August 2015

Three Essays on Enhancing Clinical Trial Subject Recruitment Using Natural Language Processing and Text Mining

Euisung Jung
University of Wisconsin-Milwaukee

THREE ESSAYS ON ENHANCING CLINICAL TRIAL SUBJECT RECRUITMENT USING NATURAL LANGUAGE PROCESSING AND TEXT MINING

by

Euisung Jung

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Management Science

at

The University of Wisconsin-Milwaukee

August 2015

ABSTRACT
THREE ESSAYS ON ENHANCING CLINICAL TRIAL SUBJECT RECRUITMENT USING NATURAL LANGUAGE PROCESSING AND TEXT MINING

By

Euisung Jung

The University of Wisconsin-Milwaukee, 2015
Under the Supervision of Dr. Hemant Jain and Dr. Atish Sinha

Patient recruitment and enrollment are critical factors for a successful clinical trial; however, recruitment tends to be the most common problem in most clinical trials. The success of a clinical trial depends on efficiently recruiting suitable patients to conduct the trial. Every clinical trial research has a protocol, which describes what will be done in the study and how it will be conducted. Also, the protocol ensures the safety of the trial subjects and the integrity of the data collected. The eligibility criteria section of clinical trial protocols is important because it specifies the necessary conditions that participants have to satisfy.

Since clinical trial eligibility criteria are usually written in free text form, they are not computer interpretable. To automate the analysis of the eligibility criteria, it is therefore necessary to transform those criteria into a computer-interpretable format. Unstructured format of eligibility criteria additionally create search efficiency issues. Thus, searching and selecting appropriate clinical trials for a patient from relatively large number of available trials is a complex task.

A few attempts have been made to automate the matching process between patients and clinical trials. However, those attempts have not fully integrated the entire matching process and have not exploited the state-of-the-art Natural Language Processing (NLP) techniques that may improve the matching performance. Given the importance of patient recruitment in clinical trial research, the objective of this research is to automate the matching process using NLP and text mining techniques and, thereby, improve the efficiency and effectiveness of the recruitment process.

This dissertation research, which comprises three essays, investigates the issues of clinical trial subject recruitment using state-of-the-art NLP and text mining techniques.

Essay 1: Building a Domain-Specific Lexicon for Clinical Trial Subject Eligibility Analysis

Essay 2: Clustering Clinical Trials Using Semantic-Based Feature Expansion

Essay 3: An Automatic Matching Process of Clinical Trial Subject Recruitment

In essay1, I develop a domain-specific lexicon for n-gram Named Entity Recognition (NER) in the breast cancer domain. The domain-specific dictionary is used for selection and reduction of n-gram features in clustering in eassy2. The domain-specific dictionary was evaluated by comparing it with Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT). The results showed that it add significant number of new terms which is very useful in effective natural language processing In essay 2, I explore the clustering of similar clinical trials using the domain-specific lexicon and term expansion using synonym from the Unified Medical Language System (UMLS).

I generate word n-gram features and modify the features with the domain-specific dictionary matching process. In order to resolve semantic ambiguity, a semantic-based feature expansion technique using UMLS is applied. A hierarchical agglomerative clustering algorithm is used to generate clinical trial clusters. The focus is on summarization of clinical trial information in order to enhance trial search efficiency. Finally, in essay 3, I investigate an automatic matching process of clinical trial clusters and patient medical records. The patient records collected from a prior study were used to test our approach. The patient records were pre-processed by tokenization and lemmatization. The pre-processed patient information were then further enhanced by matching with breast cancer custom dictionary described in essay 1 and semantic feature expansion using UMLS Metathesaurus. Finally, I matched the patient record with clinical trial clusters to select the best matched cluster(s) and then with trials within the clusters. The matching results were evaluated by internal expert as well as external medical expert.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Essay 3

ACKNOWLEDGEMENTS

I extend my sincere thanks and appreciation to my advisors, Dr. Hemant Jain and Dr. Atish Sinha. This dissertation could not have been written without their support and encouragement. They have provided me with tremendous support from my first day as a doctoral student. I also thank the members of my PhD committee, Dr. Huimin Zhao, Rashmi Prasad, and Carmelo Gaudioso for their helpful advice and suggestions. I cannot thank my committee enough.

Additionally, I especially want to thank my parents and in-laws for believing in me and emotionally supporting me throughout my studies. They have sacrificed their lives for my wife and myself and provided unconditional love and care.

Most importantly, I wish to thank my wife Dr. Eun Ju Jung. In many ways, the pursuit of a doctorate degree is a lonely process. Without her continued patience, support, encouragement, and sacrifice, my dissertation could not have been completed. There are no words that can express my gratitude and appreciation for all she has done for me.

Extra special thanks goes to my lord God.

CHAPTER 1

INTRODUCTION

"Never before in history has innovation offered
promise of so much to so many in so short a time."
Bill Gates

Basic science research has flourished over the past few decades and transferred knowledge into dramatic scientific advances for the treatment and prevention of human disease. As a result of these advances, new therapeutic agents, procedures, and devices have appeared. The healthcare industry has experienced decades of growth and success (Nussenblatt and Meinert, 2010).

Ever since the evidence-based practice was adopted, efforts have increased to base medical care as much as possible on the evidence of scientific research rather than on expert opinion or personal experience (National Research Council, 2001). A scientific experiment that provides one of the least biased type of clinical research evidence is the randomized controlled trial (RCT) (Sim et al., 2004). Moreover, RCTs are the most rigorous way to decide the existence of a cause-effect relationship between treatment and outcome (Sibbald and Ronalrd, 1998). In this sense, RCTs help to move basic scientific research from the laboratory into treatment for humans.

An RCT is also called a randomized clinical trial when it is applied to clinical research (Peto et al., 1976). A clinical trial is defined as “Research studies that explore whether a medical strategy, treatment, or device is safe and effective for humans” (National Institutes of Health, http://www.nhlbi.nih.gov/health/health-

topics/topics/clinicaltrials/). The main objective of a clinical trial is to evaluate the efficacy and / or effectiveness of a medical intervention with human subjects. Thus, new treatment can be proven safe and effective before public deployment. Cautiously conducted clinical trials are considered the fastest and safest way to find new treatments (NLM, http://www.nlm.nih.gov/medlineplus/tutorials/clinicaltrials/).

It is clear that a clinical trial is one of most important resources of practical medical knowledge. Over the past decade, the total number of clinical studies registered on ClinicalTrials.gov, based on the First Received Date, has dramatically increased (Figure 1). ClinicalTrials.gov, run by the U.S. National Library of Medicine (NLM), is the official public registry of clinical trials. To date (as of July, 2015), there were more than 190,000 trials for about 5,000 diseases on ClinicalTrials.gov (ClinicalTrials.gov).

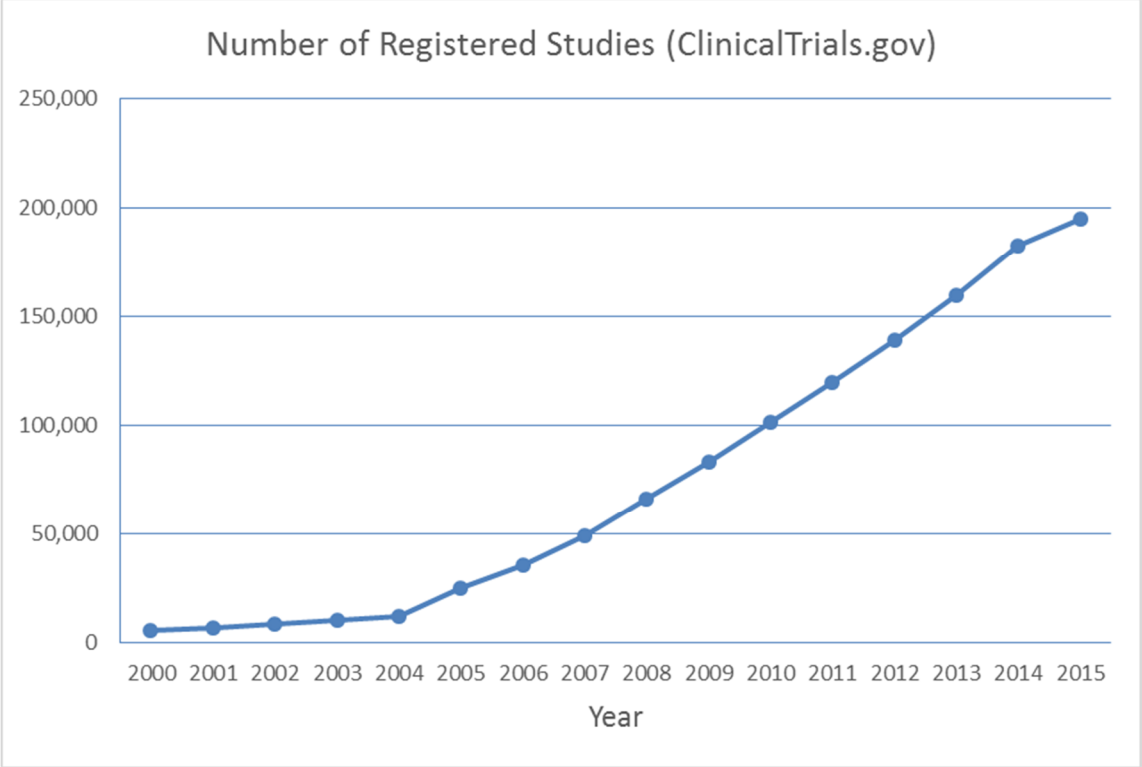

Figure 1. Number of Registered Studies
(ClinicalTrials.gov, 2015)

Patient recruitment and enrollment are critical factors for successful clinical trial research (Fran, 2004), and it is well known that subject recruitment is the most common problem in most clinical trials (Ashery and Mcauliffe, 1992). Inadequate recruitment can disrupt a clinical trial research timetable, waste resources, reduce the trial's ability to detect treatment effectiveness, and perhaps result in the failure of a clinical trial research project (Ashery and Mcauliffe, 1992). Accordingly, it is essential to achieve clinical trial research participant enrollment to conduct a successful trial (Frank, 2004).

In other words, the success of a clinical trial depends on efficiently recruiting suitable patients to conduct the trial. Insufficient patient participation from the time of a study's initiation to closeout might incur lack of statistical power to prove or disprove the goal of the clinical trial research (Frank, 2004). The main cause of recruitment problems includes the need for large samples and multiple eligibility criteria, subject reluctance, low patient treatment motivation, client dislike of research procedures, clinicians' distrust of research, and difficulties collaborating with treatment agencies.

Like other scientific research, every clinical research has a protocol that describes what will be done in the study and how it will be conducted. Also, the protocol ensures the safety of the trial subjects and integrity of the data collected. For this reason, it is a critical document for everyone involved in conducting the trial. In particular, the protocol of clinical trials should be followed precisely, since they deal with human subjects. In the U.S., diverse organizations, including the Office of Human Subjects Research Protection (OHRP) and the Food and Drug Administration (FDA), have the authority to determine whether certain clinical studies are adequately conducted according to their protocols.

The eligibility criteria section of clinical trial protocols is important because it specifies the necessary conditions of clinical research participants (Luo et al., 2011). According to the definition from the U.S. National Library of Medicine (U.S. NLM, ClinicalTrials.gov), eligibility criteria for clinical trials are "the medical or social standards determining whether a person may or may not be allowed to enter a clinical trial; they are based on such factors as age, gender, the type and stage of a disease, previous treatment history, and other medical conditions."

Since clinical research eligibility criteria are usually written in free text form, they are not computer interpretable. A popular method for achieving computable eligibility criteria is knowledge representation, which often requires labor-intensive manual effort and medical expert encoders in identifying the semantics of the eligibility criteria (Luo et al. 2010; Samson et al. 2011). No one can deny that standard-based formal computer understandable representation of eligibility criteria would provide obvious benefits for supporting clinical research and care use cases (Ross et al. 2010; Weng et al 2010). Therefore, the necessity for transforming free text eligibility criteria into a computable format has increased. In the last few years, a considerable number of attempts have been made at formal representations of eligibility criteria (Samson et al. 2011; Luo et al. 2010; Luo et al. 2010; Weng et al. 2010; Ross et al. 2010; Luo et al. 2013).

Unstructured characteristics of eligibility criteria raise other issues for search efficiency. It is not a simple task for a patient to search a huge repository and select appropriate clinical trials because subject eligibility criteria are not in a structured form but in free text form. The results from existing trial search engines usually are not satisfactory and require a manual process to refine relevant studies (Boland et al. 2013).

Boland et al. (2013) proposed feature-based indexing, clustering, and search of clinical trials, but their work still depends on a manual process by an expert for selection eligibility criteria features. To the best of my knowledge, no attempts have so far been made to build an entire automatic matching process for clinical trial clusters and patient information using state of art NLP and text mining algorithms. Given the importance of patient recruitment in clinical trial research, the objective of this dissertation is to build an integrated automatic matching process for clinical trials and patient information that enhances efficiency and effectiveness of the clinical trial subject recruitment process by using a NLP and Text Mining technique.

Essay 1 examines the building of a breast cancer domain-specific lexicon for n-gram Named Entity Recognition (NER). The domain-specific dictionary is used for selection and reduction of word n-gram features in the clinical trial clustering and matching patients to clusters and clinical trials. Essay 2 explores clustering of similar clinical trials using the domain-specific lexicon built in Essay1 and a synonym relationship from the Unified Medical Language System (UMLS). I generated word n-gram features and modified the features with the domain-specific dictionary matching process. In order to resolve semantic ambiguity, all synonym tags from the UMLS are annotated to the original features. A hierarchical agglomerative clustering (HAC) algorithm is used to generate clinical trial clusters. The focus of essay 2 is to examine the summarization of clinical trial information at cluster level to enhance trial search efficiency. Finally, essay 3 investigates an automatic matching process for patient information with clinical trial clusters and clinical trials within matched clusters.

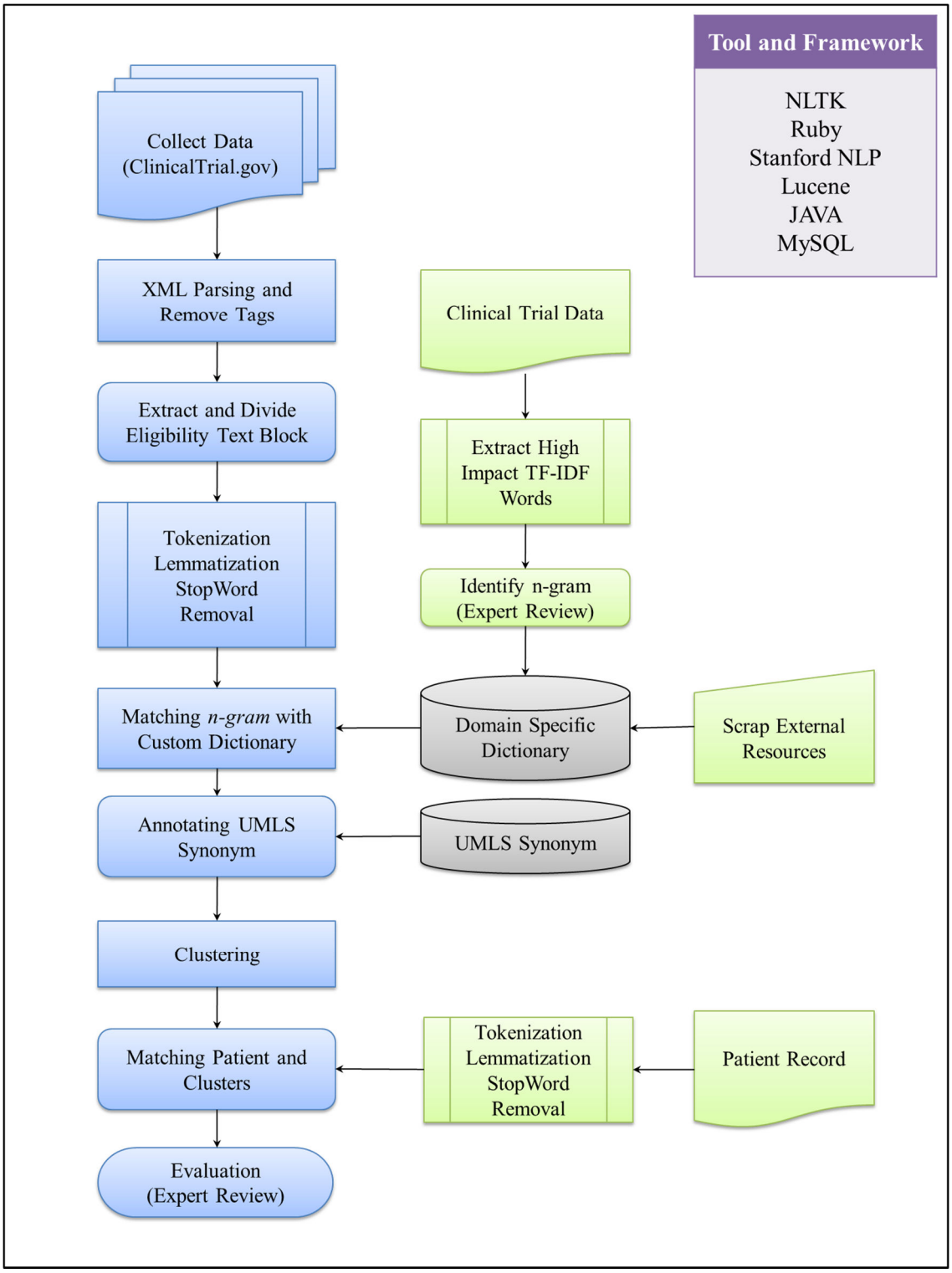

Figure 2. Steps in Patient and Clinical Trial Matching
Using a Domain Specific Dictionary and UMLS Synonyms

Figure 2 shows all the research steps included in essays 1, 2, and 3.

