that do specify content within the article, but can be removed as they mark content that is not relevant to the task at hand, such as `ref-list`, `x-ref`, `graphic`.

A total of 22 such tags are identified and used for removal. These tags are simply ignored during the traversal when extracting text from the article.

2.2 Regular Expression-based Preprocessing

The text obtained from the XMLs can be further parsed to remove elements that are not relevant to the task, and the remaining text can also be cleaned to make extraction simpler. Regular expression-based preprocessing is primarily geared at removing irrelevant or extraneous numbers that might lead to false positives for sample sizes, as well as normalizing the representations of numbers that may lead to false negatives, i.e., the correct values being missed. This is performed as such numbers are prolific in scientific literature. Examples of these include:

- *Removal of commas:* Commas that mark the thousandth, etc. digits of a number are removed. For example, “$12,696$” becomes “12696”.
- *Mathematical or scientific notation:* Numbers that can be inferred to be irrelevant to sample size extraction based on the surrounding context are removed, such as “$p = 8.14 \times 10(-05))$”, “$3\log 10$”.
- *In-line references:* References within the text to other publications or elements within the same article are removed, such as “[10, 21]” and “Fig. 12”.
- *Units and elements:* Descriptions of units or known elements are removed from the text, such as “1,020 SNPs”, “23 mg/L”.

A total of 23 such patterns are parsed and removed or rewritten using regular expressions.

2.3 Tokenization

The text is first tokenized into sentences using NLTK’s [30] implementation of the Punkt Sentence Tokenizer [24, 6]. Each sentence is then tokenized into words and punctuation using NLTK’s implementation of the Treebank Word Tokenizer [6].

2.4 Part-of-Speech Tagging

The tokens obtained from the previous step are then marked with their part-of-speech (POS) tag. We use the NLTK implementation of POS tagger [6], which is trained on the Penn Treebank tag set [32]. A complete list of the POS tags is given in Appendix A.

Chapter 3

Task 1: Identifying Stage and Ethnicity Groups

A GWAS involves study samples drawn from one or more *ethnicity groups*. This chapter is concerned with the problem of extracting these ethnicity groups that the experiment pertains to. There are conventionally two *stages* in the study: initial and replication, and each of these can be associated with several distinct sample populations. We therefore represent this problem as that of extracting tuples of the form $\langle \texttt{stage, ethnicity} \rangle$ from the free text of a GWAS article, with the entities in the tuple corresponding to the attributes of the study sample.

3.1 Data

The articles are selected from the Catalog of GWAS. We selected articles that satisfy the following criteria:

- *Curated data available*: 2,185 PubMed articles were curated with the data available.

- *NXMLs or PDFs available*: We used NXML versions of the articles if they are available through PubMed Central. These versions have high-quality text. Otherwise, we transcribed PDF versions of the remaining articles to text.

- *No missing values*: the characteristics of the samples are available for whichever stage is mentioned in the article, and the curated data contain no blank entries.
- *Ethnicity group not "NR"*: When unable to find a conclusive ethnicity group for the sample, the entries state "NR" ("not reported").
- *Ethnicity mentions in text*: Terms that correspond to ethnicity groups must be available in text (and not inferred from affiliations of authors, for example).
- *Do not contain errors*: The curated data was found to contain errors in the entries for some articles. Those were excluded.

The final dataset consists of 1,311 articles, comprising 2,357 $\langle\texttt{stage, ethnicity}\rangle$ tuples.

The curated data is normalized to remove spelling errors and inconsistent wording primarily to ensure that there is only one top-level term for a given ethnicity entity. For example, ethnicity group entries in the curated data stating "North African / Middle East" or "Middle East / North African" are both considered to correspond to "Middle East / North African", with this choice of the eventual top-level entry being made arbitrarily.

The curation teams at NHGRI and the European Bioinformatics Institute (EBI) provide us an "extraction guideline" (see [19]), which helps us in the design of the selection criteria and data preparation steps.

3.2 Method

We apply the same pipeline given in Figure 1.1, but we employ two committees to extract the tuples:

Step (A) *Passage extraction*: Mentions in the text corresponding to ethnicity entities are tagged and their surrounding passages extracted. These instances are (weakly) labeled according to curated data as positive or negative.

Step (B) *Feature creator*: The ethnicity instances are featurized and made suitable for classification.

Step (C) *Committee of positive/negative classifiers*: A committee of weak learners are exploited to generate noisy labels, for cost-sensitive learner to classify ethnicity instances as positive or negative instances.

Committee of initial/replication classifiers: Ethnicity instances classified as positive are further classified into the initial and replication experimental stages of the GWAS.

For both committees, we perform **Step (D)** *Label estimator* using the EM algorithm, followed by **Step (E)** *Cost-sensitive learner* to predict the ethnicity and stage of the mentions.

Post-processing: Instances of *ethnicity* classified into a particular *stage* are grouped as $\langle$`stage, ethnicity`$\rangle$ and duplicates removed. The performance of this method is evaluated upon this final set of results.

These steps are described in detail below.

Step (A): Passage Extractor. Mentions in text are generally not exact string matches (or even exact synonyms) of ethnicity groups, necessitating a dictionary mapping of mentions in text (e.g., “German”) to the top-level ethnicity entity (e.g., “European”). Mentions in the text that correspond to a top-level ethnicity entity are mapped to their corresponding entities and tagged with the help of a constructed dictionary as described below, followed by passage extraction. Mentions corresponding to a stage are not tagged.

We construct the dictionary of ethnicity mappings as follows. A multitude of terms can refer to the ethnicity of an individual, including the country of origin (*e.g.*, “Germany”), the specific ethnicity group (*e.g.*, “European”), an adjectival for the country (*e.g.*, “German”), a demonym for the country (*e.g.*, “Germans”), and similar sets of terms for cities and other regions. We handle these terms through the conventions:

- *Country/region name*: Not every mention of a region, say, a country, maps to a specific ethnicity term. The NHGRI curation guideline [19] stating that a given set of individuals belong to an ethnicity group only if it is directly stated in the study, or if at least 90% of the population of the region is known to belong to a single ethnicity group, with this knowledge being based on the CIA World Factbook [1].

- *Adjectivals and demonyms*: An extensive list of the adjectivals and demonyms for countries are obtained from Wikipedia [2], and a dictionary is constructed to map the terms to their corresponding countries. These countries are then mapped to the corresponding ethnicity group (or discarded if no mapping exists).