Essay 1: Building a Domain-Specific Lexicon for Clinical Trial Subject Eligibility Analysis

It is well understood that an NLP application requires sophisticated lexical resources to support its processing goals. Different solutions have been proposed to identify multi-gram disease named entities in the healthcare informatics literature. Jimeno et al. (2008) found that dictionary look-up provides competitive results with statistical approach and MetaMap solution, indicating that the use of disease terminology is highly standardized throughout the terminologies and the literature. Although there has been extensive effort made in the identification of protein- and gene-named entities (PGNs) in the biomedical literature, little research has been done on the recognition and resolution of terminologies in the clinical trial subject eligibility analysis.

A lexicon plays a significant role in all forms of medical language processing (Luo et al., 2010). At present, there is no comprehensive lexicon to capture multi-gram medical terminology in clinical trial eligibility criteria, especially in the breast cancer domain.

The goal of essay 1 is to build a breast cancer specific lexicon to cover clinical trial eligibility criteria and complete the multi-gram medical terminology.

Essay 2: Clustering Clinical Trials Using Semantic-Based Feature Expansion

With so much data and information around us, it becomes a problem to find pieces that are relevant. Therefore, a great deal of effort has been made to reduce the clinical trial search space. However, most of the proposed solutions require users to understand data structure and to generate complex database queries. The need for

understanding various medical terminologies remains an unsettled issue. The Unified Medical Language System® (UMLS®) was initiated and is now being maintained by The National Library of Medicine (NLM). The objective of UMLS is to facilitate the development of computer systems that deal with the semantics of the language of biomedicine and healthcare. In essay 2, I propose a novel clustering method to narrow the clinical trial search space using a custom dictionary and the UMLS Semantic Network.

Essay 3: An Automatic Matching Process for Clinical Trial Subject Recruitment

The process of new treatment and new drug development is extremely time consuming and expensive. A key bottleneck in this process is subject recruitment in clinical trials. Of all clinical trials conducted globally, more than 80% are delayed due to slow patient recruitment. This delay may cost the pharmaceutical companies millions of dollars per day in terms of lost sales. Speeding up patient recruitment in clinical trials can result in lower drug development costs and, ultimately, new drugs that are more affordable to patients.

In essay 3, I propose a novel automatic matching process of clinical trials and patient medical records. First, patient records were collected from a prior study and were pre-processed for tokenization and lemmatization. Second, the pre-processed patient records were matched with breast cancer custom dictionary and UMLS Metathesaurus for semantic feature expansion. Finally, I compared each pre-processed patient record with clinical trial clusters and each clinical trial study within matched clusters. The matching results are evaluated by internal expert as well as external medical expert.

CHAPTER 2

Essay 1: Building a Domain-Specific Lexicon for Clinical Trial Subject Eligibility Analysis

"Not everything that can be counted counts
and not everything that counts can be counted"
Albert Einstein

2.1. Introduction

There is a growing number of healthcare-related corpora and document data in the free text form, and with this comes the need to analyze and draw meaningful information. However, it is not easy to retrieve and query relevant information from text data. There have been several attempts at applying NLP and text mining techniques to the healthcare domain.

Clinical trials are designed to answer specific questions about the effects of a therapy or technique designed to improve human health. They rely on eligibility criteria, which specify who is qualified for clinical research study participation and who is disqualified. However, analysis of clinical trial subject eligibility text is not a typical text analysis task since it has some intriguing characteristics. In particular, the clinical trial subject eligibility section comprises a variety of biomedical terms that include abbreviations and acronyms. Moreover, clinical trial subject eligibility texts are not usually complete syntactically. They are not depicted by complete sentences, but outlined by succinct and fragmented phrases. For example, a sentence in the inclusion criteria of the clinical trial id 'NCT01068483' is 'Progressive, recurrent unresectable disease' which is not a grammatically complete sentence.

There is an increasing need to efficiently transform these free text clinical research eligibility criteria into computable formats to support the subject recruitment process. Various approaches have been proposed to achieve high-performance text analysis of clinical trial subject eligibility criteria. Prior work has typically used the Bag of Words (BOW) model as features for text analysis. However, the BOW approach does not recognize multi-word terms, which are typical in medical and healthcare domains.

The term dependency is the issue in the general BOW approach. The n-gram model takes into consideration the context information of a word, which depends on a previous or next word (Khan, 2010). But while the n-gram model improves the text analysis performance, it decreases the performance if the word length of n is greater than 3 (Liu, 2008).

A lexicon is fundamental to all forms of medical language processing and plays a significant role (Lou et al. 2010). Dictionary-based n-gram features induction, in which only those n-gram*s* that appear in a pre-defined dictionary are used as features (Remus and Rill, 2013). The n-gram feature induction approach yields the most accurate discriminative model for machine learning-based text analysis within a specific domain. Moreover, the dictionary-based n-gram feature induction leads to large dimensionality reductions. Thus, this feature selection may significantly reduce both noise and feature space size.

At present, there is no lexicon resource for identifying the n-gram terms in breast cancer clinical trial eligibility. In this essay, I build a domain-specific lexicon to facilitate analysis of a breast cancer clinical trial subject eligibility section. To the best of my knowledge, such a study has not been carried out before.

This essay is structured as follows. The next section reviews prior research in lexicon-building. In section 2.3, I describe short representations of textual documents, term frequency, inverse document frequency, data-driven n-gram feature induction, and the dictionary-based word n-gram feature induction approach. In section 2.4, the n-gram lexicon building process is described, and I compare the proposed lexicon and UMLS to evaluate its effectiveness in section 2.5. Finally, I draw conclusions and point out possible directions for future work in section 2.6.

2.2. Background Literature

Over the past few years, there have been several studies on clinical trials that use the text mining approach. One of the salient research streams is formal representation of eligibility criteria (Weng et al. 2009). Tu et al (2011) examined formalizing eligibility criteria in a computer-interpretable language to facilitate eligibility determination for study subjects and the identification of studies on similar patient populations. ERGO (Eligibility Rule Grammar and Ontology) annotation is used for capturing the semantics of criteria. Luo et al. (2013) examined a semi-automatic process to extract Common Data Elements (CDEs) in eligibility criteria of clinical trials. Luo et al. (2013)'s study is the first study using text mining in CDE discovery from free text clinical trial eligibility criteria.

There have been foundational studies on enhancing eligibility criteria representation. Luo et al. (2010) presented a corpus-based approach to create a semantic lexicon for clinical research eligibility criteria using UMLS. The main purpose of that research was to reduce the ambiguity in UMLS sematic-type assignment while building a

semantic lexicon for clinical trial eligibility criteria. A total of 20 UMLS semantic types, representing about 17% of all the distinct semantic types assigned to corpus lexemes, covered about 80% of the vocabulary of our corpus.

Temporal knowledge representation from temporal express in clinical research eligibility criteria is also a topic being actively investigated. Boland et al (2012) identified the temporal knowledge representation requirements of eligibility criteria by reviewing annotated 100 eligibility criteria. They developed EliXR-TIME, a frame-based representation designed to support semantic annotation for temporal expressions in eligibility criteria by reusing applicable classes from well-known clinical temporal knowledge representations (Boland et al, 2012). Luo et al. (2011) presented an ontology-based approach for extracting temporal information from clinical trial eligibility criteria. They developed a Conditional Random Field (CRF)-based parser, which is based on Temporal Awareness and Reasoning Systems for Question Interpretation (TARSQI) toolkit and the TimeText project, to automatically annotate the elements of temporal constraints, specifically focusing on clinical trial eligibility criteria. The results were evaluated with an additional 60 randomly selected eligibility criteria.

Another active research topic is effective and efficient search of clinical trials. Korkontzelos et al (2012) presented Assisting Search and Creation Of clinical Trials (ASCOT), a search application focused on clinical trials. Text mining and data mining methods were applied to ASCOT and an eligibility criteria recommendation component was included.

There has been much research on application, usage, and evaluation of UMLS. Wu et al. (2012) examined characteristics of UMLS Metathesaurus terms in clinical

notes. A 51 million document corpus of Mayo Clinic clinical notes was analyzed with modified Aho-Corasick algorithm and the occurrences of UMLS terms were statistically computed in terms of string attributes, source terminologies, semantic types, and syntactic categories. They found that on average 44.64 term matched per document and only 3.56% of the available case-insensitive terms in the UMLS were utilized. Aronson et al. (2001) depicted a MetaMap program developed by the NLM to map biomedical text to the UMLS Metathesaurus or to discover Metathesaurus concepts referred to in text. Fung et al. (2010) investigated the problem list terminologies (PLT) of large healthcare institutions and identified a subset of concepts based on standard terminologies. Data were acquired from six large-scale healthcare institutions and mapped with the UMLS Metathesaurus.

Feature selection and summarization of clinical trial is an emerging research topic. Boland et al. (2013) investigated the feasibility of feature-based indexing, clustering, and search of clinical trials. They argued that concept-oriented eligibility features could enhance user search effectiveness, facilitating meaningful and efficient indexing for clinical trials. In their study, concept-oriented eligibility features are a clinically meaningful atomic patient state, such as diagnosis, symptom, or demographic characteristics, which are derived from eligibility criteria. They argued that no studies have ever examined feature-based indexing for clinical trials system; thus, their work could set a baseline.

The ultimate goal of eligibility criteria analysis is directed at the automatic matching process between a clinical trial and patients. Wilcox et al. (2009) presented a model, electronic Participant Identification and Recruitment Model (ePaIRing), which

uses patient information to enhance patient recruitment in clinical trials. The model was created by grounded theory analysis, which is a qualitative approach. It iteratively collect and interpret data to arrive at explanation of data (Wilcox et al., 2009).

Over the past decades a considerable number of studies have been done on clinical trial and its subject eligibility criteria. However, no studies have ever tried to generate a domain-specific lexicon resource, even though it is recognized that a domain-specific lexicon is fundamental of medical text analysis and foundation of NLP and text mining. Thus, in the first essay, I generated a breast cancer-specific multi-gram lexicon by inducing high impacted multi-gram terms from clinical trial description as well as integrating heterogeneous online resources.

Table 1 show the selected research on clinical trial

Table 1. Selected Research on Clinical Trial

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Boland MR, Miotto R, Gao J, Weng C,	Methods of Information in Medicine (2013)	Feasibility of Feature-based Indexing, Clustering, and Search of Clinical Trials on ClinicalTrials.gov: A Case Study of Breast Cancer Trials,				
Luo Z, Miotto R, Weng C,	Journal of Biomedical Informatics (2012)	A Human-Computer Collaborative Approach to Identifying Common Data Elements in Clinical Trial Eligibility Criteria	To identify Common Data Elements (CDEs) in eligibility criteria	association rule-learning algorithm , UMLS, dice coefficient	Clinicaltrials.org (breast cancer and cardiovascular)	
Weng C, Wu X, Luo Z, Boland M, Theodoratos D, Johnson SB	Journal of the American Medical Informatics Association, (2011)	EliXR: An Approach to Eligibility Criteria Extraction and Representation				
Luo Z, Yetisgen-Yildiz M, Weng C,	Journal of Biomedical Informatics, (2011)	Dynamic Categorization of Clinical Research Eligibility Criteria by Hierarchical Clustering				

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Weng C, Tu SW, Sim I, Richesson R	Journal of Biomedical Informatics (2010)	Formal Representations of Eligibility Criteria: A Literature Review	Review eligibility criteria knowledge representation	Analyze publications	PubMed, Google, 27 systems	
Thadani S, Weng C, Bigger JT, Ennever J, Wajngurt D	Journal of the American Medical Informatics Association (2009)	Electronic Screening Improves Efficiency of Clinical Trials Recruitment	evaluate the performance of an electronic screening (E-screening) method		125 patients, investigator review	significantly reduced the screening burden associated with the ACCORD trial
Weng C, McDonald DW, Gennari JH,	International Journal of Medical Informatics (2007)	Participatory Design of a Collaborative Clinical Trial Protocol Writing System				
Mary Regina Boland, Samson W. Tu, Simona Carini, Ida Sim, Chunhua Weng	Proc of 2012 AMIA Clinical Research Informatics Summit	ELIXR-TIME: A Temporal Knowledge Representation for Clinical Research Eligibility Criteria	temporal expressions is needed to facilitate temporal information extraction		100 eligibility criteria from ClinicalTrials.gov	EliXR-TIME, a frame-based representation designed to support semantic annotation for temporal expressions in eligibility criteria

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Luo Z, SB Johnson, AM Lai, Weng C	Proc of 2011 AMIA Fall Symposium	Extracting Temporal Constraints from Clinical Research Eligibility Criteria Using Conditional Random Fields	develop automated approaches for extracting the primary constructs of temporal constraints in clinical research eligibility criteria	Conditional Random Fields (CRFs) to train a temporal parser from manually-annotated criteria	150 temporal eligibility criteria randomly selected from ClinicalTrails.gov	
Weng C, Batres C, Borda T, Weiskopf NG, Wilcox AB, Bigger JT, Davidson K,	Proc of 2011 AMIA Fall Symposium	A Real-Time Screening Alert Improves Clinical Trial Recruitment Efficiency				
Weng C, Bigger JT, Busacca L, A Wilcox, A Getaneh,	Proc of AMIA 2010 Fall Symposium	Comparing the Effectiveness of a Clinical Data Warehouse and a Clinical Registry for Supporting Clinical Trial Recruitment: A Case Study				
Luo Z, Johnson SB, Weng C,	Proc of AMIA 2010 Fall Symposium	Semi-Automatic Induction of Semantic Classes from Free-Text Clinical Research Eligibility Criteria Using UMLS				
Luo Z, Duffy R, Johnson SB, Weng C	Proc of AMIA Clinical Research Informatics Summit 2010	Corpus-based approach to create a semantic lexicon for clinical research eligibility criteria using UMLS				

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Wilcox AB, Natarajan K, Weng C	Proc of AMIA Translational Bioinformatics Summit 2009	Using Personal Health Records for Automated Clinical Trials Recruitment: the ePaIRing Model				
Li, L, Chase H, Patel C, Friedman C, and Weng C	Proc of 2008 AMIA Fall Symposium	Comparing ICD9-Encoded Diagnoses and NLP-Processed Discharge Summaries for Clinical Trials Pre-Screening: A Case Study.				
Weng C, Becich M, Fridsma D	The 2nd International Conference on Information Technology and Communications in Health, Feb 2007,	Collective Domain Modeling across Clinical Trials Standards: Needs, Challenges, and Design Implications				
Weng C, Gennari JH, McDonald DW	11th World Congress on Medical Informatics (MedInfo'04)	A Collaborative Clinical Trial Protocol Writing System				

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Gennari JH, Weng C, McDonald DW, Benedetti J, Green S	11th World Congress on Medical Informatics (MedInfo'04)	An Ethnographic Study of Collaborative Clinical Trial Protocol Writing				
Weng C, McDonald DW, Gennari JH	IT in Health Care: Socio-technical Approaches 2nd International Conference, 13-14 September 2004	Participatory Design of A Collaborative Clinical Trial Protocol Writing System				
Weng C, Kahn MG, Gennari JH	Proc of AMIA 2002 Fall Symposium	Temporal Knowledge Representation for Scheduling Tasks in Clinical Trial Protocols.				

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
D.W. Lonsdale, C. Tustison, C.G. Parker, D.W. Embley	Data & Knowledge Engineering (2008)	Assessing clinical trial eligibility with logic expression queries	identification, extraction, and query formulation of information regarding medical clinical trials	web-based information extraction		Query generation
Marc Cuggia, Paolo Besana, David Glasspool	International journal of medical informatics (2011)	Comparing semi-automatic systems for recruitment of patients to clinical trials	review decision support systems for automatic recruitment of patients to clinical trials			
Ida Sim, Ben Olasov, and Simona Carini	Journal of Biomedical Informatics (2004)	An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method	developing RCT Bank to capture detailed information about the design, execution, and results of RCTs	competency decomposition		RCT Schema using UMLS
Y. Megan Kong, Carl Dahlke, Qun Xiang, Yu Qian, David Karp, Richard H. Scheuermann	Journal of Biomedical Informatics(2011)	Toward an ontology-based framework for clinical research databases	integrate data standards and ontology structures of knowledge representation	database implementation of the OBX model		Ontology-Based eXtensible (OBX) data model

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Guoqian Jiang, Harold R. Solbrig, Christopher G. Chute	Journal of Biomedical Informatics(2011)	Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network	relationship between terminological annotations and the UMLS Semantic Network (SN) that can be exploited to improve those annotations	UMLS SN	caDSR CDE Browser	the UMLS SN based profiling approach is feasible for the quality assurance and accessibility of the cancer study CDEs
P.J. Embi et al.	Arch Intern Med, (2005)	Effect of a clinical trial alert system on physician participation in trial recruitment				
P.A. Harris et al.	Acad Med (2012)	ResearchMatch: a national registry to recruit volunteers for clinical research				
S.W. Tu et al.	Journal of Biomedical Informatics (2011)	A practical method for transforming free-text eligibility criteria into computable criteria	creating computer-interpretable languages for eligibility criteria	ERGO annotations	1000 eligibility criteria randomly drawn from ClinicalTrials .gov	incrementally capturing the semantics of free-text eligibility criteria

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
J. Nahar *et al.*	J Med Syst,	Significant cancer prevention factor extraction: an association rule discovery approach				
Peter J. Embi, MD, MS; Anil Jain, MD; Jeffrey Clark, BS; Susan Bizjack, MSN; Richard Hornung, DrPH; C. Martin Harris, MD, MBA	Arch Intern Med. (2005)	Effect of a Clinical Trial Alert System on Physician Participation in Trial Recruitment	the resources of a comprehensive EHR can be leveraged for the benefit of clinical trial recruitment	EHR-based clinical trial alert (CTA) system	From Cleveland Clinic	The CTA intervention was associated with significant increases in the number of physicians
Stephanie Heinemann, Sabine Thüring, Sven Wedeken, Tobias Schäfer, Christa Scheidt-Nave, Mirko Ketterer, Wolfgang Himmel1	BMC Medical Research Methodology 2011,	A clinical trial alert tool to recruit large patient samples and assess selection bias in general practice research	evaluate the recruitment performance of the practice staff when using the CTA tool according to 4 criteria	clinical trial alert (CTA) tool	GP's data	

2.3. Background

Tokenization

Tokenization is the process of breaking up a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing, such as parsing or text mining (Manning et al., 2008). Manning et al. (2008) defined a token as "an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing." Table 2 explains an example of tokenization.