The final dictionary comprises 449 terms that map to 14 top-level ethnicity groups. These terms cover a majority of the mentions in text, and we omit publications that do not contain language that can be matched to this dictionary. A more comprehensive dictionary may include lists of tribes and indigenous peoples of the world.

This dictionary is used to match mentions in text to ethnicity entities through string matching. The tagged instances are extracted along with their corresponding passages, which consist of the 10 words on either side of the entity in the sentence.

For training and testing, these instances are weakly labeled from curated data by checking if, for a given article, the ethnicity group is present in either experimental stage, initial or replication. If so, this is considered a positive instance, and negative otherwise.

Step (B): Feature Creator. The following types of features are generated for each instance:

- *Token-based features*: A set of binary features each of which turn on for a specific

[1] http://www.cia.gov/library/publications/the-world-factbook/

[2] http://en.wikipedia.org/wiki/List_of_adjectival_and _demonymic_forms_for_countries_and_nations

ethnicity entity (*e.g.*, a feature will be 1 for “East Asian”.)

- *Context-based features*: These include normalized term frequency-inverse document frequency (TF-IDF) representations of unigrams and bigrams of 10 words in either direction of the ethnicity mention, as long as the words are within the same sentence. The words are stemmed using the Porter stemmer [38].

- *Position-based features*: These include features like section title (also in TF-IDF form), the distance (normalized) of the ethnicity mention from the start of the article, or from the start of the section.

- *Additional features*: These include features that do not fit into the above categories, such as the number of times the ethnicity entity was observed (tagged) in the same article.

This results in sparse feature vectors of approximately 80,000 dimensions.

Step (C): Committee of Positive/Negative Classifiers. We use the cost-sensitive learning approach described in Chapter 1 to classify instances as positive or negative. The committee members of weak labelers include:

- *Binary classifier*: A Logistic Regression binary classifier trained on the weak labels from curated data. The predictions of this classifier on the training data is used as the values for this committee member.

- *Rule-based classifier*: this classifier predicts a positive example if the features meet any criteria, such as the presence of words that are commonly found in descriptions of a sample (e.g., “stage”, “cohort”). 65 such terms are used in total. A simple approach to this classifier is to simply predict a negative label when none of the rules match. However, this introduces some error that we would like to minimize by imputing more accurate values for the instances where the rule-based classifier cannot predict a positive label confidently. The problem of imputing values to fill in for missing data for EM algorithms is well-studied [14, 48], with various approaches available in the literature. The approach that works for us is to impute the missing values with the output of the first committee member, i.e., the Logistic Regression binary classifier, as in [15].

- *Weak labels from curated data*: the labels obtained by exact-matching the ethnicity to the curated data.

The committee matrix obtained from concatenating the outputs of all the members is used to estimate the cost to be assigned to each training instance, as in previous sections. These costs are used to train a cost-sensitive, L_2-regularized, linear support vector machine classifier to classify instances as positive or negative instances.

Committee of Initial/Replication Classifiers. Training a cost-sensitive classifier to classify positive instances of ethnicity entities into the corresponding stage of a study is performed in a similar fashion to that for the ethnicity instance classifier.

In this case, the positive training instances are now relabeled "initial" or "replication". The committee members are:

- *Binary classifier*: as above, but trained to distinguish "initial" from "replication" instances.

- *Rule-based classifier:* the rule-based classifier is modified to use the presence of stage-specific words to make its prediction (e.g., "discovery" for the initial stage, or "follow-up" or "second stage" for replication). 8 such terms are used. As in the first stage, the missing values are imputed from the corresponding predictions of the binary Logistic Regression classifier.

- *Weak labels from curated data:* as above, but containing classes "initial" and "replication" instead.

The outputs of the members are used to construct the committee matrix and estimate the cost assigned to each training instance, which is then used to train a cost-sensitive, L_2-regularized, linear SVM to classify the test data into the initial or replication stages with the same set of features.

The output of this step is a classification of each positive *ethnicity* instance into a specific *stage* (initial or replication).

Post-Processing. Either stage in a GWAS may have multiple ethnicity groups. Hence, the extraction can possibly result in multiple tuples of the form $\langle$`stage, ethnicity`$\rangle$ for each study. We compile a list of such tuples for each article, with duplicates being discarded.

3.3 Results

We evaluate the performance by comparing them with the $\langle$`stage, ethnicity`$\rangle$ tuples known to correspond for each GWAS article. Further, we also compare the results with the following alternative approaches. The evaluation methodology and metrics are described below.

1. *Baseline*: All ethnicity instances tagged by the dictionary in an article are assigned to both experimental stages, and the results measured.

2. *Cost-insensitive classification*: the framework described above is used in a cost-insensitive fashion by excluding the committees and directly training the classifiers on the weak labels derived from curated data in **Step (A)**. This provides a candidate for comparison to the cost-sensitive approach for evaluating the performance of cost-based learning. The classifiers for each stage in this method are chosen by

grid-searching over various combinations of loss functions (log loss, or hinge loss), regularization (L_2 or elastic net), and regularization parameters, and choosing the model with the best performance as evaluated by 3-fold cross validation over the training data.

3. *Cost-sensitive classification*: the framework described above, including committee classification, is used.

In each of the methods (excluding the baseline), five-fold article-based cross validation (5-fold CV) is performed. The articles in the dataset are randomly shuffled, and each fold of the 5-fold CV utilizes all $\langle$`stage, ethnicity`$\rangle$ tuples belonging to 80% of the articles in the dataset as training data, and the tuples in the remaining 20% of articles as test data.

The results from each fold are then collected to obtain $\langle$`stage, ethnicity`$\rangle$ tuples for all the articles in the dataset. These results are compared against the curated data and the F1 score calculated in the standard way:

- If a $\langle$`stage, ethnicity`$\rangle$ tuple in the result for a specific article is present in curated data for that article, it is considered a true positive (TP); otherwise, it is considered a false positive (FP).
- If a $\langle$`stage, ethnicity`$\rangle$ tuple in the curated data for a specific article does not have a counterpart in the extracted results, it is considered a false negative (FN).

Using this, we calculate the precision, recall and the Macro F1 score for each method on 1,311 articles comprising 2,357 $\langle$`stage, ethnicity`$\rangle$ tuples from approximately 35,000 mentions of ethnicity-related terms. The resultant values are tabulated below.

Table 3.2 presents the macro precision, recall and the F1 score for the methods. The precision and recall are calculated as above, but for each article individually, and then averaged to obtain the macro precision and recall. The harmonic means of these two values for each method are the respective Macro F1 scores.

Table 3.1. Performance of Ethnicity Group Extraction (Micro).

Method	Precision	Recall	Micro F1 Score
Baseline	0.4898	1.0000	0.6576
Cost-insensitive	0.6965	0.7077	0.7020
Cost-sensitive	0.7471	0.7711	**0.7589**

Table 3.2. Performance of Ethnicity Group Extraction (Macro).

Method	Precision	Recall	Macro F1 Score
Baseline	0.5972	1.0000	0.7478
Cost-insensitive	0.7408	0.7943	0.7666
Cost-sensitive	0.7893	0.8757	**0.8302**

3.4 Discussion of Results

The results in Tables 3.1 and 3.2 indicate that the cost-sensitive approach is able to significantly outperform the similar but cost-insensitive approach, which performs only close to a brute-force baseline. Not only is the cost-sensitive approach able to achieve a much higher degree of recall, but the improvement is accompanied by an increase in overall precision as well.

As the recall gets closer to the limit, the results also indicate that further improvements to the method will be gained by focusing not on extracting relevant ethnicity groups, but on eliminating the ones that are irrelevant to the article.

3.5 Challenges

The challenges faced in the task of extracting ethnicity groups of sample populations from the Catalog of GWAS are described below:

- *Entity normalization*: There are various ways of representing the same entity, and it is necessary to normalize these representations to a single representative entity. However, there exist degrees of difficulty with respect to normalization; for example, it is relatively easy to equate “African American” to “African-American”, but much harder to equate the two represntations with “American citizen of African origin”.

- *Studies with several target entities to extract*: Many GWAS in the U.S. use a highly ethnically diversified study sample with, for example, “52% Caucasian, 24% Latino, 11% African, 9% Eastern Asian and 4% Indigenous Americans”, and studies may also divide into more than two stages. How to flexibly identify and deal with these situations is challenging.

- *Varying concept granularity*: Mentions of ethnicity terms might not be correctly tagged in an article as the authors may report ethnicities in specific terms such as names of tribes and indigenous people, etc., which may not map perfectly to a top-level ethnicity group. This introduces ambiguity which can be an issue for ethnicity background identification.

- *Inadequate reporting of ethnicity data*: Often, the importance of a study to a specific population or application only becomes apparent after the article is published. Hence, the text in the article may never refer to the specific ethnicity group that their experimental sample was drawn from, but simply describe it in terms of the city, state or region, or even the hospital that the population was recruited at. This is doubly challenging as it requires an indefinite expansion of the dictionary of ethnicity-related terms, and also a standardized mapping from each such term to a top-level ethnicity group.

Material in part from Chapter 3 is currently being prepared for submission for publication. The thesis author was the primary investigator and author of this material.

Chapter 4

Task 2: Identifying Stage, Ethnicity Groups and Sample Size

This chapter extends upon the information extraction task in the previous chapter to also include the *sample size* in the extracted relations. We represent this problem as that of extracting tuples of the form $\langle$`stage, ethnicity, size`$\rangle$ from the free text of a GWAS article.

4.1 Data

Again, our articles are selected from the Catalog of GWAS. We selected articles that satisfy the following criteria:

- *Curated data available*: 2,185 PubMed articles were curated with the data available.
- *NXMLs or PDFs available*: We used NXML versions of the articles if they are available through PubMed Central. These versions have high-quality text. Otherwise, we transcribed PDF versions of the remaining articles to text.
- *No missing values*: the characteristics of the samples are available for whichever stage is mentioned in the article, and the curated data contain no blank entries.

- *Ethnicity group not "NR"*: When unable to find a conclusive ethnicity group for the sample, the entries state "NR" ("not reported").
- *Ethnicity mentions in text*: Terms that correspond to ethnicity groups must be available in text (and not inferred from affiliations of authors, for example).
- *Sample size mentions in text*: The sample size is present in text as a number (and not inferred from the article's supplementary material, the text as a sum of the number of cases and controls, or families and couples, or multiple sample population pools, or some other description in words.)
- *Sample size mentions in context:* Aside from the value of sample size being present in text, it is also important for the value to be present in a textual context (as opposed to being in a table composed of many numeric values). This is more difficult to measure and described further in **Step (A)**.
- *Do not contain errors*: The curated data was found to contain errors in the entries for some articles, such as the size of the case or control groups being entered as that of the entire pool. Such articles, when found, were excluded.

Of the dataset, only 409 articles, comprising 657 $\langle$`stage, ethnicity, size`$\rangle$ tuples, meet the first 6 basic criteria of having complete data, and these are used for training. The dataset for evaluation is further reduced to 92 articles and 166 tuples for which contextual clues are present that enable the CRF to identify the mentions for sample size, based on analysis of the output of the CRF tagger. This is described in Section 4.2.

As before, the curated data is normalized to remove spelling errors and inconsistent wording primarily to ensure that there is only one top-level term for a given ethnicity entity.

As in Task 1, the extraction guideline provided to us by the curation teams at NHGRI and EBI was followed in designing the selection criteria and data preparation steps.

4.2 Method

We apply the same pipeline given in Figure 1.1, again employing two committees. The major differences between the extraction of $\langle$`stage, ethnicity`$\rangle$ tuples as in Chapter 3 and the extraction of $\langle$`stage, ethnicity, size`$\rangle$ tuples lie in **Step (A)** *Passage extraction*.

Step (A) *Passage extraction*: Mentions in the text potentially corresponding either to ethnicity entities or a sample size are tagged and their surrounding passages extracted. The instances of ethnicity entities are paired with instances of sample size and then (weakly) labeled according to curated data as positive or negative.

Step (B) *Feature creator*: The $\langle$`ethnicity, size`$\rangle$ tuples are featurized and made suitable for classification.

Step (C) *Committee of positive/negative classifiers*: A committee of weak learners are exploited to generate noisy labels, for the cost-sensitive learner to classify ethnicity instances as positive or negative instances.

Committee of initial/replication classifiers: $\langle$`ethnicity, size`$\rangle$ tuples classified as positive are further classified into the initial and replication experimental stages of the GWAS.

For both committees, we perform **Step (D)** *Label estimator* using the EM algorithm, followed by **Step (E)** *Cost-sensitive learner* to predict the ethnicity, size and stage of the mentions.