Table 2. Example of Tokenization (Manning et al., 2008)

Input	Friends, Romans, Countrymen, lend me your ears
Output	Friends Romans Countrymen lend me your ears

Lemmatization

For grammatical reasons, there are diverse forms of a word, such as *organize*, *organizes*, and *organizing*. Likewise, families of derivationally related words with similar meanings, such as *democracy*, *democratic*, and *democratization*, are common in textual data. In NLP and text mining, it is useful for a search of words to return documents that contain another word in the set. The goal of lemmatization is to reduce inflectional forms and derivationally related forms of a word to a common base form.

There are two different approaches to obtain a common base form: stemming and lemmatization. Stemming refers to a crude process that cuts off the end of words in order to acquire a base form and includes the removal of derivational affixes. Lemmatization refers to a process that uses a vocabulary and morphological analysis of words for the purpose of removing inflectional endings and returning the base or dictionary form of a word called lemma. Lemmatization makes use of full morphological analysis to accurately identify the lemma for each word. In this study, I use lemmatization rather than stemming, which does not guarantee returning grammatically correct words.

Table 3 shows an example of lemmatization, and Table 4 shows an example of lemmatization with a sentence.

Table 3. Example of Lemmatization (Manning et al., 2008)

Base Form	Inflectional or Derivationally related form
be	am, are, is
car	car, cars, car's cars'

Table 4. Example of Lemmatization with a Sentence (Manning et al., 2008)

Original Sentence	the boy's cars are different colors
Lemmatization	the boy car be differ color

Stop Word Removal

A document is a combination of sentences, and a sentence is a set of words. A word is a complicated combination of characters. There is a variety of words and special characters that do not have significant meaning. Some extremely common words called stop words would appear to be of little value in NLP and text mining process. Examples of stop words are "the," "of," "to," and "a. " These are required to satisfy English grammar rules even though they have no semantic meaning. Moreover, special characters such as the period and question mark used to indicate the end of a sentence or an interrogative sentence, respectively, are considered to be noise in NLP. Therefore, all of the stop words need to be removed in the pre-processing step.

The general strategy for removing stop words in English is to use a stop list that is a negative dictionary. Fox (1989) reported a stop list based on the Brown corpus of 1,014,000 words drawn from a broad range of literature in English. The final product of Fox's work is a list of 421 stop words that would be maximally efficient and effective in filtering the most frequently occurring and semantically neutral words.

This study adopts Fox's stop list to cull all insignificant words in data.

Representation of Textual Documents and Vector Space Model

Text representation is one of the pre-processing processes that is used to reduce the complexity of documents and make them easier to handle. To implement any technique of text mining, it is initially necessary to transform the digitized texts

in an efficient and meaningful way so that they can be analyzed.

The space vector model is the most commonly used approach to represent textual documents. This approach represents a text by a numerical vector obtained by counting the most relevant lexical elements present in the text (Amine et al., 2008).

All documents d_j will be transformed into a vector:

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{|T|j}) \qquad (1)$$

Where T is the whole set of terms (or descriptors) that appear at least once in the corpus (|T| is the size of the vocabulary), and w_{kj} represents the weight (frequency or importance) of the term t_k in the document d_j

Table 5 represents a Document Term Matrix model.

Table 5. Document Term Matrix

Documents	**Terms or Descriptors**						
d_1	w_{11}	w_{21}	w_{31}	…	w_{j1}	…	w_{n1}
d_1	w_{12}	w_{22}	w_{32}	…	w_{j2}	…	w_{n2}
…	…	…	…	…	…	…	…
d_m	w_{1m}	w_{2m}	w_{3m}	…	w_{jm}	…	w_{nm}

I represented each clinical trial document by a vector in a multidimensional space. Each word constitutes a dimension in this space. When a word is absent in a

clinical trial document, its value along the corresponding dimension is 0. When a word occurs in the document, the value along the dimension is determined by a weight factor indicating its importance.

The simplest representation of texts introduced within the framework of the vector space model is called Bag of Words (BOW) (Salton and McGil, 1986). It consists of texts transformed into vectors where each component represents a word. This representation of texts excludes any grammatical analysis and any concept of distance between the words, and syntactically destructures texts by making them understandable to the machine.

Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. There are many methods to calculate the weight w_{kj} knowing that for each term, it is possible to calculate not only its frequency in the corpus but also the number of documents that contain this term.

Most approaches (Sebastiani, 2002) are centered on a vectorial representation of texts using the TF-IDF measure. The frequency *TF* of a term *T* in a corpus of textual documents corresponds to the number of occurrences of the term *T* in the corpus. The frequency *IDF* of a term *T* in a corpus of textual documents corresponds to the number of documents containing *T*. These two concepts are combined (by product) in order to assign a stronger weight to terms that appear often in a document and rarely in the complete corpus.

$$TF \times IDF(t_k, d_j) = Occ(t_k, d_j) \times log \frac{Nb\_doc}{Nb\_doc(t_k)}$$

Where $\mathrm{Occ}(t_k, d_j)$ is the number of occurrences of the term t_k in the document d_j, *Nb_doc* is the total number of documents of the corpus and *Nb_doc*(t_k) is the number of documents of this unit in which the term t_k appears at least once (Amine et al., 2008).

N-gram and N-gram Induction

N-gram (Damashek, 1995) is a character sequence of length n extracted from a document. To generate the n-gram vector for a document, a window n characters in length is moved through the text, sliding forward by a fixed number of characters (usually one) at a time. At each position of the window, the sequence of characters in the window is recorded. For example, the first three 5-grams in the phrase "character string" are "chara," "harac," and "aract." Damashek (Damashek, 1995) suggested the use of character n-grams instead of words for gauging text similarity. N-gram retrieval promises lower vulnerability to data entry errors, spelling varieties, word conjugations, and other morphological varieties.

The concept of n-grams was first discussed in 1951 by Shannon (Shannon, 1951). Since then, n-grams have been used in many areas, such as spelling-related applications, string searching, prediction and speech recognition. Word n-gram is a sequence of consecutive tokens, with the length of n. Mostly words are taken as tokens, but in recent works, characters could also be token (Trenkle and Cavnar, 1994).

A word n-gram feature induction, sometimes also referred to as feature extraction, induces features on textual data based on a set of word n-grams. With feature induction, the textual data is represented in a feature space, usually encoding the existence of these word n-grams or their frequency. The word n-grams to be used as features may be chosen by either using a data driven approach or dictionary-based approach.

In a data driven feature induction, every word n-gram combination from the textual data is created. Thus, the feature size equals the word n-gram vocabulary size. Such a data driven feature induction does not require prior domain knowledge to recognize meaningful word n-grams.

In a dictionary approach, n-gram tokens are selected based on a custom lexicon database that focuses on a specific domain. In this approach, it is proposed that an n-gram feature selection that maps all bigram and trigram tokens to the custom lexicon database be used.

Named Entity Recognition (NER)

Named Entity Recognition (NER) is the task of identifying and classifying entities such as person names, place names, organization names, etc., in a given document. Named entities play a major role in information extraction. A well-performing NER is important for further levels of NLP techniques. Many techniques have been applied in English for NER. Some of them are rule-based systems (Krupka and Hausman, 1998), which make use of dictionary and patterns of named entities. Examples are Decision trees (Karkaletsis et al., 2000), Hidden Morkov Model (HMM) (Biker, 1997), Maximum Entropy Morkov Model (MEMM) (Borthwick et

al., 1998), and Conditional Random Fields (CRF) (Andrew McCallum and Wei Li, 2003). The approaches can be classified as a rule-based approach, machine learning approach, or hybrid approach.

NER has been done generically but can also be domain-specific where a finer tagset is needed to describe the named entities in a domain. Domain-specific NER is common and has been in existence for a long time in the bio-domain (Settles 2004) for identification of protein names, gene names, DNA names, etc. The NER task is also viewed as the first step of information extraction of free text clinical studies describing shock, trauma, inflammation, and other related states (Apostolova et al., 2008). The proposed custom dictionary also supports the NER process in the healthcare domain.

2.4. Research Method

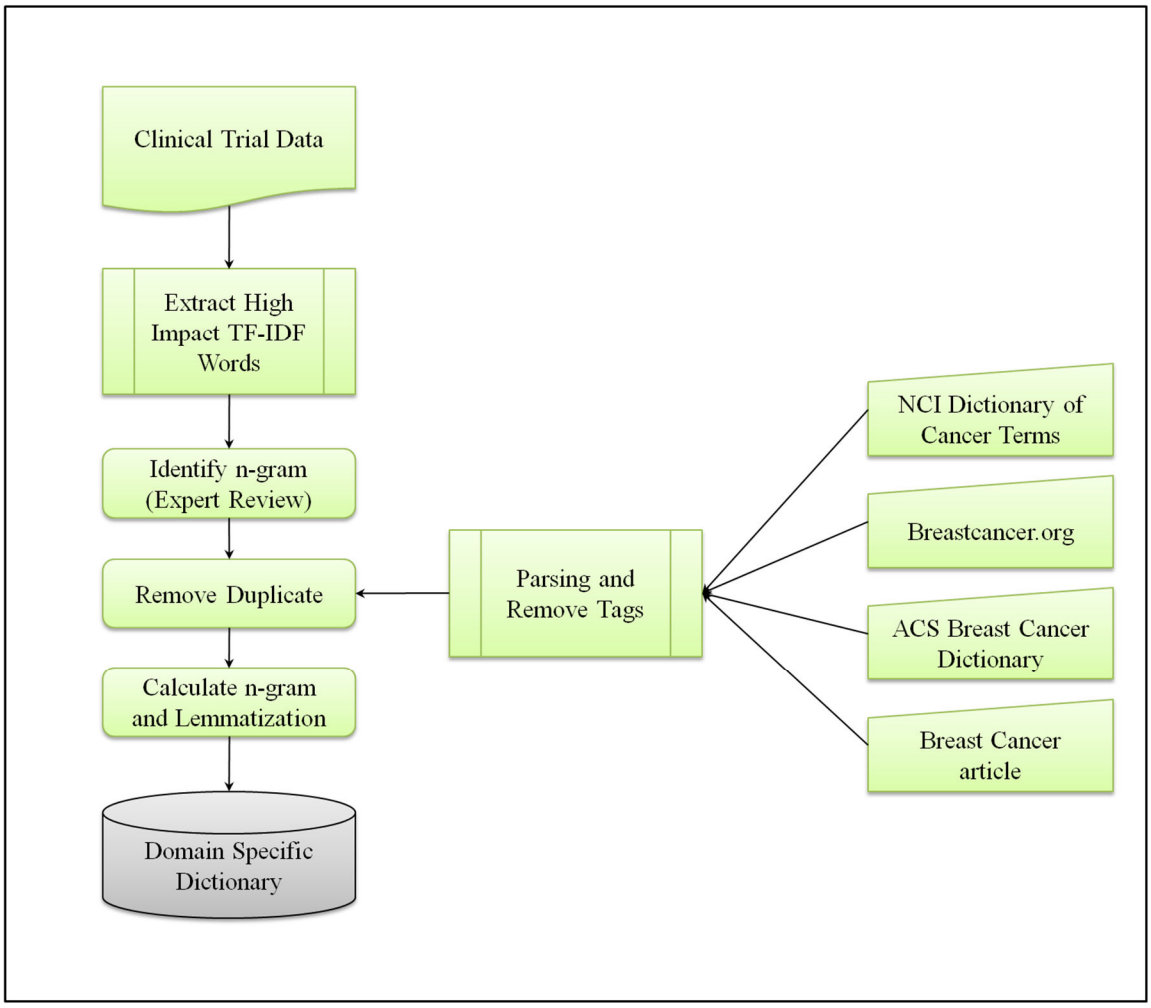

Figure 3. Steps for Building Domain Specific Dictionary

Figure 3 represents the steps for building breast cancer domain-specific dictionary in essay 1.

2.4.1 Data Set

I collected 378 clinical trials using search term 'Breast Cancer' that was listed in ClinicalTrials.gov between January 1, 2010 and January 1, 2011, and downloaded all the related information as a collection of individual XML files.

All XML tags and metadata were removed and only the <eligibility> - <criteria> - <textblock> section was extracted. Since subject eligibility criteria text is in free text format and contains two opposite criteria, "inclusion" and "exclusion", I separated subject eligibility criteria text blocks based on the key words "Inclusion criteria" and "Exclusion criteria."

2.4.2 Building Dictionary

I constructed a domain-specific n-gram term dictionary for the breast cancer domain. The custom dictionary was based on high TF-IDF score words from the clinical trial eligibility data set and other online resources (i.e., NCI Dictionary of Cancer Terms, Breastcancer.org, and ACS Breast Cancer Dictionary). The n-gram term dictionary for breast cancer domain can be a resource for dictionary-based feature induction that uses a pre-defined dictionary as well as an NER process.

First, during pre-processing, tokenization, lemmatization, and stop word removal were performed over the selected data set.

Second, I calculated the TF-IDF score for all unigram features that drew from the breast cancer clinical trial eligibility text data set. Three experts reviewed all 26,193 unigram list organized in descending order by TF-ID score, and they manually identified bigram and trigram terms from the unigram list. The review was conducted sequentially. The output of first reviewer was forwarded to second reviewer and the

results from second reviewer was validated by the third reviewer. The review process was iterated among three reviewers until all reviewer agreed on identified bigram and trigram. The final review was conducted by expert who is medical doctor as well as Ph.D. Only adjacent words were considered. After expert review, a total of 1,506 multi-gram terms were identified.

Third, an online medical term crawler was developed by the author in Ruby language to gather breast cancer terms from web sites. The crawler automatically collected web documents from the targeted site and parsed the documents to extract medical terms. All unnecessary tags were removed. The crawler collected 4,704 terms from the NCI dictionary of Cancer Terms, 910 terms from Breastcancer.org, 155 terms from the ACS Breast Cancer Dictionary, and 28 terms from breast cancer glossary of Terms in emedicinehealth.com. All the collected items were stored in MySql database and duplicates were removed by SQL query. Table 6 shows the number of dictionary items by the source.

Table 6. Number of Item by Source

Source	Number of item
NCI Dictionary of Cancer Terms	4,704
Clinical Trial cluster	1,506
Breastcancer.org	910
ACS Breast Cancer Dictionary	155
emedicinehealth.com	28
Total	7,303

The custom dictionary included total 7,303 items. The dictionary included 707 trigram, 2,098 bigram, 4,162 unigram, and 336 n-grams terms consisting of more than three words that were identified. Table 7 shows the number of dictionary items by the type of n-gram.

Table 7. Number of Item by N-Gram Type

Type of n-gram	Number of item
1	4,162
2	2,098
3	707
4	191
5	93
6	37
7	11
8	3
9	1
Total	7,303

2.5. Evaluation

Two experiments were conducted to evaluate the efficiency of the domain-specific dictionary. First, all items in the custom dictionary were directly matched with the SNOMED CT in UMLS Metathesaurus to examine uniqueness of the custom dictionary items. Only English terms in SNOMED CT were used for evaluation. I created a database query with Structured Query Language (SQL) and ran the query to evaluate uniqueness. The SQL query selected all items in the custom dictionary and matched each item with terms in SNOMED CT. According to the query result, 4,243 items in the custom dictionary were unique and 3,060 items overlapped with SNOMED CT. This evaluation showed that around 58% of the custom dictionary items are newly introduced as a lexicon resource. This is significantly high and should not be overlooked. Table 8 shows number of unique item for different types of n-gram in the custom dictionary.

Table 8. Number of Unique Items in the Custom Dictionary

Type of n-gram	Number of Unique Item
1	2168
2	1299
3	486
4	159
5	84
6	32
7	11
8	3
9	1
Total	4,243

Second, the items in the custom dictionary and in SNOMED CT were matched with test data set to validate usefulness of custom dictionary for processing clinical trial data. A total 1,058 clinical trial studies from January 1, 2011, to January 1, 2013 were collected from the ClinicTrial.gov and the subject eligibility criteria section was divided into two parts, inclusion criteria and exclusion criteria. These two data sets were pre-processed with tokenization, lemmatization, and stop word removal. All possible trigram and bigram combination were generated to match with

the proposed custom dictionary and SNOMED CT. The matching results for trigram and bigram are presented in Table 9 and **Error! Reference source not found.**

Table 9. Trigram Matching

Trigram Matching				
	Number of Matched Items Using the Custom Dictionary Only	**Number of Matched Items Using SNOMED CT Only**	**Number of Matched Items Using the Custom Dictionary and SNOMED CT**	**Additional Number of Unique Items Matched by Custom Dictionary**
Inclusion Data	828	904	1,439	535
Exclusion Data	748	984	1,226	242
Total	1,576	1,888	2,665	777

Table 10. Bigram Matching

Bigram Matching				
	Number of Matched Items Using the Custom Dictionary Only	**Number of Matched Items Using SNOMED CT Only**	**Number of Matched Items Using the Custom Dictionary and SNOMED CT**	**Additional Number of Unique Items Matched by Custom Dictionary**
Inclusion Data	4,842	6,932	9,610	2,678
Exclusion Data	4,158	5,958	8,198	2,240
Total	9,000	12,890	17,808	4,918

According to the matching results, the SNOMED CT matched most items which were expected since the size of SNOMED CT is much larger than the custom

dictionary in both trigram and bigram. However, the number of matched items by using both the custom dictionary and SNOMED CT is greater than the number of matched items using SNOMED CT only. As shown in Table 9, 777 additional trigram matches were done by adding custom dictionary to SNOMED CT only match. This represents a 41% increase over SNOMED CT only match. Similarly as shown in Table 10, 4,918 additional bigram matches were done by adding custom dictionary to SNOMED CT only match. This represents a 38.6% increase over SNOMED CT only match. Thus, custom dictionary significantly increases the size of matches.