Post-processing: Tuples of the form $\langle$`ethnicity, size`$\rangle$ classified into a particular *stage* are grouped as $\langle$`stage, ethnicity, size`$\rangle$ and duplicates removed. The performance of this method is evaluated upon this final set of results.

These steps are described in detail below.

Step (A): Passage Extractor. For extracting candidate ethnicity entities, dictionary tagging is performed as in **Step (A)** *Passage Extractor* in Chapter 3. The dictionary comprising 449 ethnicity terms that map to 14 top-level ethnicity groups is also constructed in the same fashion.

The requirement of extracting arbitrarily-valued sample sizes renders the dictionary tagging approach ineffective. A typical GWAS article may contain hundreds of numeric values, and considering each instance of a numeric value a potential candidate for sample size leads to a huge increase in the number of false positives. This indicates the need to target potential sample sizes by syntactic and semantic features of the text, i.e., by making use of clues from the surrounding textual context. We therefore use a Conditional Random Field model to tag instances of numeric values in text that appear to correspond an experimental sample.

Conditional Random Fields have been widely used in relational learning [46], and for Named Entity Recognition (NER) in the Biomedical domain [21, 42, 17]. To effectively make use of the textual context around a potential *size* entity, CRFs require a rich feature set [47]. These features are described below:

- *Orthographic features:* These features seek to model specific syntactic characteristics of the token under consideration, where token is a specific piece of text (a word, for our purposes). This includes binary features such as whether the token is alphanumeric, hyphenated, etc., numeric features such as word length, and textual features such as prefixes and suffixes of the token of fixed character lengths and the stemmed version of the token (eventually represented as binary features as whether a feature value equals a known value).

- *Local knowledge:* It is necessary to consider the surrounding context of a token to correctly classify it as a *size* entity or not. Therefore, a set of features exist that simply indicate the values for all other features for the immediately preceding and following 2 tokens. Bigrams and Trigrams of the tokens are also included.

- *External knowledge:* To take advantage of the semantics of tokens, features must be provided that are based upon some external knowledge of the text. This includes features such as Part-of-Speech (POS) tagging the token, indicating if the token is a known *ethnicity* term, if the word belongs to a lexicon of related words (e.g., "cohort", "individuals", "participants", are all comparable in their semantics and carry the same weight).

This results in a set of 412 feature functions that are used to featurize the tokens for the CRF tagger, described in Appendix C.1. To avoid overfitting, only features that are observed a minimum of 5 times in the training dataset are included. The output of the CRF tagger is a set of tokens in text that are likely to correspond to the sizes of the experimental samples in the GWAS.

The training data for the CRF model is obtained by marking all exact matches of the true sample size (as known from curated data) for an article in the text of that article, and labeling them as positive mentions. This results in over 75,000 lines containing positive mentions over the entire dataset of 409 articles. The CRF model is evaluated on this dataset using 5-fold cross validation.

The performance of the CRF model is shown in Table 4.1. The performance is described using four criteria to handle the "Annotation vs. Curation" problem. This problem refers to the fact that while curated data provides us with information on which entities and relations are to be correctly extracted, it is not sufficient information in itself to decide which mentions of the entities must be labeled as positive instances of that entity, or if the tagged mentions exhibit the relation at hand. While it is often sufficient to weakly label the tagged mentions, this is a particularly insidious problem for extracting numeric entities such as sample size, which often appear outside of any textual context in lists, tables, etc. in the article, which leads to some mentions either never being tagged, or leading to many false positives if the model is tuned for very high recall. This is also discussed further in Section 4.4. The first two metrics used to evaluate the performance of the tagger are:

- *Micro (All):* This is a measure of the precision, recall and F1 score over all the candidate mentions of sample size in the dataset labeled as positive, as compared with the weak labels derived from curated data.
- *Macro (All):* This is a measure of the precision and recall over the candidate mentions in each of the articles in the dataset labeled as positive as compared with the weak labels derived from curated data, which is then used to obtain the F1 score as a harmonic mean of the two figures.

These standard metrics prove to be a weak indicator of the performance of the model as not only is it unlikely that a high score will be achieved on these metrics due to the aforementioned reasons, but it is also unnecessary to attempt to tag all possible mentions in an article. Instead, it is sufficient to check if the sample sizes as known from curated data have been extracted from at least one corresponding mention in the article that is tagged as positive. Thus, the set of distinct values of the mentions tagged in the text of an article is compared with the set of distinct values of sample sizes expected from curated data to calculate the following two metrics. However, it is important to note that this assumes that the mentions tagged as positive with high probability are likely to be the right mentions (in terms of context) and not, say, mentions that appear in a table.

It is for this reason that features based on external knowledge are included in the CRF model as described above.

The two metrics thus obtained are:

- *Micro (Distinct):* This is a measure of the precision, recall and F1 score over the set of distinct values of the candidate mentions of sample size labeled as positive in the entire dataset, as compared with the set of distinct values of sample sizes expected from curated data (while ensuring that no distinct value occurs in multiple articles).

- *Macro (Distinct):* This is a measure of the precision and recall over the set of distinct values of the candidate mentions of sample size labeled as positive in each of the articles in the dataset as compared with the set of distinct values of sample sizes expected from curated data, which is then used to obtain the F1 score as a harmonic mean of the two figures.

Table 4.1. Performance of CRF Model for Sample Size Extraction.

Method	Precision	Recall	F1 Score
Micro (All)	0.6293	0.2879	0.3951
Macro (All)	0.4425	0.2958	0.3545
Micro (Distinct)	0.5263	0.5109	0.5185
Macro (Distinct)	0.4132	0.4634	0.4368

These metrics are used to tune the conditional random field. However, as observed in Table 4.1, not all the values of sample size required are extracted (as measured by recall). As we are primarily interested in high recall at the tagging stage, it is possible to lower the probability threshold (generally 0.5) required for a mention to be tagged as positive, with some loss of precision. Heuristics such as the following were explored to achieve this:

- *Statically lowering threshold:* The probability threshold is statically set to a lower value (in the range (0.0, 0.5)) and all mentions with a probability of being positive exceeding the threshold are labeled positive.

- *Adaptively lowering threshold:* If an article does not have at least k positive mentions, with k being some fixed value (common to the entire dataset) in (1, 10), then the probability threshold is set to the probability of the k^{th} mention (ordered by probability in descending order) in that article.

These techniques were found to increase recall, but were accompanied by a drastic lowering in the precision of the mentions. This not only leads to a large increase in the number of false positives, but also to a combinatorially larger increase in the number of candidate relations as each mention of sample size is paired with each mention of ethnicity group (as described below).