2.6. Discussion

One of the most time consuming and high labor cost tasks in text mining research is the creation, compilation, and customization of the necessary lexicons (Jonnalagadda et al., 2013). Lexical resources are requisite to improve the performance of text mining, especially in NER. For the healthcare informatics researchers, it is required to implement modularized systems that cannot be generalized, therefore the building of customized lexical resources is needed for these highly specific systems (Stanfill et al., 2010).

This research has attempted to build a domain-specific lexicon focusing on breast cancer and has shown the semi-automated dictionary building process. The evaluations for breast cancer domain-specific dictionary using the clinical trial subject eligibility documents revealed that even though the total number of matched items using the custom dictionary is than the number of matched items using SNOMED CT, about 30% of matched items using the custom dictionary and SNOMED CT were

derived from the custom dictionary. This shows the importance of the domain specific dictionary and expert knowledge in lexicon resources.

There is no research that is free from limitation. First, coverage rate of domain-specific dictionary is relatively low. The domain-specific dictionary included limited online sources. Thus, if more extensive resource such as NCI Thesaurus is included in future research, it will result in better performance. The evaluation of this research only calculated the matched terms with test data set. If an annotated data as gold standard is available, more sophisticated evaluation metrics such as precision, recall, and F-measure could be included. In future research, with expert's annotation for test data set, the most popular performance measures in information retrieval, accuracy, precision, recall, and F-measure could be evaluated.

CHAPTER 3

Essay 2: Clustering Clinical Trials Using Semantic-Based Feature Expansion

"The problems are solved, not by giving new information,
but by arranging what we have known since long."
Ludwig Wittgenstein

3.1. Introduction

The subject eligibility criteria section is one of the essential parts of clinical research protocols since it specifies the inclusion and exclusion characteristics of clinical research participants. Since clinical trial protocols and result data have been digitized and made publicly available by the National Institutes of Health (NIH), there has been an increasing need for developing novel approaches that exploit such an invaluable resource. However, there are several challenges to acquiring meaningful knowledge from an unstructured data source (Bollier, 2010).

One of the salient issues in data analysis is information overload. When searching relevant clinical trials in the one of largest online clinical trial repositories, ClinicalTrials.gov, which includes more than 190,000 clinical trial studies, the same information overload problem was encountered. Many scholarly methods such as EmergingMed, SearchClinicalTrials.org, and TrialX application have been developed to address this problem. Although a large number of studies have been made on narrowing the clinical trial search scope, they required users to create complex queries (Hao, 2014).

An alternative option for a query-based clinical trial search is a case-based search by clustering trials, which can identify and suggest similar trial to an example trial (Hao, 2014). This approach can alleviate user burden to create complex query and can be useful for multiple usage cases. Clinical trial participants, clinical trial investigators, and meta-analysis researchers can benefit from the case-based search approach (Hao, 2014). To support case-based clinical trial search, it is necessary to develop an automated method for identifying and grouping semantic classes that belong to clinical trial subject eligibility criteria.

Interpretation of a subject eligibility section by means of a computer has received considerable attention for its promising applications in clinical trial research, especially in automatically matching patients to clinical trial studies (Luo, 2010). Inducing sematic classes from text data is an efficient way to understand text data and it is required to induce semantic classes from clinical trial eligibility criteria to understand that. Clustering is a popular solution for inducing semantic classes for various applications, such as ontology development, content organization, and thesaurus construction (Cheng et al., 2004; Pratt and Fagan, 2000; Lin, 1998).

In this research, we present a novel approach for reducing clinical trial information search space, which uses the result of hierarchical clustering with the n-gram model and semantic-based feature expansion technique.

3.2. Background Literature

Table 11 presents the selected research on clinical trial using NLP and text mining.

Table 11. Selected Research on Clinical Trial using NLP and Text Mining

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Tianyong Hao et al.	Journal of Biomedical Informatics (2014)	Clustering clinical trials with similar eligibility criteria features	Identify and cluster clinical trials with similar eligibility features.	Center-based clusters	From ClinicalTrials.gov	useful for clinical trial eligibility criteria designs and for improving clinical trial recruitment
S.W. Tu et al.	Journal of Biomedical Informatics, (2011)	A practical method for transforming free-text eligibility criteria into computable criteria	Creating computer-interpretable languages for eligibility criteria	ERGO annotations	1000 eligibility criteria randomly drawn from ClinicalTrials.gov	incrementally capturing the semantics of free-text eligibility criteria
Guoqian Jiang, Harold R. Solbrig, Christopher G. Chute	Journal of Biomedical Informatics (2011)	Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network	Relationship between terminological annotations and the UMLS Semantic Network (SN) that can be exploited to improve those annotations	UMLS SN	caDSR CDE Browser	the UMLS SN based profiling approach is feasible for the quality assurance and accessibility of the cancer study CDEs

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Boland MR, Miotto R, Gao J, Weng C,	Methods of Information in Medicine(2013)	Feasibility of Feature-based Indexing, Clustering, and Search of Clinical Trials on ClinicalTrials.gov: A Case Study of Breast Cancer Trials				
Luo Z, Yetisgen-Yildiz M, Weng C,	Journal of Biomedical Informatics (2011)	Dynamic Categorization of Clinical Research Eligibility Criteria by Hierarchical Clustering				
Luo Z, Johnson SB, Weng C,	Proc of AMIA 2010 Fall Symposium	Semi-Automatic Induction of Semantic Classes from Free-Text Clinical Research Eligibility Criteria Using UMLS				
Weng C, Wu X, Luo Z, Boland M, Theodoratos D, Johnson SB	Journal of the American Medical Informatics Association, (2011)	EliXR: An Approach to Eligibility Criteria Extraction and Representation				

3.3. Background

Unified Medical Language System (UMLS)

UMLS was initiated in 1989 by the National Library of Medicine (NLM), which continues to maintain it. It is an attempt to fill the gap among the medical vocabularies from heterogeneous sources. The purpose of UMLS is to facilitate the development of computer systems that deal with the semantics of the language of biomedicine and health. NLM provides system developers with the UMLS Knowledge Sources (database) and related software applications (programs) for building healthcare information systems that create, process, retrieve, integrate, and aggregate biomedical and health data, as well as for use in academic research (Kohler 2008; http://www.nlm.nih.gov/pubs/factsheets/umls.html).

UMLS consists of three knowledge sources, which are the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon. Moreover, the three knowledge sources comprise several tools that facilitate the use of UMLS.

Metathesaurus

The Metathesaurus is a very large, multipurpose, and multilingual vocabulary database that is organized by concepts. The current release contains more than 1.5 million biomedical terms from over 150 different sources. Synonymous terms are clustered together to form a concept. For example, "breast cancer," "breast tumor malignant," and "malignant neoplasm of breast" belong to the same UMLS concept. The concept unique identifier (CUI) for "breast cancer" is C0006142.

There are various types of relationships that link concepts to other concepts. Inter-concept relationships are not only inherited from the vocabulary sources but are also created by the Metathesaurus editors. All concepts in the Metathesaurus are assigned to at least one semantic type from the Semantic Network to keep consistent categorization at the general level depicted in the Semantic Network.

Semantic Network

The main purpose of the Semantic Network is to provide a consistent categorization of all concepts stored in the Metathesaurus and information about a set of basic semantic types or categories. The Network contains 133 semantic types and 54 relationships. There are major groupings of semantic types under topics such as organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas. The scope of the UMLS semantic types is quite wide; therefore, it permits the semantic categorization to include a wide range of terminologies over multiple domains.

The Semantic Network is organized using a directed graph, where the semantic types represent the nodes and the relationships among them are the edges. Figure 4 illustrates a portion of the Network. The semantic type "Biologic Function" has two children, "Physiologic Function" and "Pathologic Function," and each of these in turn has several children. Each child and parent in the hierarchy is linked by an "is-a" link. Figure 5 illustrates a portion of the hierarchy for Network relationships. The "affects" relationship has six children, including "manages", "treats," and "prevents." Figure 6

shows a portion of the Semantic Network illustrating the relations, either hierarchical or associative, that exist between semantic types.

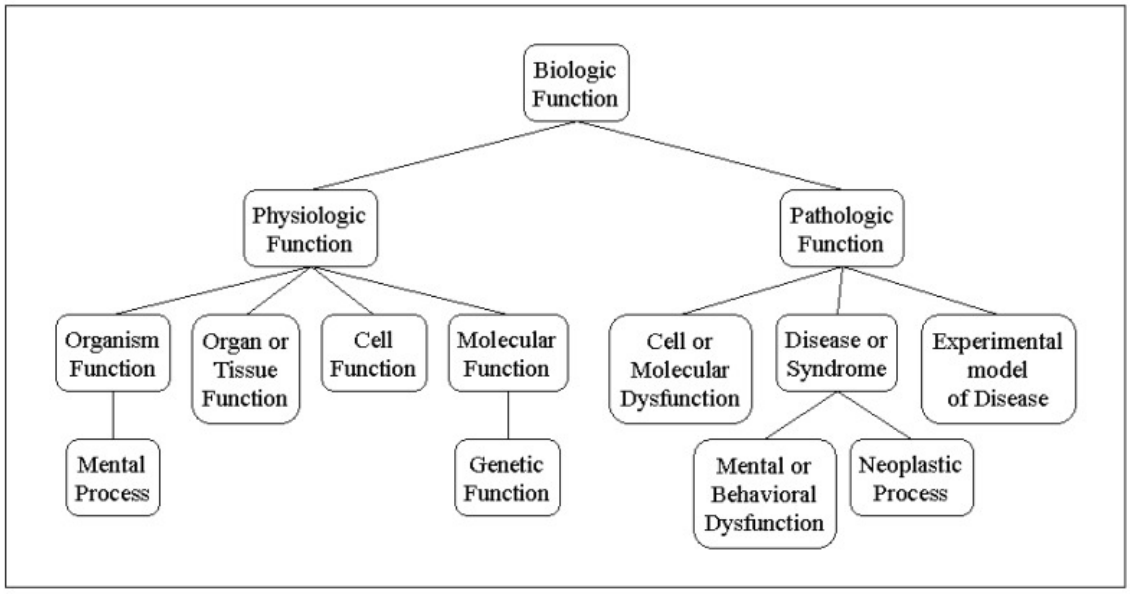

Figure 4. A Portion of the UMLS Semantic Network: "Biologic Function" Hierarchy
(UMLS Reference Manual, 2009)

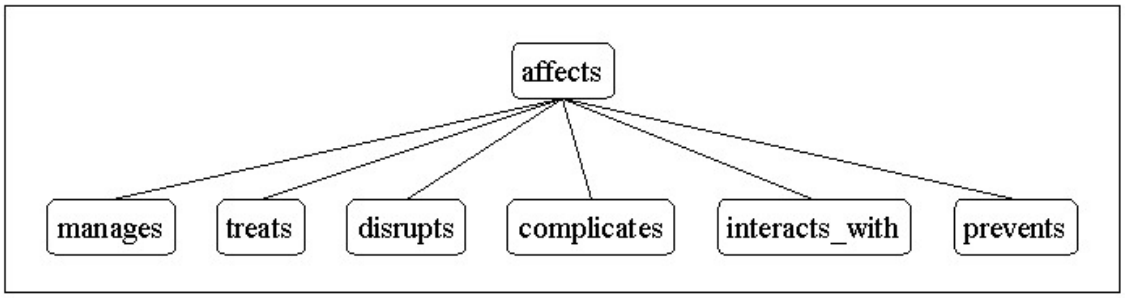

Figure 5. A Portion of the UMLS Semantic Network: "Affects" Hierarchy
(UMLS Reference Manual, 2009)

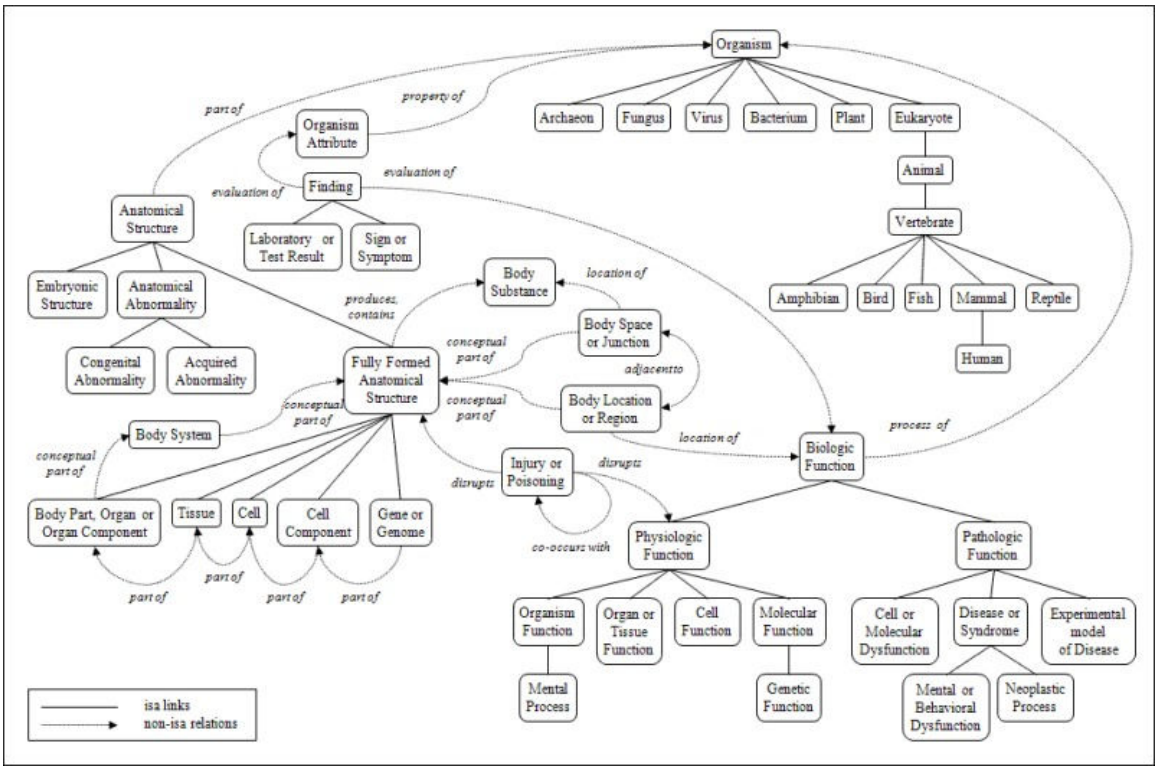

Figure 6. A Portion of the UMLS Semantic Network: Relations
(UMLS Reference Manual, 2009)

SPECIALIST Lexicon and Lexical Tools

The SPECIALIST Lexicon is a general English lexicon including many biomedical vocabularies, and the Lexical Tools are designed to resolve the high degree of variability in natural language. The syntactic, morphological, and orthographic information for each term in the Lexicon is recorded by the SPECIALIST NLP system. The inflected forms of words or terms are appropriately considered instances of the same word.

3.4. Research Method

Figure 7 shows the steps of clinical trial clustering process using UMLS. First, I collected clinical trial information for breast cancer from ClinicalTrial.gov and parsed original XML format files. Next, only eligibility criteria section from clinical trial was extracted and pre-processed using tokenization, lemmatization, and stop word removal. Breast cancer specific dictionary and UMLS Metathesaurus were used for finding n-gram terms and semantic feature expansion. Agglomerative hierarchical clustering algorithms were applied to create clusters for the inclusion and exclusion data set and then intersectional clusters were derived. Finally, a label for intersectional cluster was created.

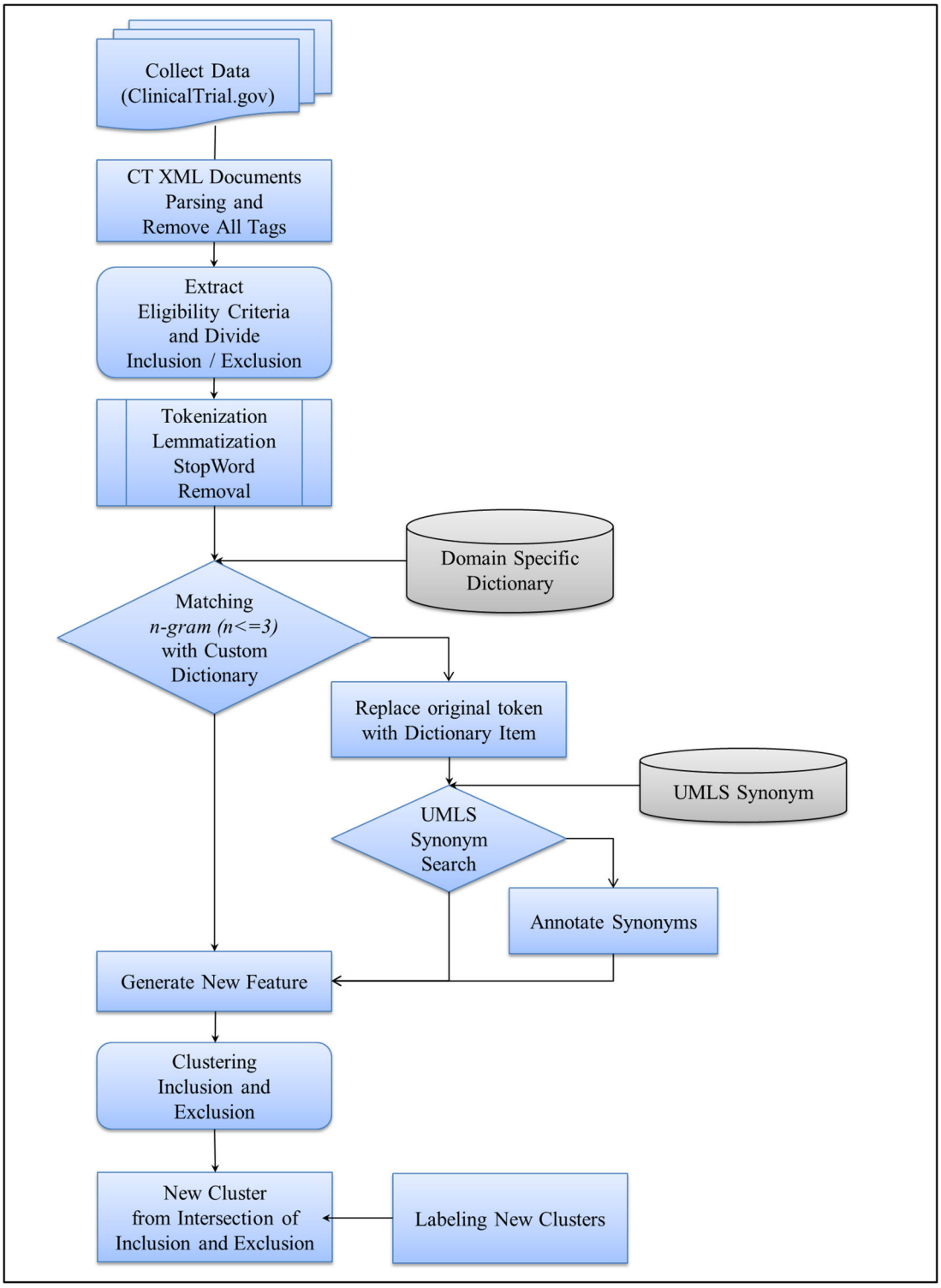

Figure 7. Steps for Clinical Trial Clustering Using Domain-Specific Dictionary and UMLS

3.4.1. Data Set

I collected the clinical trials from ClinicalTrials.gov, which is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world. I used the search term "breast cancer" to limit clinical trials to only the breast cancer domain and then collected three years of data from January 1, 2010, to January 1, 2013. The total number of clinical trials collected is 1,660, all information on the trials were downloaded as a collection of individual Extensible Markup Language (XML) format files. XML is a markup language that defines a set of rules for encoding a document in a format that is both human-readable and machine-readable. The World Wide Web Consortium (W3C) produces the specifications for XML 1.0 and XML and has come into common use for the interchange of data over the Internet (http://en.wikipedia.org/wiki/XML).