Investigating the causes of this led to numerous articles being found unusable due to the sample size being mentioned in tables without sufficient context and some instances of curated data containing an incorrect value such as the number of cases or controls in the experiment. As this limits the performance of the entire system, we chose 92 articles (comprising 166 relations) in which the true positives were consistently tagged over multiple runs of the model as trained on different splits of the dataset, and this is used to evaluate the relative performance of the cost-sensitive approach against alternatives (which also share the same tagger).

Next, each *ethnicity* entity obtained from the dictionary tagger and *size* entity extracted by the CRF tagger are then paired to obtain all possible pairs of the form $\langle$`ethnicity, size`$\rangle$, as in the standard method in relation extraction [53]. Each such instance of a pair is labeled as described below and forms the basis for the rest of the process.

For training and testing, these instances are weakly labeled from curated data by checking if, for a given article, the ethnicity group is present in either experimental stage, initial or replication, and if the sample size is also present in the same stage. If so, this is considered a positive instance, and negative otherwise.

Step (B): Feature Creator. The following types of features are generated for each instance:

- *Token-based features*: A set of binary features each of which turn on for a specific ethnicity entity (*e.g.*, a feature will be 1 for "East Asian".)
- *Context-based features*: These include normalized term frequency-inverse document frequency (TF-IDF) representations of unigrams and bigrams of 10 words in either direction of the ethnicity mention and the size mention, as long as the words are within the same sentence. The words are stemmed using the Porter stemmer [38].
- *Position-based features*: These include features like section titles (also in TF-IDF form) of the mentions, the distance (normalized) of the ethnicity and size mentions from the start of the article, or from the start of the section, and the width (normalized) of the instance (i.e., number of words between the ethnicity mention and size mention).
- *Additional features*: These include features that do not fit into the above categories, such as the number of times the same pair of ethnicity and size mentions was observed in the same article.

This results in sparse feature vectors of approximately 108,000 dimensions.

Step (C): Committee of Positive/Negative Classifiers. We use the cost-sensitive learning approach described in Chapter 1 to classify instances as positive or negative. The committee members of weak labelers include:

- *Binary classifier*: A Logistic Regression binary classifier is trained on the weak labels from curated data. The predictions of this classifier on the training data is used as the values for this committee member.
- *Rule-based classifier*: this classifier predicts a positive label if the features meet any criteria, such as the presence of words that are commonly found in descriptions of a sample (e.g., “stage”, “cohort”). 65 such terms are used in total. As in the first task, we again impute the missing values from the predictions of the binary Logistic Regression classifier, as per the approach in [15].
- *Weak labels from curated data*: the labels obtained by exact-matching the ethnicity to the curated data.

The committee matrix obtained from concatenating the outputs of all the members is used to estimate the cost to be assigned to each training instance, as in previous sections. These costs are used to train a cost-sensitive, L_2-regularized, linear support vector machine classifier to classify instances as positive or negative instances.

Committee of Initial/Replication Classifiers. Training a cost-sensitive classifier to classify positive instances of ethnicity entities into the corresponding stage of a study is performed in a similar fashion to that for the ethnicity instance classifier.

In this case, the positive training instances are now relabeled “initial” or “replication”. The committee members are:

- *Binary classifier*: as above, but trained to distinguish “initial” from “replication” instances.
- *Rule-based classifier:* the rule-based classifier is modified to use the presence of stage-specific words to make its prediction (e.g., “discovery” for the initial stage,

or “follow-up” or “second stage” for replication). 8 such terms are used. As in the rule-based classifier for the first committee, we also impute the remaining values from the binary classifier.

- *Weak labels from curated data:* as above, but containing classes “initial” and “replication” instead.

The outputs of the members are used to construct the committee matrix and estimate the cost assigned to each training instance, which is then used to train a cost-sensitive, L_2-regularized, linear SVM to classify the test data into the initial or replication stages with the same set of features.

The output of this step is a classification of each positive $\langle$`ethnicity, size`$\rangle$ instance into a specific *stage* (initial or replication).

Post-Processing. Either stage in a GWAS may have multiple ethnicity groups. Hence, the extraction can possibly result in multiple tuples of the form $\langle$`stage, ethnicity, size`$\rangle$ for each study, and for a given value of *size*. We compile a list of such tuples for each article, with duplicates being discarded.

4.3 Results

We evaluate the performance by comparing them with the $\langle$`stage, ethnicity, size`$\rangle$ tuples known to correspond for each GWAS article. Further, we also compare the results with the following alternative approaches. The evaluation methodology and metrics are described below.

1. *Baseline*: All $\langle$`ethnicity, size`$\rangle$ instances tagged by the dictionary in an article are assigned to both experimental stages, and the results measured. Note that as the baseline also depends on the output of the CRF tagger (which may not tag all

positive mentions), the recall need not be 1.0. However, the recall for the baseline does represent the recall limit for all the methods.

2. *Cost-insensitive classification*: the framework described above is used in a cost-insensitive fashion by excluding the committees and directly training the classifiers on the weak labels derived from curated data in **Step (A)**. This provides a candidate for comparison to the cost-sensitive approach for evaluating the performance of cost-based learning. As in Task 1, the classifiers for each stage in this method are chosen by grid-searching over various combinations of loss functions (log loss, or hinge loss), regularization (L_2 or elastic net), and regularization parameters, and choosing the model with the best performance as evaluated by 3-fold cross validation over the training data.

3. *Cost-sensitive classification*: the framework described above, including committee classification, is used.

As before, in each of the methods (excluding the baseline), five-fold article-based cross validation (5-fold CV) is performed over the entire dataset, and measured on the evaluation set. The 409 articles in the dataset are randomly shuffled, and each fold of the 5-fold CV utilizes all ⟨`stage, ethnicity, size`⟩ tuples belonging to 80% of the articles in the dataset as training data, and the tuples in the remaining 20% of articles as test data.

The results from each fold are then collected to obtain ⟨`stage, ethnicity, size`⟩ tuples for all the articles in the dataset. The overall performance is observed on the test results for the evaluation set of 92 articles. These results are compared against the curated data and the F1 score calculated in the standard way:

- If a ⟨`stage, ethnicity, size`⟩ tuple in the result for a specific article is present in curated data for that article, it is considered a true positive (TP); otherwise, it is considered a false positive (FP).