Figure 8 shows a sample of an original clinical trial XML document.

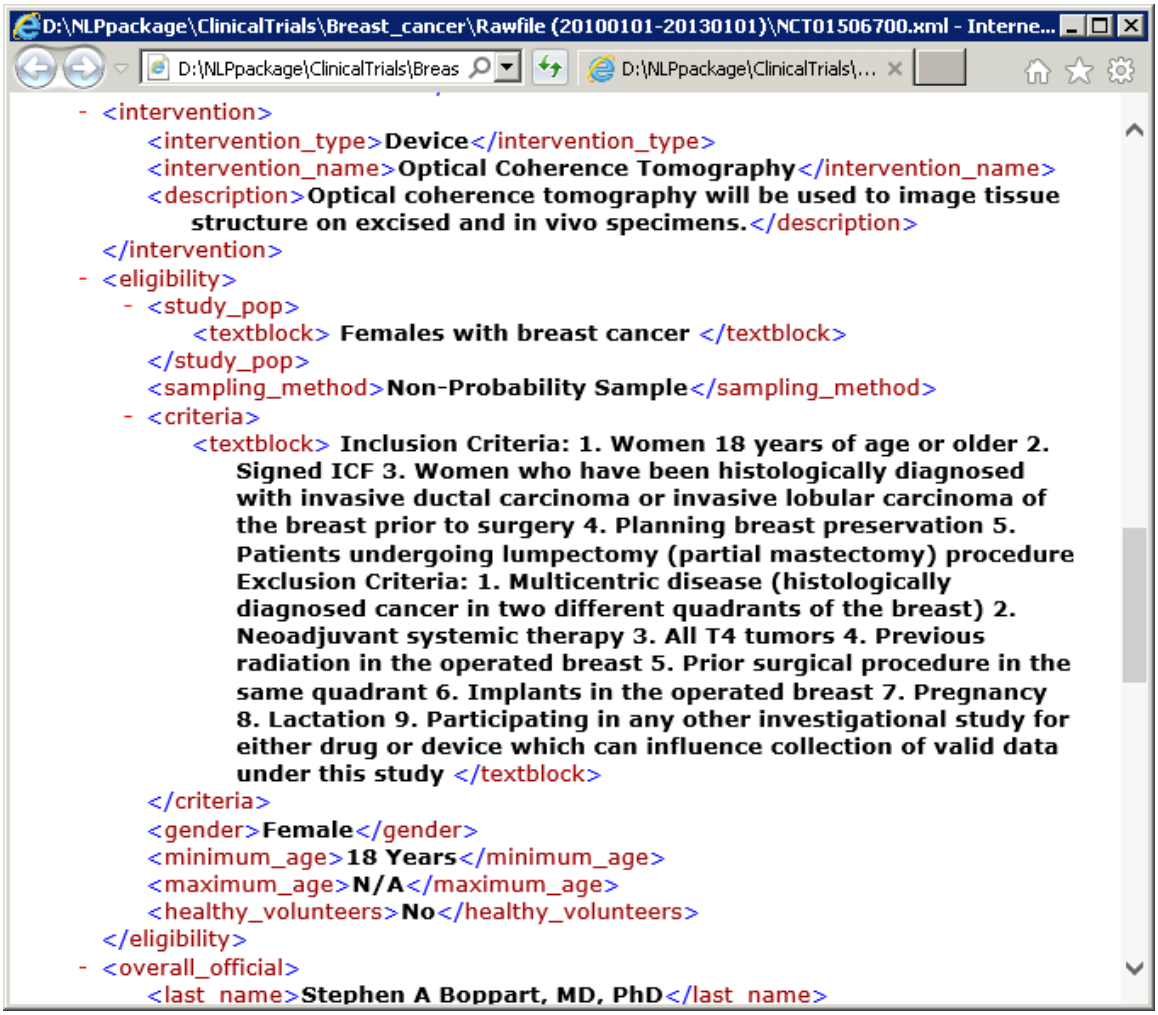

Figure 8. Sample of Original Clinical Trial XML Document (NCT01483196.xml)

To parse an XML document and remove unnecessary tags, I developed a custom parser using the Ruby programming language. All XML tags metadata were removed and only the <eligibility> - <criteria> - <textblock> section was extracted. Subject eligibility criteria text was in a free text format and could be divided by two opposite criteria: "Inclusion" and "Exclusion." I separated the subject eligibility criteria text block based on the key word "Inclusion criteria" and "Exclusion criteria." There were several upper and lower case variations in the keyword such as "INCLUSION CRITERIA," "inclusion

criteria," and "inclusion Criteria." The regular expression was used to capture all letter case variations. Once the eligibility criteria text was divided into two sections, all key words representing inclusion and exclusion were removed by the pre-processing step. Table 12 presents a sample of a criteria section extracted from the clinical trial id NCT0506700. The inclusion criteria set and exclusion criteria set were managed separately. The gender and age range were basic structured eligibility criteria that gave significant information during the matching process between the clinical trial and patient information. Thus, I extracted those two sections and included them with the data file naming rule. The naming rule for each eligibility criteria was <*clinical trial ID_gender criteria_minimum age_maximu age*>; as a result, the data file name for NCT050670 was modified to "NCT0506700_Female_18_NA." This file name implied that females over age 18 were eligible to participate in the clinical trial NCT0506700.

Table 12. Sample of Extracted Eligibility Criteria Text (ID; NCT0506700)

Original criteria	Inclusion Criteria: 1. Women 18 years of age or older 2. Signed ICF 3. Women who have been histologically diagnosed with invasive ductal carcinoma or invasive lobular carcinoma of the breast prior to surgery 4. Planning breast preservation 5. Patients undergoing lumpectomy (partial mastectomy) procedure Exclusion Criteria: 1. Multicentric disease (histologically diagnosed cancer in two different quadrants of the breast) 2. Neoadjuvant systemic therapy 3. All T4 tumors 4. Previous radiation in the operated breast 5. Prior surgical procedure in the same quadrant 6. Implants in the operated breast 7. Pregnancy 8. Lactation

	9. Participating in any other investigational study for either drug or device which can influence collection of valid data under this study
Inclusion only	1. Women 18 years of age or older 2. Signed ICF 3. Women who have been histologically diagnosed with invasive ductal carcinoma or invasive lobular carcinoma of the breast prior to surgery 4. Planning breast preservation 5. Patients undergoing lumpectomy (partial mastectomy) procedure
Exclusion only	1. Multicentric disease (histologically diagnosed cancer in two different quadrants of the breast) 2. Neoadjuvant systemic therapy 3. All T4 tumors 4. Previous radiation in the operated breast 5. Prior surgical procedure in the same quadrant 6. Implants in the operated breast 7. Pregnancy 8. Lactation 9. Participating in any other investigational study for either drug or device which can influence collection of valid data under this study

3.4.2. Pre-processing

For the first pre-processing step, I performed tokenization and lemmatization for the inclusion and exclusion data sets with Stanford CoreNLP, which is an integrated framework that provides a set of natural language analysis tools, including the part-of-speech (POS) tagger, the named entity recognizer (NER), the parser, the co-reference resolution system, the sentiment analysis tools, and model files for analysis of English (http://nlp.stanford.edu/software/corenlp.shtml). The Stanford CoreNLP code is written in Java and licensed under the GNU General Public License (v2 or later) and requires Java 1.6 or higher version.

The second pre-processing step was stop word removal. The Apache Lucene framework, a high-performance and full-featured text search engine library written entirely in Java, was used to remove stop words (http://lucene.apache.org/core/). Apache Lucene is an open source project available for free download. This study adopted Fox's stop word list to cull all insignificant words in the data.

Table 13 presents a sample of the preprocessed inclusion and exclusion criteria text NCT01506700.

Table 13. Sample of Preprocessed Inclusion and Exclusion Criteria

Pre-processed Inclusion Criteria	woman 18 year age older sign icf woman histologically diagnose invasive ductal carcinoma invasive lobular carcinoma breast prior surgery Planning breast preservation patient undergo lumpectomy partial mastectomy procedure
Pre-processed Exclusion Criteria	multicentric disease histologically diagnose cancer two different quadrant breast neoadjuvant systemic therapy t4 tumor previous radiation operate breast prior surgical procedure same quadrant implant operate breast pregnancy Lactation participate investigational study drug device influence collection valid datum under study

3.4.3. Matching with Custom Dictionary

The domain-specific dictionary for breast cancer described in essay 1 was utilized to detect n-gram terms in the inclusion or exclusion criteria data set. In this first step, I identified all trigram combinations from the preprocessed data set and then matched each trigram term with the custom dictionary that has n-gram terms for the breast cancer domain. Once the trigram term was matched with the term in the custom dictionary, three

unigram tokens that consisted of the trigram were removed from the data. I replaced the space in the trigram with an underscore (_) to transform the trigram into the single token form because all identified trigram and bigram words should be considered as unigram terms to maintain the original n-gram form. After the trigram matching step was completed, all bigram combinations from the modified data set were drawn and matched with the custom dictionary. Table 14 shows all trigram combinations from NCT0506700 and the results of the custom dictionary matching. From the pre-processed inclusion criteria presented in Table 14, I generated all trigram combinations that listed in Table 15. The first three tokens for clinical trial id 'NCT01506700' are 'woman', '18', 'year', so trigram 'woman 18 year' was generated and this trigram compared with the custom dictionary. If the trigram 'woman 18 year' was found in the custom dictionary, the three unigram tokens, 'woman', '18', and 'year' were removed from original data set and replaced with 'woman_18_year'. Otherwise, the first three token was kept and the combination window for trigram slid to next token and generated second possible trigram combination '18 year age'. All possible trigram combination from clinical trial id 'NCT01506700' inclusion criteria were compared with the custom dictionary and found two trigram matches 'invasive ductal carcinoma' and 'invasive lobular carcinoma'.

Table 14. All Trigram Combinations from NCT01506700 and Results of the Custom Dictionary Matching

Inclusion / Exclusion	All Trigram Combinations	Matching with The Custom Dictionary
Inclusion Criteria	woman 18 year 18 year age year age older	***invasive ductal carcinoma*** ***invasive lobular carcinoma***

	age older sign older sign icf sign icf woman icf woman histologically woman histologically diagnose histologically diagnose invasive diagnose invasive ductal ***invasive ductal carcinoma*** ductal carcinoma invasive carcinoma invasive lobular ***invasive lobular carcinoma*** lobular carcinoma breast carcinoma breast prior breast prior surgery prior surgery Planning surgery Planning breast Planning breast preservation breast preservation patient preservation patient undergo patient undergo lumpectomy undergo lumpectomy partial lumpectomy partial mastectomy partial mastectomy procedure	
Exclusion Criteria	multicentric disease histologically disease histologically diagnose histologically diagnose cancer diagnose cancer two cancer two different two different quadrant different quadrant breast quadrant breast neoadjuvant	No Match

	breast neoadjuvant systemic neoadjuvant systemic therapy systemic therapy t4 therapy t4 tumor t4 tumor previous tumor previous radiation previous radiation operate radiation operate breast operate breast prior breast prior surgical prior surgical procedure surgical procedure same procedure same quadrant same quadrant implant quadrant implant operate implant operate breast operate breast pregnancy breast pregnancy Lactation pregnancy Lactation participate Lactation participate investigational participate investigational study investigational study drug study drug device drug device influence device influence collection influence collection valid collection valid datum valid datum under datum under study	

Table 15 shows all possible bigram combinations from NCT0506700 and the results of the custom dictionary matching. I generated all bigram combinations that listed in Table 15 from the pre-processed inclusion criteria presented in Table 14. The first two tokens for clinical trial id ‘NCT01506700’ are ‘woman’ and ‘18’, so bigram ‘woman 18’ was generated and then compared with the custom dictionary. If the bigram ‘woman 18’ was found in the custom dictionary, the two unigram tokens, ‘woman’ and ‘18’ were removed from original data set and replaced with ‘woman_18’. Otherwise, the first two tokens were kept, and the combination window for bigram slid to next token and generated second possible bigram combination ’18 year’. All possible bigram combination from clinical trial id ‘NCT01506700’ inclusion criteria were compared with the custom dictionary and found one bigram match ‘invasive ductal carcinoma’ and ‘invasive lobular carcinoma’.

Table 15. All Bigram Combinations from NCT01506700 and Results of the Custom Dictionary Matching

Inclusion / Exclusion	All Bigram Combinations	Matching with The Custom Dictionary
Inclusion Criteria	woman 18 18 year year age age older older sign sign icf icf woman woman histologically histologically diagnose diagnose breast breast prior	***partial mastectomy***

	prior surgery surgery Planning Planning breast breast preservation preservation patient patient undergo undergo lumpectomy lumpectomy partial ***partial mastectomy*** mastectomy procedure	
Exclusion Criteria	multicentric disease disease histologically histologically diagnose diagnose cancer cancer two two different different quadrant quadrant breast breast neoadjuvant neoadjuvant systemic ***systemic therapy*** therapy t4 t4 tumor tumor previous previous radiation radiation operate operate breast breast prior prior surgical surgical procedure procedure same	***systemic therapy***

	same quadrant quadrant implant implant operate operate breast breast pregnancy pregnancy Lactation Lactation participate participate investigational investigational study study drug drug device device influence influence collection collection valid valid datum datum under under study	

3.4.4. Matching with the UMLS Semantic Network

Semantic-Based Feature Expansion Using UMLS

Identifying optimal feature sets is crucial for improving the effectiveness of text analysis (Chung 2009). There are two main research approaches to identifying optimal feature sets for text analysis. The focus of the first approach is on feature selection and extraction from relatively large documents. Usually, studies with a large corpus are concerned with reducing feature sets efficiently to identify the optimal feature sets that improve performance. The second approach focuses on expanding feature sets to find the optimal feature set that enhances performance. This approach utilizes relatively small

feature sets from small size documents and expands the features sets by adding semantically related features (Chung, 2009).

Tso et al. (2003) proposed a method of feature expansion to resolve the data sparseness problem, which is one of the most serious obstacles in research on word sense disambiguation (WSD). The experiment of using a word sense identifier with a feature expansion resulted in more than double the precision improvement over the baseline approach alone (Tso et al., 2003). A prior study (Chung, 2009) showed that expanded feature sets containing synonymous relationships significantly improved the results of text categorization. When expanding feature sets with synonyms used on classifier names, the effectiveness of text categorization considerably improved, regardless of word sense disambiguation (Chung, 2009). Fisher and Roark (2007) incorporated feature expansion techniques into their sentence-ranking framework and achieved substantial gains over the baseline framework, which does not include feature expansion steps.

Document representation through the simple BOW vector space model has a few shortcomings such as ignoring term dependencies, structure, and ordering of the terms in documents. To overcome these issues, Khan (2010) proposed Semantics Based Feature Vector using Part of Speech (POS) tags to extracts the concept of terms in feature set. Also, he used WordNet to extracts co-occurring and associated terms. The proposed method outperformed the TF-IDF with BOW feature selection method for text classification.

There have been several attempts to incorporate semantic features from the WordNet lexical database to improve the predictive performance of the text classification model (de Buenaga Rodriguez et al., 1997; Scott and Matwin, 1998; Jensen and Marinez,

2000; Kehagias et al., 2003; Hotho and Bloehdorn, 2004; Rosso et al., 2004; Peng and Choi, 2005; Mansuy and Hilderman, 2006). The rationale behind this is that the features in the training set alone are not enough to build a good model for categorization. However, if we incorporate the word relationships from WordNet, a more accurate model may be possible. Most prior studies reported that incorporating semantic features results in a statistically significant increase in accuracy (Mansuy and Hilderman, 2006).

The clinical trial eligibility criteria section is not a lengthy document but is a succinct description of clinical trial subject characteristics. Moreover, the contents in the clinical trial eligibility criteria are written by medical researchers, and the target audience are also medical experts; thus, the criteria usually include a large number of medical terms. For that reason, I incorporated synonymously related terms from the UMLS Semantic Network to expand feature sets based on semantic relatedness.