- If a ⟨`stage, ethnicity, size`⟩ tuple in the curated data for a specific article does not have a counterpart in the extracted results, it is considered a false negative (FN).

Using this, we calculate the precision, recall and the F1 score for each method over the 92 articles and 166 relations. The resultant values are tabulated in Table 4.2.

Table 4.3 presents the macro precision, recall and the F1 score for the methods. The precision and recall are calculated as above, but for each article individually, and then averaged to obtain the macro precision and recall. The harmonic means of these two values for each method are the respective Macro F1 scores.

Table 4.2. Performance of Ethnicity Group and Sample Size Extraction (Micro).

Method	Precision	Recall	Micro F1 Score
Baseline	0.1597	0.8795	0.2704
Cost-insensitive	0.5000	0.6928	0.5808
Cost-sensitive	0.5591	0.7410	**0.6373**

Table 4.3. Performance of Ethnicity Group and Sample Size Extraction (Macro).

Method	Precision	Recall	Macro F1 Score
Baseline	0.2598	0.8913	0.4023
Cost-insensitive	0.6678	0.7283	0.6967
Cost-sensitive	0.6915	0.7736	**0.7302**

4.4 Discussion of Results

The results show that a cost-sensitive committee learning approach reliably outperforms a similar, cost-insensitive approach. This holds true even when the additional committee members are simple classifiers that encode real-world domain knowledge and patterns as rules, which can compensate to some extent for the lack of data, as it is not presumable that all patterns are present in the data in significant quantity as to be learned by a model.

This improvement is also reflected at the level of the entire dataset as well as that of the individual article. The performance of the CRF model also indicates that larger improvements can be made by focusing on the task of tagging sample sizes more accurately.

4.5 Challenges

In addition to the challenges faced in extracting ethnicity groups as described in Section 3.5, this task encounters new challenges in the extraction of sample size and the task of relating the entities, as described below:

- *Extraction of numeric entities*: Extraction of numeric entities like sample size of the experiment poses a challenge: exact-matching techniques cannot be used to tag such entities, and there is no conceptual framework or hierarchy within which these numbers exist, as for other classes of entities. Further, orthographic features frequently used in taggers (e.g., token length, number of capital letters in token) are also not effective.

- *Disconnected entities*: Unlike in conventional approaches [33] to extracting such complex relations of the form $\langle \texttt{a, b, ..., z} \rangle$, it is not the case that the relation between every pair of entities in the tuple is also clearly expressed in the text (e.g., the ethnicity group may be alluded to in a different section of the article, with the stage and sample size mentioned in a table), necessitating an approach that can construct the complex relations from an incomplete set of pairs, and take into account the features of each entity and the associations between them in the classification tasks.

- *Composite numeric entities*: Genome-wide association studies often report sample sizes in terms of the cases and controls, or in numbers of families or couples. NHGRI curating guideline require a total, and the human curator is required to infer the sizes of the components and report the sum. Automating this operation is a non-trivial task.

- *Minimal context*: Characteristics of the experimental sample, and especially sample sizes, are often reported in the supplementary material adjoining an article, or in a table. This poses a challenge for relation extraction as there exists no explicit, textual context for pinpointing a certain figure as the relevant sample size, or even if so, the stage that it belongs to.

- *Semantic considerations*: A sample size can refer to the total size of multiple ethnicity groups, e.g., "1638 individuals of East Asian and European origins" for the initial stage. This requires that two tuples, ⟨`initial, East Asian, 1638`⟩ and ⟨`initial, European, 1638`⟩ be extracted to construct the information represented in the text of the article as ⟨`initial, (East Asian, European), 1638`⟩. If a tuple were to be missed or an extraneous one added with the same sample size, the semantics of the whole are diluted. This is not captured by a simple metric such as F1-score.

Material in part from Chapter 4 is currently being prepared for submission for publication. The thesis author was the primary investigator and author of this material.

Chapter 5

Conclusions

The large number of curated biomedical databases available in the public domain provides an unprecedented opportunity to train NLP systems to comprehend biomedical publications. In this thesis, we describe an approach to two such information extraction tasks for The Catalog of Genome-Wide Association Studies (GWAS): extraction of tuples of the form $\langle \texttt{stage, ethnicity} \rangle$ and $\langle \texttt{stage, ethnicity, size} \rangle$, where *stage* refers to the specific experimental stage of the GWAS, *ethnicity* to the ethnic groups of populations involved, and *size* to the size of the population pool. Our approach applies methods from learning from noisy-label and committee classifiers to assign costs to train cost-sensitive classifiers to perform these extraction tasks.

5.1 Overall Conclusions

The results show that our approach is effective and outperforms alternative conventional cost-insensitive approaches by reaching a F1 score greater than 0.8 for extracting relations of the form $\langle \texttt{stage, ethnicity} \rangle$ and 0.7 for relations of the form $\langle \texttt{stage, ethnicity, size} \rangle$. The generality of the approaches also leads us to conclude that they can be used for a variety of applications and specifically to the automated curation of biomedical databases.

5.2 Future Work

- *Extension to multi-class and multi-label classification*: An entity or relation may naturally fall into one of several classes, or even multiple classes (labels). It is necessary to extend the model to handle such multi-class and multi-label classification tasks.

- *Complex models*: Often, it is necessary to tune a model based on a metric that is close to, but not exactly, the loss function that is used to train the model, as in the case of the conditional random field model used. However, to incorporate interrelations between different instances of the dataset, it may be useful to explore refashioning the objective functions of the models, as well as more complex models themselves.

- *Extension to long tuple extraction*: Given arbitrary tuples of the form $\langle$`a, b, ..., z`$\rangle$, it is necessary to consider aspects of tagging, pairing (or otherwise combining) the entities, and the ordering of stages required to extract such tuples from a curated dataset.

- *Using annotation quality*: It would be beneficial to extend the technique to take the estimated quality of annotations, or mentions, into account in the classification tasks. For example, the probability score awarded by the CRF model to a candidate mention for sample size could be linked to that of the overall relation, say, as a simple feature of the relation instance. The problem could be approached by contrasting the value of the contextual features of a mention against that of other mentions in the same article, to weed out false positives and those mentions that are unlikely to be true positives for a given article, taking other mentions into account through means such as voting or ranking at the tagging stage.

Appendix A

List of Part-of-Speech Tags

Table A.1 enumerates the list of POS tags used to tag tokens, created by the Penn Treebank Project [32] and used in the NLTK tagger implementation.