All trigram and bigram terms that were found in the custom dictionary were passed on to the next step to find synonyms from the UMLS Semantic Network. I created a custom query to find all synonymous relationships in the UMLS Semantic Network and then ran the custom query with each trigram and bigram term.

Table 16 shows the UMLS synonym matching results for each trigram and bigram term, and Table 17 shows the final feature set for clinical trial NCT01506700.

Table 16. UMLS Synonym Matching Result for NCT01506700

Inclusion / Exclusion	Trigrams and Bigrams Found in Custom Dictionary	Matching UMLS Synonyms
Inclusion Criteria	invasive ductal carcinoma	No Match
	invasive lobular carcinoma	No Match
	partial mastectomy	Subtotal mastectomy Segmental excision of breast Excision of part of breast Partial mastectomy Segmental resection of breast Segmental excision of breast
Exclusion Criteria	systemic therapy	No Match

Table 17. Final Feature Set for NCT01506700

Inclusion	woman 18 year age older sign icf woman histologically diagnose invasive_ductal_carcinoma invasive_lobular_carcinoma breast prior surgery Planning breast preservation patient undergo lumpectomy partial_mastectomy procedure Subtotal_mastectomy Segmental_excision_of_breast Excision_of_part_of_breast Segmental_resection_of_breast
Exclusion	multicentric disease histologically diagnose cancer two different quadrant breast neoadjuvant systemic_therapy t4 tumor previous radiation operate breast prior surgical procedure same quadrant implant operate breast pregnancy Lactation participate investigational study drug device influence collection valid datum under study

To the best of my knowledge, there has been no study that applies the semantic-based feature expansion technique to clinical trial clustering. This is the first study that adopts novel approaches that can improve text analysis performance for clinical trial subject eligibility clustering.

3.4.5. Hierarchical Clustering

Classification and clustering are two different types of data mining problems (Dunham, 2003). Also, they are two typical examples of supervised and unsupervised data mining.

Given a set of objects that is partitioned into a finite set of classes, classification is the task of automatically determining the class of an unseen object, based typically on a model trained on a set of objects with known class memberships. Clustering is the process of grouping data objects together on the basis of the features they have in common. The objects are grouped into clusters with the objective of maximizing the intra-cluster similarity and the inter-cluster dissimilarity between objects.

Classification is supervised in that it typically requires labeled training data to train a classifier. The categorization or automatic classification of texts is the task of distributing a set of documents according to some common characteristics. The terms “categorization” or “classification” are used when dealing with the assignation of a document to a predefined classes or categories.

Clustering is unsupervised since it is performed on raw input data with no prior knowledge, or supervision, over method. Unsupervised classification or "clustering" is automatic and discovers latent (hidden) unlabeled classes. The term “clustering”

designates the creation of classes or groups (clusters) of a certain number of similar objects without prior knowledge. The classes are isolated from one another and are discovered automatically. A large number of unsupervised classification methods have been applied to textual documents (Amine et al., 2008).

Hierarchical clustering is the clustering in which the clusters do not simply make a partition of the set of objects, but the set of objects are organized into a tree hierarchy so that any child cluster is a subset of the parent cluster and the sibling clusters are disjoint. When applied to genomes, hierarchical clustering produces a biological taxonomy, which helps us to make sense of the enormous diversity of living organisms. In any organism, there are many different kinds of features to choose from, and in principle, all of them can be used. Unsupervised learning is one of the main strengths of the hierarchical clustering methodology, and its high performance becomes even more significant when compared to some supervised methods.

Similarity Measure

Typically, the similarity between documents is estimated by a function calculating the distance between the vectors of these documents. Two documents that are close according to this distance are regarded as similar. Several measures of similarity have been proposed (Jones and Furnas, 1987), including the following:

Cosine distance:

$$\cos(d_i, d_j) = \frac{\sum_{t_k}[TF \times IDF(t_k, d_i)] \cdot [TF \times IDF(t_k, d_i)]}{\|d_i\|^2 \cdot \|d_j\|^2}$$

Euclidean distance:

$$Euclidean(d_i, d_j) = \sqrt{\sum_{\mathrm{I}}^{\mathrm{n}}(w_{ki}, w_{kj})^2}$$

Manhattan distance:

$$Manhattan(d_i, d_j) = \sum_{\mathrm{I}}^{\mathrm{n}}|w_{ki}, w_{kj}|$$

The main purpose of this essay is to cluster clinical trials with semantic based feature expanded subject eligibility criteria. There are a number of clustering models based on connectivity, centroid, distribution, and other characteristics. In this experiment, the agglomerative hierarchical clustering model was adopted because it could show all the merging steps in the clustering process. To measure similarity between clinical trial subject eligibility, I adopted the cosine distance, which is one of the popular metrics for text analysis.

3.5. Results

Before conducting the hierarchical clustering analysis, the scatter score for all clusters was calculated to determine the optimal number of clusters. Scatter score measures the degree of within-cluster scatter for the specified clusterings with the specified distance. The within-cluster scatter is simply the sum of the scatters for each set in the clustering. As the number of clusters increases, the within-cluster scatter decreases monotonically. Typically, this is used to determine how many clusters to generate by inspecting a plot of within-cluster scatter against the number of clusters and looking for a "knee" in the graph.

Figure 9 shows the scatter score for the all inclusion criteria set and the "knee" point of the graph, which is 156. Therefore, the optimal number of inclusion criteria clusters is 156.

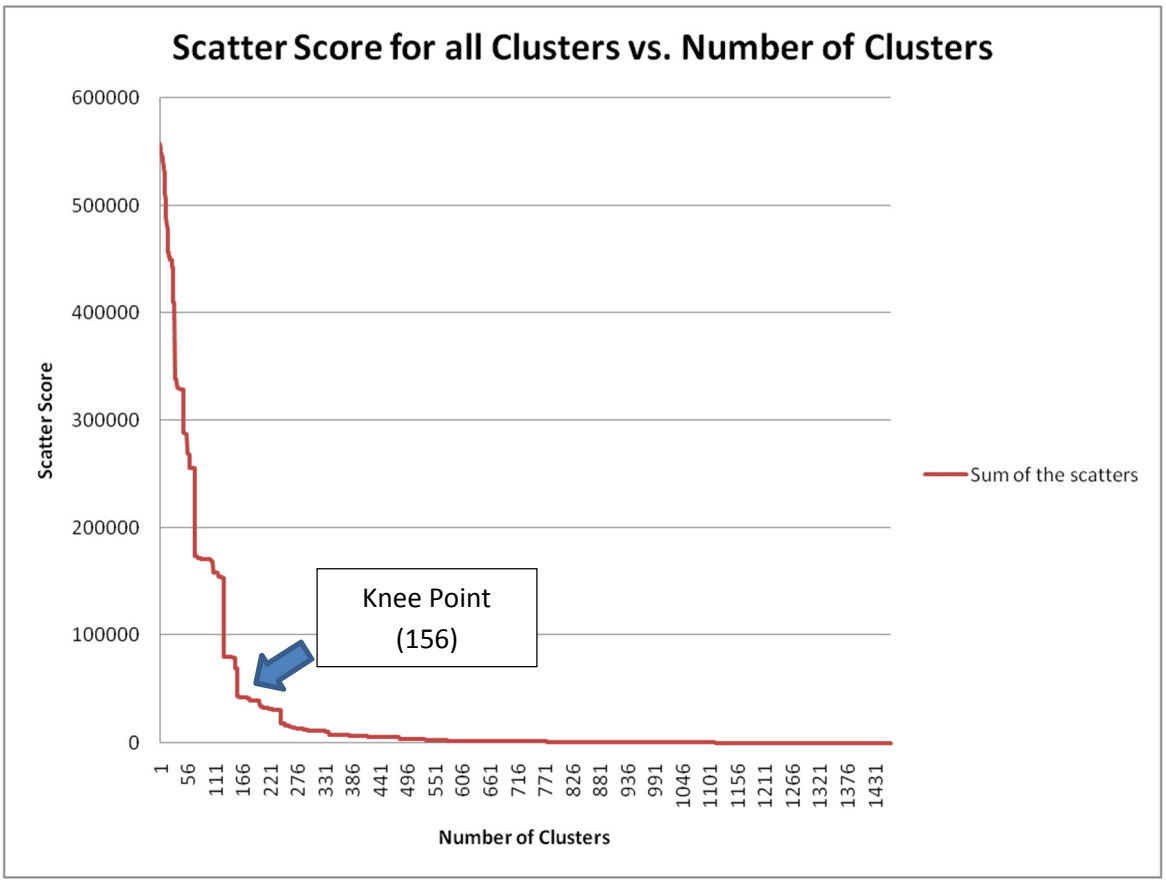

Figure 9: Scatter Score for All Inclusion Criteria Clusters

Figure 10 shows the scatter score for the all exclusion criteria set and the ‘knee’ point of the graph, which is 168. Therefore, the optimal number of inclusion criteria clusters is 168.

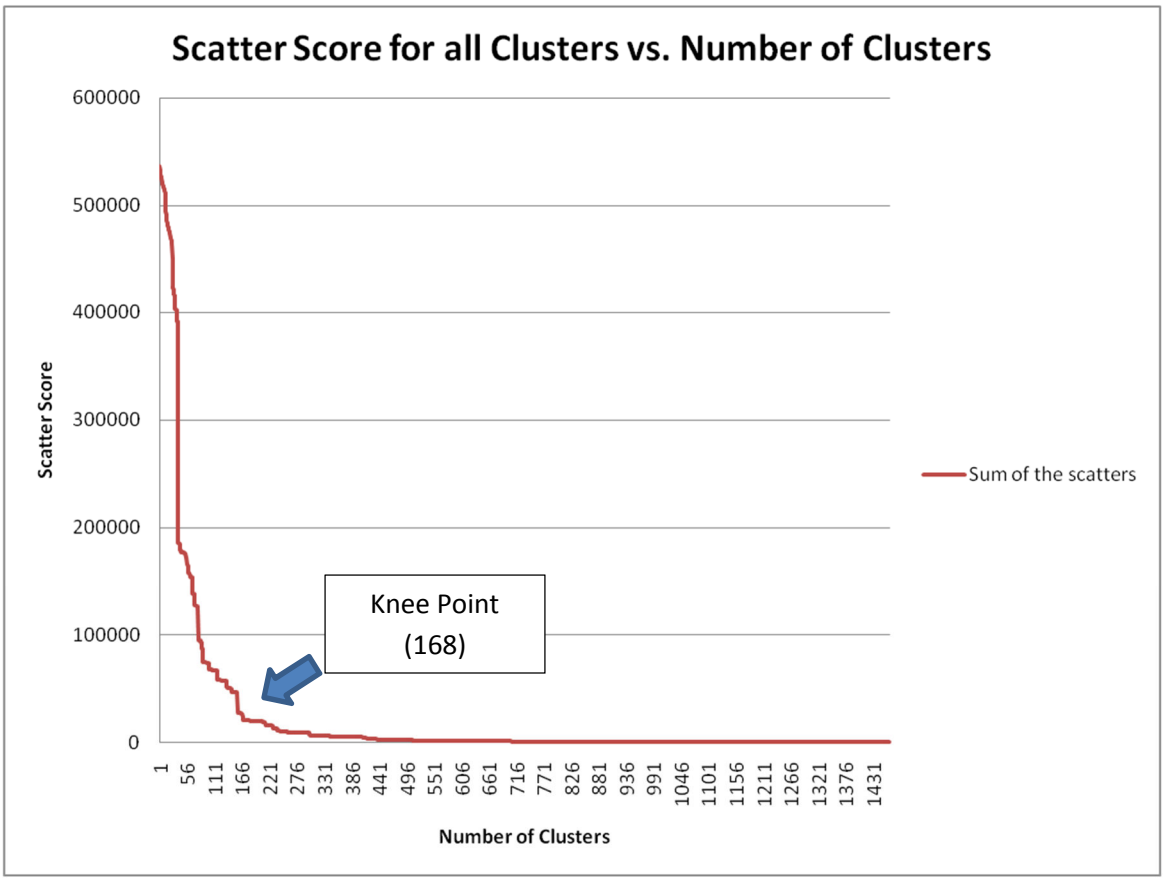

Figure 10: Scatter Score for All Exclusion Criteria Clusters

Based on the scatter score analysis, I generated 156 clusters for inclusion criteria and 168 clusters for exclusion criteria. Figure 11 shows the sample of two clinical trials inclusion criteria (NCT01642511 and NCT01668914) that clustered together at a low level because the similarity score is 1.0. Table 18 shows the original eligibility inclusion criteria of NCT01642511 and NCT01668914.

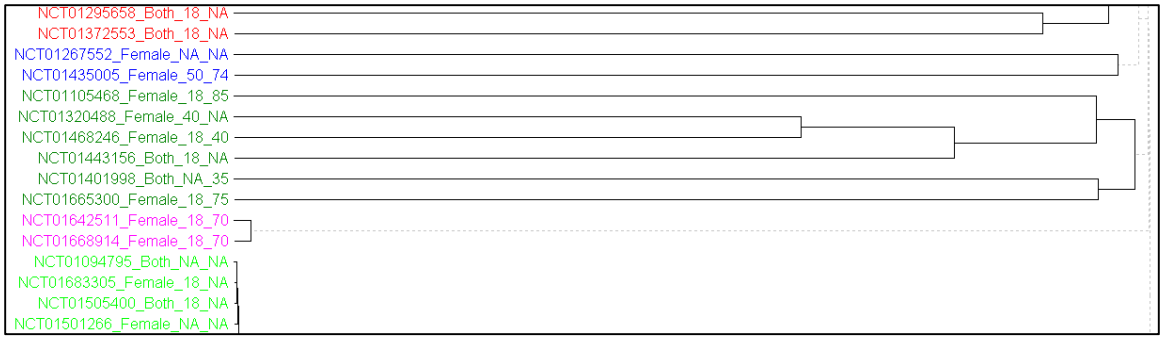

Figure 11. Tree of Hierarchical Clustering for NCT01642511 and NCT01668914

Table 18. Original Text of Two Clinical Trials (NCT01642511 and NCT01668914)

CT ID	NCT01642511	NCT01668914
	Similarity Score =1.0	
Original Text	- enlarged internal mammary nodes by imaging	- enlarged internal mammary nodes by imaging

Figure 12 shows the sample of two clinical trials exclusion criteria (NCT01510964 and NCT01691144) that merged at a high level because the similarity score was 0.56. Table 19 shows the original eligibility exclusion criteria of NCT01510964 and NCT01691144.

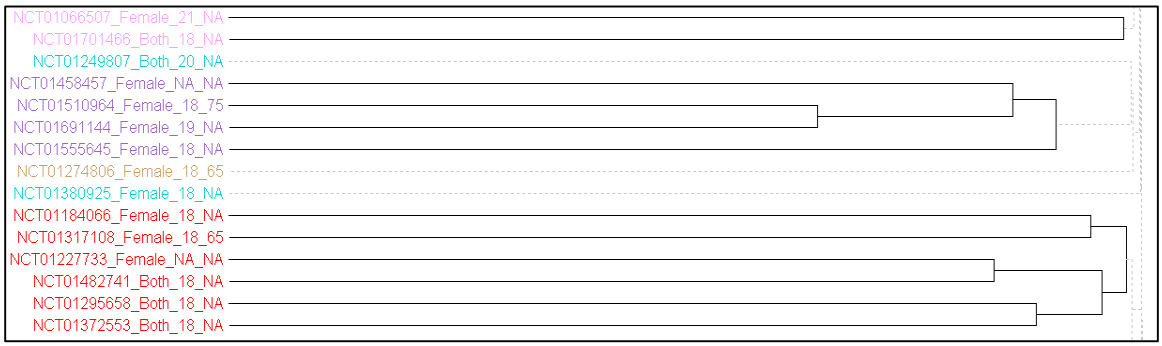

Figure 12. Tree of Hierarchical Clustering for Exclusion Criteria (NCT01510964 and NCT01691144)

Table 19: Original Text of Two Clinical Trials for Exclusion Criteria (NCT01510964 and NCT01691144)

CT ID	NCT01510964	NCT01691144
	Similarity Score =0.56	
Original Text	- presence of metastasis or relapse - severe mental deterioration - comprehension difficulties of the Italian language.	- Unability to fill out questionnaires (due to language or cognitive barriers)

Table 20 shows other examples of case comparison by the inclusion criteria similarity score.

Table 20. Example of Case Comparison

Case No.	Contents	Case No.	Contents	Score
NCT01619306	Inclusion Criteria: - Patients with early stage breast cancer - Healthcare professionals caring for breast cancer patients - Medical students /cancer researchers	NCT01619514	Inclusion Criteria: - Patients with breast cancer	0.88
		NCT01506869	Phase 1 Inclusion Criteria: 1. Age >= 40 years old 2. Gender: males and females 3. Provide written informed consent 4. Satisfactory compliance Phase 2 Inclusion Criteria: 1. Age >= 40 and =< 75 years old 2. Gender: males and females 3. Provide written informed consent 4. Satisfactory compliance Exclusion Criteria: 1. History of cancer; 2. History of LADA and other autoimmunity diseases; 3. Acute diabetic complication, acidosis, etc; 4. Moderate to severe liver, kidney dysfunction, i.e. ALT/AST > 2.5 times the upper limit of normal range or Ccr < 25ml/min; 5. Any other condition or major systemic diseases that the investigator feels would interfere with trial participation or evaluation of results.	0.09

NCT01526499	Inclusion Criteria: 1. Females with age between 18 and 70 years old 2. ECOG performance between 0-1 3. Life expectancy more than 3 months 4. Histological proven unresectable recurrent or advanced breast cancer 5. No previous chemotherapy for metastatic breast cancer;suitable for monotherapy (Neoadjuvant or adjuvant docetaxel should be completed at least one year). 6. At least one measurable disease according to the response evaluation criteria in solid tumor (RECIST1.1) 7. No anticancer therapy within 4 weeks 8. Adequate hematologic, hepatic, and renal function,No serious medical history of heart, lung, liver and kidney 9. Provision of written informed consent prior to any study specific procedures Exclusion Criteria: 1. Pregnant or lactating women (female patients of child-bearing potential must have a negative serum pregnancy test within 14 days of first day of drug dosing, or, if positive, a pregnancy ruled out by ultrasound) 2. Women of child-bearing potential, unwilling to use adequate contraceptive protection during the course of the study 3. Treatment with an investigational product within 4 weeks before the first treatment 4. Symptomatic central nervous system metastases	NCT01526512	Inclusion Criteria: 1. Females with age between 18 and 80 years old 2. ECOG performance between 0-3 3. Life expectancy more than 3 months 4. Histological proven unresectable recurrent or advanced HER2-negative breast cancer 5. At least one previous therapy regimen (including endocrine therapy) for metastatic breast cancer;suitable for monotherapy (Neoadjuvant or adjuvant docetaxel should be completed at least one year). 6. At least one measurable disease according to the response evaluation criteria in solid tumor (RECIST1.1) 7. No anticancer therapy within 4 weeks 8. Adequate hematologic, hepatic, and renal function,No serious medical history of heart, lung, liver and kidney 9. Provision of written informed consent prior to any study specific procedures 10. Previous capecitabine is permitted, however, it should be completed at least 6 months. Exclusion Criteria: 1. Pregnant or lactating women (female patients of child-bearing potential must have a negative serum pregnancy test within 14 days of first day of drug dosing, or, if positive, a pregnancy ruled out by ultrasound) 2. Women of child-bearing potential, unwilling to use adequate contraceptive protection	0.97

	5. Other active malignancies (including other hematologic malignancies) or other malignancies, except for cured nonmelanoma skin cancer or cervical intraepithelial neoplasia. 6. Patient having a history of clinically significant cardiovascular, hepatic, respiratory or renal diseases, clinically significant hematological and endocrinal abnormalities, clinically significant neurological or psychiatric conditions 7. Uncontrolled serious infection 8. Patients with bad compliance		during the course of the study 3. Treatment with an investigational product within 4 weeks before the first treatment 4. Symptomatic central nervous system metastases 5. Other active malignancies (including other hematologic malignancies) or other malignancies, except for cured nonmelanoma skin cancer or cervical intraepithelial neoplasia. 6. Patient having a history of clinically significant cardiovascular, hepatic, respiratory or renal diseases, clinically significant hematological and endocrinal abnormalities, clinically significant neurological or psychiatric conditions 7. Uncontrolled serious infection 8. Patients with bad compliance 9. Patients lack of Dihydropyrimidine Dehydrogenase(DPD)	

		NCT01569802	Inclusion Criteria: - Subject is female of any race and ethnicity - The subject is asymptomatic and presents for routine screening mammography and chooses to have a combination 2D + 3D mammogram as her standard of care. Exclusion Criteria: - Patient chooses standard 2D mammography over a combination 2D + 3D mammogram	0.07
NCT01558258	Inclusion Criteria: - women diagnosed with early, resectable breast cancer (Stage 0, I, II, or III) prior to age 50 - have completed treatment with surgery, radiation, and/or chemotherapy at least 3 months previously. Exclusion Criteria: - have a breast cancer recurrence, metastasis, or another cancer diagnosis (excluding non-melanoma skin cancer - unable to commit to intervention schedule.	NCT01627366	Inclusion Criteria: - Female - 21 years of age or older - English- or Spanish-speaking - Diagnosis of ductal carcinoma in situ (DCIS) or Stage I, II, or III BC for the first time - 12 months post-diagnosis - At least 1 month post-chemotherapy completion Exclusion Criteria: - Previous cancer except non-melanomatous skin cancers or in situ non-breast cancers - Pregnant and lactating women - Patients receiving parenteral anti-cancer therapy, except trastuzumab - Clinically apparent cognitive or psychiatric impairment -	0.59

			Participation in another research study - Current treatment for another cancer - Male	
		NCT01569802	Inclusion Criteria: - Subject is female of any race and ethnicity - The subject is asymptomatic and presents for routine screening mammography and chooses to have a combination 2D + 3D mammogram as her standard of care. Exclusion Criteria: - Patient chooses standard 2D mammography over a combination 2D + 3D mammogram	0.01

3.5.1. Intersection of inclusion and exclusion clusters

Since the inclusion and exclusion subject eligibility criteria were mutually exclusive, the eligibility criteria section was divided into two sub-sections: inclusion criteria and exclusion criteria. The two data sets were pre-processed, matched with the custom dictionary and UMLS Metathesaurus, and clustered individually. However, to achieve completed clinical trial subject eligibility clusters, it was necessary to merge the two different cluster sets.

All the elements in each inclusion cluster were compared with all the elements in each exclusion cluster and new clusters were generated based on only the elements belonging to the same inclusion and exclusion clusters. Figure 13 presents an example of new cluster generation. For instance, clinical trials A, B, C, and D belong to the inclusion cluster Inc-I, and clinical trials A, B, E, D, and F belong to the exclusion cluster Exc-I. From this example, the new cluster Inc-I is created that includes only the common elements of the inclusion cluster Inc-I and the exclusion cluster Exc-I. More specifically, when we assume that one of the criteria in the inclusion cluster Inc-I is subject's pregnancy and one of the criteria in the exclusion cluster Exc-I is subject's breast feeding, the intersection cluster of the inclusion cluster Inc-I and the exclusion cluster Exc-I will have the eligibility criteria that include subjects who are pregnant but exclude those who are breastfeeding.

As mentioned before, the total number of inclusion clusters was 156 and the total number of exclusion clusters was 168. From these clusters, 596 intersection clusters were generated. Accordingly, the number of intersection clusters that had more than two instances was 117, and the number of intersection clusters with two or less than two

instances was 479. Table 21 presents the number of intersection clusters. The name for an intersection cluster was assigned by combining the ID of the inclusion cluster and the exclusion cluster. For example, the intersection cluster Inc(16)_Exc(130) had clinical trials that appeared in both inclusion cluster(16) and exclusion cluster(13).

Table 21. Number of Intersectional Clusters

	Number of Clusters
Total number of intersection clusters	596
Number of single-instance clusters	393
Number of two-instance clusters	86
Number intersection clusters having more than two instances	117

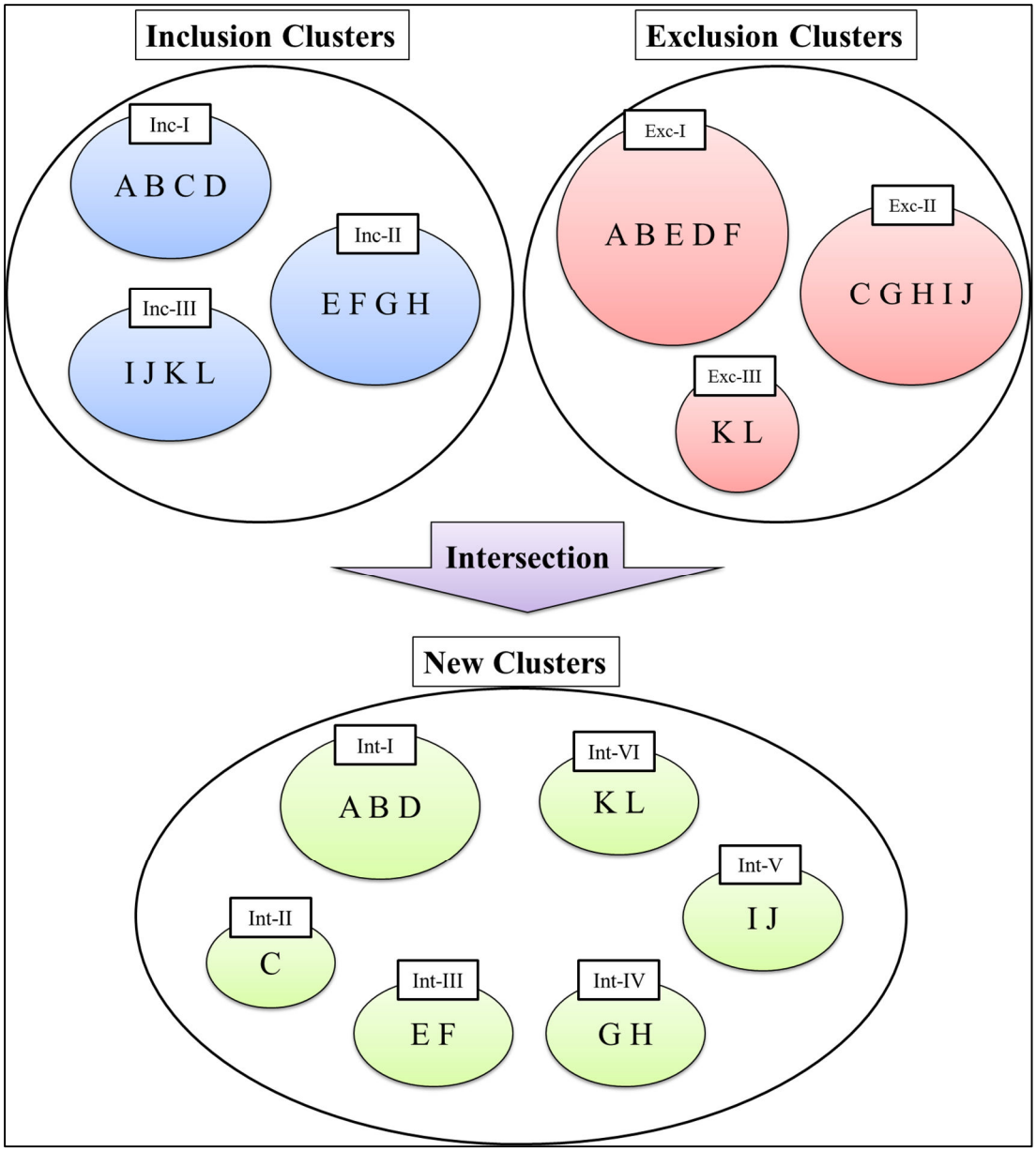

Figure 13. Intersection Clusters of the Inclusion and Exclusion Clusters

3.6. Cluster labeling

3.6.1. UMLS Synonym chunks

Identifying most frequent synonym chunks in clusters

To identify the characteristics of clusters, I generated a label for each cluster. First, I counted all the synonym chunks that were used for the semantic feature expansion in inclusion and exclusion criteria from same intersection clusters. The most frequent synonym chunk of the inclusion and exclusion clusters was selected as the representative label for the intersection cluster. If the most frequent synonym chunk in the inclusion or exclusion cluster had been already selected for another intersection cluster, the second most frequent synonym chunk was selected. If multiple most frequent synonym chunks with the same frequency are found, all of the synonym chunks in the top frequency were selected for labeling.

Next, I queried UMLS Metathesaurus with selected inclusion and exclusion synonym chunks to find the lowest concept unique identifier among synonyms. The CUI in UMLS is the concept unique identifier for a UMLS Metathesaurus concept to which strings with the same meaning are linked. The synonyms in chunks has their own CUI. To find unique name of each synonym chunk, I used lowest CUI in each synonym chunk.

Then, all the lowest CUI concepts from UMLS Metathesaurus are merged and the '|' symbol is added between concepts as a delimiter. Also, '||' is added to divide inclusion and exclusion synonym chunks.

Table 22 presents the proposed label for the intersection cluster Inc(16)_Exc(130). For the Inc(16)_ExC(13) cluster, the lowest CUI of the most frequent

synonym chunk in inclusion section is C0013216 and the note for C0013216 are ‘Chemotherapy’ and ‘Drug therapy’. The lowest CUI of the most frequent synonym chunk in exclusion section is C0006141 and the notes for C0006141 are ‘Breast anatomy’, ‘Breast’ and ‘Breast structure’. The proposed label for Inc(16)_Exc(130) is merging the notes for these two sections, ‘CT - Chemotherapy|Drug therapy|Chemotherapy|DT - Drug therapy||Breast anatomy|Breast|Breast structure’. The proposed label means that representative inclusion criteria for intersection cluster Inc(16)_Exc(130) are ‘Chemotherapy’ and ‘Drug therapy’. The representative exclusion criteria for intersection cluster Inc(16)_Exc(130) are ‘Breast anatomy’, ‘Breast’ and ‘Breast structure’. All the trials in intersection cluster Inc(16)_Exc(130) require ‘Chemotherapy’ or ‘Drug therapy’ experience for patients as inclusion criteria, and the patient experienced ‘Breast anatomy’ should be excluded all trials in intersection cluster Inc(16)_Exc(130).

Table 22. Proposed label for the cluster Inc(16)_Exc(130)

Intersection Cluster ID	Label
Inc(16)_Exc(130)	CT - Chemotherapy|Drug therapy|Chemotherapy|DT - Drug therapy||Breast anatomy|Breast|Breast structure

Table 23 shows the algorithm to generate labels for the intersectional clusters. First, all the UMLS synonym chunks were counted for all the inclusion and exclusion clusters. Second, the most frequent synonym chunk in each inclusion and exclusion cluster was selected as a candidate for label. If there were more than two synonym chunks in one cluster in the same frequency, all the synonym chunks were selected. If the

most frequent synonym chunk had been already selected for another cluster, the second most frequent synonym chunk was the selected candidate. Third, the synonym chunk for inclusion and exclusion was merged to generate the full label. If inclusion and exclusion synonym chunks were the same, I selected the second most frequent synonym chunk for the exclusion cluster.

Table 23. Pseudo Code for generating cluster label

```
Function GenerateIntersectionClusterLabel(ClusterID) : returns ClusterLabel
Begin
        Set ClusterLabel to null
        Set ClinicalTrials to null
        Set SynonymInclusion to null
        Set SynonymExclusion to null
        Set MostFrequentSynonymsInInclusion
        Set MostFrequentSynonymsInExclusion
        Set SynonymCount to 0
        Create Queue, Q
        Query DB AllClinicalTrials in ClusterID
        Add AllClinicalTrials to Q
        While Q is not empty
                De-queue AllClinicalTrials CT from Q
                For each CT in AllClinicalTrials
                        Query DB Synonym in CTInclusion
                        Count SynonymInclusion
                        For each Synonym in CTInclusion
                                If CountIncSynonym >= MostFrequentSynonymsInInclusion
                                and LowesetCUIIncSynonym is not exist in Label list
                                        MostFrequentSynonymsInInclusion = SynonymInclusion
                                        Qeury DB newLowestCUI in UMLS
                                        Add ClusterLabel
                                Else If
                                        Break
                                End If
                        Next
                        Query DB Synonym in CTExclusion
                        Count SynonymExclusion
                        For each Synonym in CTExclusion
                                If CountExcSynonym >= MostFrequentSynonymsInExclusion
                                and LowesetCUIExcSynonym is not exist in Label list
```

```
                                        MostFrequentSynonymsInExculision = SynonymExclusion
                                        Qeury DB newLowestCUI in UMLS
                                        Add ClusterLabel
                                Else If
                                        Break
                                End IF
                        Next
                Next
        End While
        Return ClusterLabel
End
```

The clinical trial NCT01202851 belongs to the intersection cluster Inc(16)_Exc(130). Table 24 presents the original text of the subject eligibility section in NCT01202851; it has 'adjuvant radiation' in the inclusion criteria and 'surgical treatment' in the exclusion criteria, both corresponding to the proposed label.

Table 24. Subject Eligibility of NCT01202851

CT ID	Subject Eligibility Text
NCT01202851	Inclusion Criteria: 1. Women with stage 0 - III breast cancer who will be undergoing daily ***adjuvant radiation*** for 4-6 weeks (patients only). 2. 18 years of age or older (patient and spouse/partner). 3. Able to read, write, and speak English or Spanish (patient and spouse/partner). Exclusion Criteria: 1. Patients who have any major psychiatric diagnoses (e.g., schizophrenia, bipolar disorder). 2. Patients who have not undergone any ***surgical treatment*** for their cancer. 3. Patients with extreme mobility issues (e.g., unable to get in and out of a chair unassisted). 4. Patients who have practiced yoga or taken yoga classes in the year prior to study enrollment or who are currently engaged in a regular mind-body practice

3.7. Discussion

The broad objective of this work was to group and summarize clinical trial subject eligibility using a hierarchical clustering approach. This essay has also presented a framework for clustering clinical trial and labeling clusters.

In this research, I examined 1,660 breast cancer clinical trials and derived 596 intersectional clusters. Also, I generated a label for each cluster to identify the characteristics of the cluster. The full text information of clinical trial studies were collected from ClinicalTrials.gov, and the original XML format documents were parsed with the author-developed parser. The subject eligibility section was extracted from the parsed documents and it was divided into two data sets for inclusion and exclusion criteria.

The agglomerative hierarchical clustering algorithm with cosine similarity metric was used to generate two sets of clusters, one for inclusion and the other for exclusion criteria; sets and intersection clusters were derived from those two cluster sets. The cluster labels were generated based on the most frequent UMLS synonym chunks in each inclusion and exclusion cluster to understand the characteristics of clusters.

While healthcare and IS researchers have made substantial progress in clustering clinical trial subject eligibility, little has been done to examine the semantic feature expansion technique in the healthcare domain and the contrary characteristics of inclusion and exclusion criteria. Inclusion criteria should be found and exclusion criterial should not be found in patient records.

This essay has also made practical contributions by providing groups of similar clinical trials that can reduce a physician's search space to find relevant trials to help clinical trial research as well as to provide alternative treatment to terminal disease patients.

The clusters can also be utilized by initiators of new clinical trial study for finding similar trials currently in progress. When a primary investigator starts a new clinical research study, he or she is required to review all the relevant prior clinical studies. The clusters from this study can reduce the cost and effort for future clinical trial researchers by providing clinical trial clusters that have been labeled with the main features. Furthermore, the total number of clinical trials is increasing, and research in the healthcare domain is becoming more competitive. A clinical study usually requires a huge amount of resources with respect to financial support, expert involvement, and subject participation. Therefore, repeating the same type of clinical study should be avoided; each study should have its unique contributions. The clusters from this research could be exploited for finding research on similar topics and help to screen research topics that have been already conducted by other researchers. Furthermore, when a new trial study is proposed, the primary investigator usually estimates the required number of subjects. The cluster information can provide a similar trial group, and the primary investigator can use that to identify other trials that are looking for similar patients. In this vein, clinical trial cluster information enables for researchers to estimate probability of successful recruitment of required number of participants.