Table A.1. List of Part-of-Speech tags.

Part-of-Speech Tag		Description
1	C	Coordinating conjunction
2	C	Cardinal number
3	D	Determiner
4	E	Existential there
5	F	Foreign word
6	I	Preposition or subordinating conjunction
7	J	Adjective
8	JJ	Adjective, comparative
9	JJ	Adjective, superlative
10	L	List item marker
11	M	Modal
12	N	Noun, singular or mass
13	NN	Noun, plural
14	NN	Proper noun, singular
15	NNP	Proper noun, plural
16	PD	Predeterminer
17	PO	Possessive ending
18	PR	Personal pronoun
19	PRP$	Possessive pronoun
20	R	Adverb
21	RB	Adverb, comparative
22	RB	Adverb, superlative
23	R	Particle
24	SY	Symbol
25	T	to
26	U	Interjection
27	V	Verb, base form
28	VB	Verb, past tense
29	VB	Verb, gerund or present participle
30	VB	Verb, past participle
31	VB	Verb, non-3rd person singular present
32	VB	Verb, 3rd person singular present
33	WD	Wh-determiner
34	W	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WR	Wh-adverb

Appendix B

Feature Set for Ethnicity Group Extraction

Table B.1 describes the feature set used in extracting ethnicity groups of the population samples used in the Catalog of GWAS. The set consists of real-valued TF-IDF vectors, integer-valued features, and Boolean features (represented as 0 or 1). This feature set is common to both stages, and is represented as a sparse matrix of approximately 80,000 dimensions (or columns).

B.1 Feature Scaling

The features are normalized by removing the mean and scaling to unit variance across the values of each feature, or dimension of feature vector.

B.2 Feature Selection

As the feature set is mostly composed of various TF-IDF vectors, truncated Singular Value Decomposition (SVD), or Latent Semantic Analysis, was explored as a feature selection technique to reduce the dimensionality of the feature vectors. However, this did not affect the performance significantly (and in fact degraded performance slightly) and hence the complete feature vectors were retained.

Table B.1. Feature Set for Ethnicity Group Extraction.

Feature		Type
1	Ethnicity mention token(s)	TF-IDF
2	POS tag of mention	TF-IDF
3	Section title of mention	TF-IDF
4	Extracted passage for mention (windowed)	TF-IDF
5	Extracted sentence for mention (full)	TF-IDF
6	Number of similar mentions in article	Integer
7	Position of mention in article	Integer
8	Position of mention in section	Integer
9	Length of sentence that mention appears in	Integer
10	Section title of mention is unknown	Boolean
11	Section title of mention is the beginning of article	Boolean
12	Section title of mention is not the beginning of article	Boolean
13	Institution words present (e.g., “Foundation”)	Boolean
14	Institution phrases present (e.g., “National Health”)	Boolean
15	Funding words present (e.g., “grant”)	Boolean
16	Funding phrases present (e.g., “we thank”)	Boolean
17	“Initial” words present (e.g., “meta-analyses”)	Boolean
18	“Initial” phrases present (e.g., “discovery phase”)	Boolean
19	“Replication” words present (e.g., “follow-up”)	Boolean
20	“Replication” phrases present (e.g., “control checks”)	Boolean
21	Persons words present (e.g., “individuals”)	Boolean
22	Persons phrases present (e.g., “study participants”)	Boolean
23	Ethnicity words present (e.g., “demographic”)	Boolean
24	Address words present (e.g., “Telephone”)	Boolean
25	Time period words present (e.g., “year”)	Boolean

Appendix C

Feature Set of CRF Tagger for Sample Size

Table C.1 describes a selection of feature functions from the feature set for the Conditional Random Field tagger for extracting sample sizes from the free text of the GWAS articles. This results in a final set of 412 feature functions.

C.1 Feature Selection

As the number of features generated by the feature functions can be orders of magnitude larger than simply the number of feature functions, the model is prone to overfitting. Therefore, we only include features that are observed at least k times in the training dataset, where k is a fixed integer value. $k = 5$ was found to work best for our purposes.

Table C.1. Feature Set of CRF tagger for Sample Size.

Feature		Example(s)
1	Stemmed token	“Association“ → “Assoc”
2	Part-of-Speech tag	“314“ → “CD”
3	Lemmatized token	“Persons” → “Person”
4	Suffix	“Persons” → {“ns”, “ons”, “sons”}
5	Prefix	“Persons” → {“Pr”, “Per”, “Pers”}
6	Initial capitalized	“Persons”
7	End capitalized	“PersonS”
8	All capitals	“ACCA”
9	Lowercase	“word”
10	Mixed-case	“Robert”
11	Roman numerals	“IV”
12	Hyphenated	“named-entity”
13	Word length	“Mediterranean” → 13
14	Greek	“Phi”
15	Units	“mg/L”, “kg”
16	Institution	“University”, “Agency”, “Council”
17	Funding	“Fund”, “Grant”, “Thank”
18	Time periods	“February”, “Years”
19	“Initial” stage	“genotyped”, “Discovery”
20	“Replication” stage	“follow-up”, “quality-control”
21	Persons	“individual”, “adult”, “subjects”
22	Groups	“couples”, “families”, “twins”
23	Ethnicity-related	“self-reported”, “ancestry”, “descent”
24	Address	“E-mail”, “Telephone”, “Box”

Appendix D

Feature Set for Ethnicity Group and Sample Size Extraction

Table B.1 describes the feature set used in extracting instances of ⟨`stage, ethnicity, size`⟩ from the articles. The set consists of real-valued TF-IDF vectors, integer-valued features, Boolean features (represented as 0 or 1), and other real-valued features. This feature set is common to both stages, and is represented as a real-valued sparse matrix of approximately 108,000 dimensions (or columns).

D.1 Feature Scaling

The features are normalized by removing the mean and scaling to unit variance across the values of each feature, or dimension of feature vector.

D.2 Feature Selection

As the feature set is mostly composed of various TF-IDF vectors, truncated SVD was explored as a feature selection technique to reduce the dimensionality of the feature vectors. However, this did not affect the performance significantly (and in fact degraded performance slightly) and hence the complete feature vectors were retained.

Table D.1. Feature Set for Ethnicity Group and Sample Size Extraction.

Feature		Type
1	Ethnicity mention token(s)	TF-IDF
2	Section title of ethnicity mention	TF-IDF
3	Section title of size mention	TF-IDF
4	Extracted passage for ethnicity mention (windowed)	TF-IDF
5	Extracted passage for size mention (windowed)	TF-IDF
6	Extracted sentence for ethnicity mention (full)	TF-IDF
7	Extracted sentence for size mention (full)	TF-IDF
8	Number of similar ethnicity mentions in article	Integer
9	Number of similar size mentions in article	Integer
10	Width of relation (in number of words)	Integer
11	Width of relation (in number of sentences)	Integer
12	Position of ethnicity mention in article	Integer
13	Position of size mention in article	Integer
14	Position of ethnicity mention in section	Integer
15	Position of size mention in section	Integer
16	Length of sentence that mention appears in	Integer
17	Number of identical relations of same width in article	Integer
18	Number of total identical relations in article	Integer
19	Cosine similarity between sentences of mentions	Real
20	Cosine similarity between passages of mentions	Real
21	Both mentions are in the same section	Boolean
22	Both mentions are not in the same section	Boolean
23	Both mentions are in the same sentence	Boolean
24	Both mentions are not in the same sentence	Boolean
25	Width (in words) is less than 5	Boolean
26	Width (in words) is not less than 5	Boolean
27	Width (in words) is less than 10	Boolean
28	Width (in words) is not less than 10	Boolean
29	Width (in words) is less than 20	Boolean
30	Width (in words) is not less than 20	Boolean
31	Size mention appears after ethnicity mention	Boolean
32	Size mention appears before ethnicity mention	Boolean
33	Size sentence appears after ethnicity sentence	Boolean
34	Size sentence appears before ethnicity sentence	Boolean

Table D.2. Feature Set for Ethnicity Group and Sample Size Extraction (contd.)

Feature		Type
35	Section of ethnicity mention is beginning of article	Boolean
36	Section of size mention is beginning of article	Boolean
37	Section title of mention is not the beginning of article	Boolean
38	Institution words present (e.g., “Foundation”)	Boolean
39	Institution phrases present (e.g., “National Health”)	Boolean
40	Funding words present (e.g., “grant”)	Boolean
41	Funding phrases present (e.g., “we thank”)	Boolean
42	“Initial” words present (e.g., “meta-analyses”)	Boolean
43	“Initial” phrases present (e.g., “discovery phase”)	Boolean
44	“Replication” words present (e.g., “follow-up”)	Boolean
45	“Replication” phrases present (e.g., “control checks”)	Boolean
46	Persons words present (e.g., “individuals”)	Boolean
47	Persons phrases present (e.g., “study participants”)	Boolean
48	Ethnicity words present (e.g., “demographic”)	Boolean
49	Address words present (e.g., “Telephone”)	Boolean
50	Time period words present (e.g., “year”)	Boolean

Bibliography

[1] 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.

[2] 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.

[3] Lora Aroyo and Chris Welty. Measuring crowd truth for medical relation extraction. In *2013 AAAI Fall Symposium Series*, 2013.

[4] Nguyen Bach and Sameer Badaskar. A review of relation extraction. *Unpublished manuscript*, 2007.

[5] 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.

[6] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[7] Charles Bouveyron. Weakly-supervised classification with mixture models for cervical cancer detection. In *Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence*, IWANN '09, pages 1021–1028, Berlin, Heidelberg, 2009. Springer-Verlag.

[8] Razvan C. Bunescu. Learning to extract relations from the web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, 2007.

[9] 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 crowdsourcing. *Database : the journal of biological databases and curation*, 2014.

[10] 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.

[11] 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.

[12] Benoît Frénay and Michel Verleysen. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, May 2014.

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

[14] John W Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576, 2009.

[15] JohnW. Graham. Multiple imputation and analysis with spss 17-20. In *Missing Data*, Statistics for Social and Behavioral Sciences, pages 111–131. Springer New York, 2012.

[16] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 427–434. Association for Computational Linguistics, 2005.

[17] Ying He and Mehmet Kayaalp. Biological entity recognition with conditional random fields. *AMIA Annual Symposium proceedings / AMIA Symposium*, pages 293–297, 2008.

[18] 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.

[19] Lucia A. Hindorff, Jacqueline A. L. MacArthur, Joannella Morales, Emily H. Bowler, Peggy Hall, Kent Klemm, Heather A. Junkins, Tony Burdett, Danielle Welter, Teri A. Manolio, and Helen Parkinson. Comprehensive curation and visualization of ethnicity information from published genome-wide association studies (GWAS): an improved GWAS catalog. October 2014.

[20] 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.

[21] Chun-Nan Hsu, Yu-Ming Chang, Cheng-Ju Kuo, Yu-Shi Lin, Han-Shen Huang, and I-Fang Chung. Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics*, 24(13):i286–i294, July 2008.

[22] Jing Jiang and Chengxiang Zhai. A systematic exploration of the feature space for relation extraction. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT07*, pages 113–120, 2007.

[23] 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.

[24] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, 32(4):485–525, December 2006.

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

[26] 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.

[27] Mengwen Liu, Yuan Ling, Yuan An, Xiaohua Hu, A. Yagoda, and R. Misra. Relation extraction from biomedical literature with minimal supervision and grouping strategy. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 444–449, Nov 2014.

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

[29] 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.

[30] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

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

[32] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[33] Ryan Mcdonald, Fernando Pereira, Seth Kulick, Scott Winters, Yang Jin, and Pete White. Simple algorithms for complex relation extraction with applications to biomedical ie. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 491–498, 2005.

[34] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

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

[36] 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.

[37] 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.

[38] Martin F. Porter. An algorithm for suffix stripping. *Program: Electronic Library & Information Systems*, 40(3):211–218, 1980.

[39] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo H. Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning From Crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[40] Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378, 2013.

[41] Rocco A. Servedio. Smooth Boosting and Learning with Malicious Noise. *Journal of Machine Learning Research*, 4:633–648, December 2003.

[42] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, JNLPBA '04, pages 104–107, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[43] 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.

[44] 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.

[45] 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.

[46] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Machine Learning*, 4(4):267–373, 2011.

[47] Maksim Tkachenko and Andrey Simanovsky. Named entity recognition: Exploring features. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 118–127. ÖGAI, September 2012. Main track: oral presentations.

[48] Paul T Von Hippel. Biases in spss 12.0 missing value analysis. *The American Statistician*, 58(2), 2004.

[49] 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.

[50] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Advances in Neural Information Processing Systems*, 2009.

[51] 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.

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

[53] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106, 2003.

[54] 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.

[55] 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.