There are several ways in which future research could strengthen the results of this study. First, this research was confined to the breast cancer domain. Future studies

could investigate the proposed framework in the context of different kinds of diseases. I used the hierarchical clustering algorithm and applied the cosine theta as the document similarity metric. However, prior studies have proposed different approaches for clustering and document similarity metrics. For example, Latent Dirichlet allocation (LDA), latent semantic indexing, independent component analysis, probabilistic latent semantic indexing, non-negative matrix factorization, and Gamma-Poisson distribution techniques have been used in bioinformatics research. These new techniques could be applied in future research.

CHAPTER 4

Essay 3: Automatic Matching Process of Clinical Trials Subject Recruitment

"With enough information, it is almost impossible 'not' to predict people's action."

Idries Shah

4.1. Introduction

About 85% of people with cancer were either unaware or unsure that participation in clinical trials was an option, although about 75% of these people said they would have been willing to enroll had they known it was possible (Harris Interactive, 2001). Previous research by UC Davis Cancer Center (UC Davis Cancer Center, 2001) investigators, published in 2001 in the *Journal of Clinical Oncology*, found that both doctors and patients sometimes hold misconceptions that can discourage enrollment in clinical trials. In the UC Davis Cancer Center study, more than one third of the doctors declined to refer patients to clinical trials, mistakenly believing that no trials were available. In reality, more than 150 clinical trials were available during the study period.

Another common barrier of clinical trial participation is distrust or suspicion about research. This is despite the fact that many investigational treatments are at least as effective as conventional therapy, and cancer patients who participate in clinical trials frequently have higher survival rates than those who receive standard care (UC Davis Cancer Center, 2001). Because of this unwarranted distrust or suspicion, four out of five clinical trials are delayed, and 50% of the delays are due to participant recruitment challenges (Patel et al., 2010).

These low rates represent a significant barrier to speeding progress in cancer treatment by delaying the dissemination of new therapies. Low participation in clinical trials is a critical issue in healthcare research, where participation rates range between 5% and 10% for most trials (Patel et al., 2007). In the domain of oncology, for example, fewer than 3% of potentially eligible patients enroll in clinical trials, and patient enrollment for clinical trials is as low as 2% of the patient recruitment goal (Embi et al., 2005).

Although the cost of running trials is now approaching 30% of pharmaceutical companies' entire drug development budgets, 75% of patient studies fail to make their timelines, often causing expensive delays in regulatory approval and market launch. Also, testing on humans is a sensitive and a difficult issue as it involves many legal and ethical issues. Difficulties in patient recruitment are the major reason for failure of clinical research (Spilker and Cramer, 1992).

Low and slow recruitment has serious negative impacts on the translation of the clinical trial results. It could produce inadequate statistical analyses of outcomes, lead to premature closure of trials, delay trial duration, incur higher costs of drug production, and cause loss of accreditation of the research institution that performs these studies (Penberthy et al. 2010).

Patient and physician factors can also be barriers to the enrollment in clinical trials (Breitfeld et al. 1999). Patient factors include lack of access to a healthcare institute offering clinical trials, economic and social barriers, and attitudes and beliefs. Among the diverse reasons physicians may fail to offer clinical trial participation to patients is lack of time. For example, to determine whether new patients may be eligible for a clinical

trial, physicians need to search multiple clinical trial repositories and read through the eligibility sections of several protocol documents. Physicians who participate in a busy oncology practice may find that they do not have sufficient time to do this and identify eligible subjects efficiently. Their lack of time for these activities, which may interrupt the flow of patients, constitutes a substantial barrier to trial enrollment. They may also simply forget to offer and enroll patients in possible trials.

Determining the eligibility of every patient is the first step in assuring adequate and unbiased clinical trial research. Yet, not all eligible patients are evaluated or invited to participate in a clinical trial despite the fact that patients who are offered a trial are likely to participate (Albrecht, 2008). One of the major impediments to participation is that this process of matching a patient to a clinical trial is manual and physician-driven. Traditionally, in this process, clinical trial research staff manually review multiple clinical data sources from patient medical records and match them with subject eligibility criteria. Eligible patients are often missed by this manual review process (Penberthy, 2010). Thus, helping identify potentially eligible subjects increases the likelihood of patient participation in a clinical trial and is critical to the issue of under-representation.

In this essay, I propose a novel framework for clinical trial subject recruitment using NLP and text mining techniques for automating the clinical trial and subject matching process, which is currently labor intensive and error prone. The proposed approach could be very helpful for expediting and improving the clinical trial subject recruitment process.

The rest of the chapter is organized as follows. The literature review that serves as the overview of research stream in patient and clinical trial matching is presented in

section 4.2. Document similarity measurement techniques underlying the patient and clinical trial matching process, as well as the entire research framework, are presented in sections 4.3 and 4.4. Section 4.5 presents the matching and evaluation results. Section 4.6 discusses the implications, limitations, and future directions of this research.

4.2. Background Literature

Table 25 presents the selected research on matching clinical trials and patient information.

Table 25. Selected research on matching clinical trials and patient information

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Patel et al.	IBM Research (2007)	Matching Patient Records to Clinical Trials Using Ontologies	Case study for clinical trial subject selection using ontologies and semantic technology		SNOMED CT, One year patient data from Columbia Medical Center	
Patel et al.	Elsevier (2012)	TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records	TrialX, a consumer-centric tool that matches patients to clinical trials			
Embi et al.	American Medical Association (2005)	Effect of a clinical trial alert system on physician participation in trial recruitment	Evaluation of electronic health record based clinical trial alert system		4 month intervention with 114 physicians	The CTA intervention was associated with significant increases in physicians' referrals and enrollments
Breitfeld et al.	Journal of the American Medical Informatics Association (1999)	Pilot Study of a Point-of-use Decision Support Tool for Cancer Clinical Trials Eligibility	Development of point-of-use portable decision support tool (DS-TRIEL) to automate this matching process		pilot-test with academic medical oncologist	

Authors	Journal	Title	Topic / Research Question	Theory/ Model	Data	Finding / Implication
Brigitte Séroussi and Jacques Bouaud	Artificial Intelligence in Medicine (2003).	Using OncoDoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials	Development of OncoDoc decision support system designed to provide best therapeutic recommendations for breast cancer patients			Significantly improved physician compliance and enhanced physician awareness of open trials.
Penberthy et al.	Contemporary Clinical Trials (2010)	Automated matching software for clinical trials eligibility: Measuring efficiency and flexibility	A pilot project evaluating the efficiency, flexibility, and generalizability of an automated clinical trials eligibility screening tool		5 different clinical trials and clinical trial scenarios.	Automation offers an opportunity to reduce the burden of the manual processes required for CT eligibility screening
Fink et al.	Artificial Intelligence in Medicine (2004)	Selection of patients for clinical trials: an interactive web-based system	Development of a web-based expert system that assigns cancer patients to clinical trials		187 past patients and 74 current patients for Knowledge base 261 breast-cancer patients for test	
Korkontzelos et al.	BMC Medical Informatics and Decision Making (2012)	ASCOT: a text mining-based web-service for efficient search and assisted creation of clinical trials	ASCOT, clinical trial search and creation tool.		1800 clinical trials	

There has been limited research on the topic of clinical trial and patient matching process. Korkontzelos et al. (2012) presented ASCOT (Assisting Search and Creation Of Clinical Trials), a clinical trial search application that employs text mining technology, clustering, and term extraction algorithms.

Patel et al. (2010) published an article on the clinical trial and patient matching process. In that paper, they introduced TrialX, a consumer-centric tool that matches patients to clinical trials by extracting Personal Health Records (PHR) from Microsoft HealthVault (MHV) and Google Health (GM), and linking patients to the most relevant clinical trials using semantic web technologies.

Penberthy et al. (2010) evaluated the efficiency, flexibility, and generalizability of a clinical trials eligibility screening tool with five different clinical trials. The results of their study demonstrated that the automated tool could reduce the time and cost of the manual processes required for clinical trial eligibility screening and assure clinical trial participation opportunity. During the study period in evaluating patients for eligibility by research staff, there was a substantial total savings ranging from 165 hours to 1,329 hours. The ratio of mean staff time for identifying eligible patients ranged from 0.8 to 19.4 for the manual versus the automated process.

In 2007, Patel et al. tried to formulate the clinical trial and patient matching process as a problem of semantic retrieval. They focused on the applicability of SNOMED CT ontologies, which define classes of disorders, drugs, and organisms. The case study, conducted with one year of anonymized patient records from Columbia University Medical Center, reported that using an ontology to automate the matching process is feasible and practical. However, that research focused only on the ontology-

based mapping. No text mining or NLP techniques were examined for the matching process.

Embi et al. (2005) investigated the effects of an electronic health record (EHR)-based clinical trial alert (CTA) system in selected outpatient clinics of a large US academic healthcare system. CTA was tested during the subsequent 4-month intervention period when a patient's EHR data met selected trial criteria. One hundred fourteen physicians practicing at selected EHR-equipped clinics participated in the study. The researchers compared the number of physicians participating in recruitment and their recruitment rates before and after CTA intervention. The results of the study showed that CTA intervention was associated with significant increases in the number of physician referrals and enrollment. However, Embi et al.'s research only focused on the evaluation of CTA intervention, and the clinical trial eligibility matching was conducted by the trial's principal investigator.

Breitfeld et al. (1999) developed a point-of-use portable decision support tool (DS-TRIEL) to automate the matching process. A two-level hierarchic decision framework was used for the identification of eligible subjects for two open breast cancer clinical trials.

Séroussi and Roland (1998) developed OncoDoc, which is a decision support system designed to provide the best therapeutic recommendations for breast cancer patients. OncoDoc is a browsing tool of a knowledge base, structured as a decision tree, which allows physicians to control the contextual instantiation of patient characteristics to build the best formal equivalent of an actual patient. It provides either evidence-based therapeutic options or relevant patient-specific clinical trials.

Fink et al (2004) developed a rule-based expert system that helps assign patients to clinical trials. The experiment results showed that their system can increase the efficiency of the patient selection process.

There have been several research studies that developed expert systems for helping select clinical trials for cancer and AIDS patients. Musen et al. (1996) built a rule-based system, called EON, that matched AIDS patients to clinical trials.

Ohno-Machado et al. developed the AIDS2 system, which matched AIDS patients to clinical trials (Ohno-Machado et al., 1993). The integrated logical rules with Bayesian networks was used for the AIDS2 system, and the system helped decision-making with incomplete data and to quantify the decision quality.

Bouaud et al. created ONCODOC, a cancer expert system that suggested alternative clinical trials and allowed a physician to choose one of the alternatives (Bouaud et al 1998, 2000). S´eroussi et al. used ONCODOC to evaluate usefulness of the system at two hospitals and found that ONCODOC helped increase the number of matched patients (S´eroussi et al. 1999, 2001)

Hammond and Sergot (1996) developed OaSiS, which has a graphical interface for entering patient data and extending the knowledge base. Papaconstantinou et al. (1998) developed a Bayesian system that selected clinical trials for cancer patients (Papaconstantinou et al., 1998, Theocharous et al. 1996). Their system learned conditional probabilities of medical test outcomes and evaluated the probability of a patient's eligibility for each trial. Learning accurate probabilities requires sufficient medical records, but the available medical records were limited in volume. Moreover, the underlying Bayesian network needs to be modified when a new clinical trial is added.

Fallowfield et al. investigated physicians' cancer patient selection process for clinical trials, and compared manual and automatic selection (Fallowfield et al. 1997). Their study showed that expert systems could improve clinical trial patient selection accuracy. Carlson et al. (1995) conducted research on AIDS trials and showed that expert systems could improve patient selection.

In this section, I reviewed selected research on clinical trial and patient matching processes as well as decision support systems for clinical trial subject recruitment. Only a few attempts have so far been made on using NLP and text mining techniques. However, these studies only used basic level techniques or vaguely described the research process. In this essay, I propose a novel approach for the clinical trial and patient matching process using state-of-art NLP and text mining techniques.

4.3. Background

Document Similarity Measurement

In the text mining and NLP fields, text similarity measurement plays an increasingly significant role. It measures the similarity between words, sentences, paragraphs, and documents. It is also an important component in tasks such as information retrieval, text classification, document clustering, topic detection, text summarization, word-sense disambiguation, automatic grading, and machine translation (Gomaa and Fahmy 2013). Over the past few decades, a large number of studies on measuring text and document similarity were conducted. Gomaa and Fahmy (2013) partitioned this issue into three approaches: string-based, corpus-based, and knowledge-based.

There are two different types of similarity in words: lexical and semantic similarity. If two words have a similar character sequence, these two words are similar lexically. If two words have the same meaning or are used in the same context or the same way, they are similar semantically. String-based algorithms are used for lexical similarity, while corpus-based and knowledge-based algorithms are used for semantic similarity.

String-Based Similarity Measures

A string metric measures similarity or dissimilarity (distance) between two text strings for string matching or comparison. Figure 14 shows 14 algorithms of string-based similarity measures; seven of them are character-based measures, while the other seven are term-based distance measures.

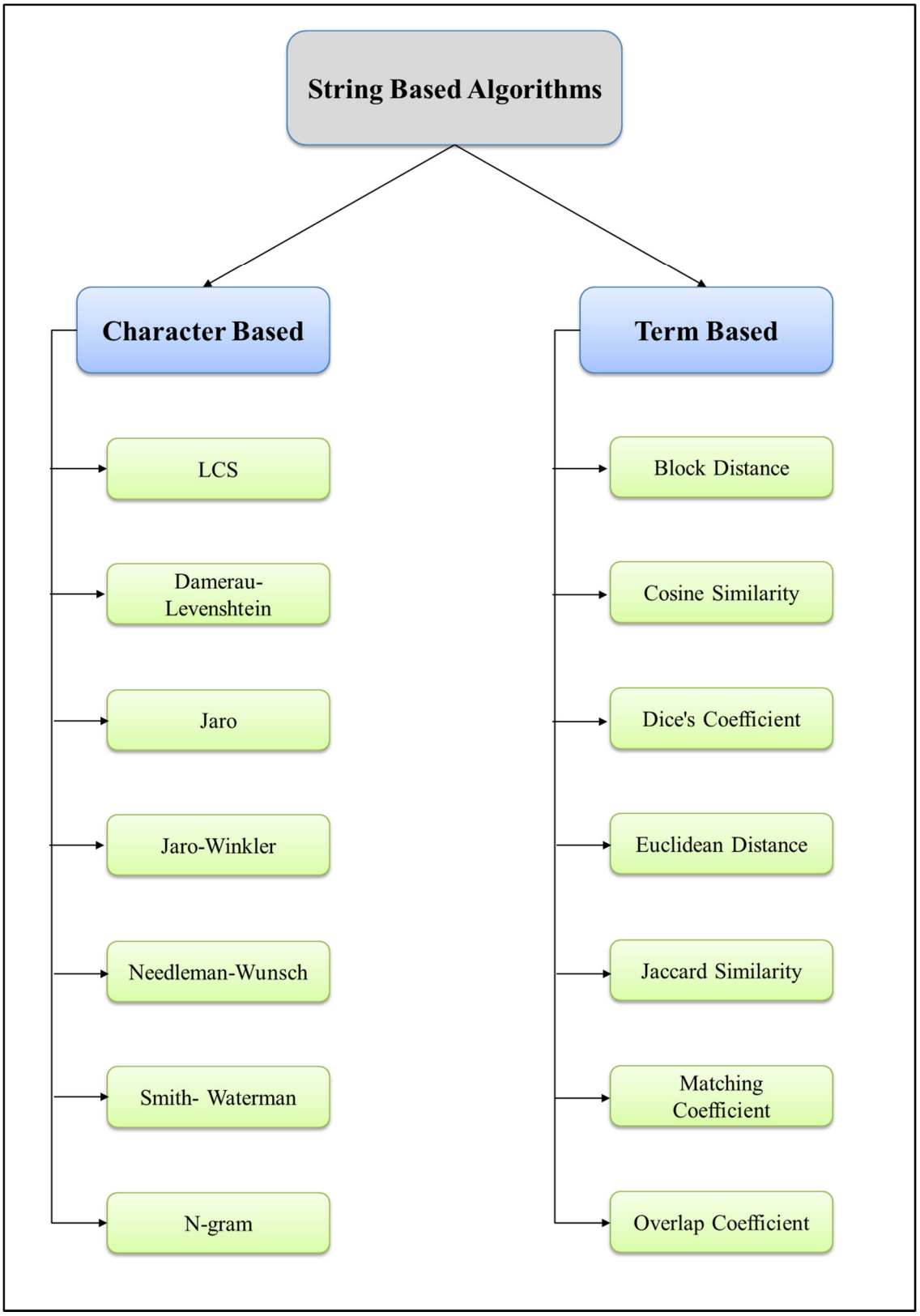

Figure 14. String-Based Similarity Measures (Gomaa and Fahmy 2013)

The ***Longest Common SubString*** (LCS) algorithm is used to find the longest string (or strings) that is a substring (or are substrings) of two or more strings. The similarity between two strings is based on the length of contiguous chain of characters that exist in both strings. The longest common substring of the strings "ABABC", "BABCA", and "ABCBA" is string "ABC" of length 3. Other common substrings are "A", "AB", "B", "BA", "BC", and "C". Table 26 shows an output of the LCS algorithm.

Table 26. Output of LCS algorithm

```
ABABC
  |||
 BABCA
  |||
  ABCBA
```

The problem definition for LCS can be described as follows.

Given two strings, *S* of length *m* and *T* of length *n*, find the longest strings that are substrings of both *S* and *T*. A generalization of this problem is the k-common substring problem. Given the set of strings $S = \{S_1, \ldots, S_K\}$ where $|S_i| = n_i$ and $\sum n_i = N$, find each $2 \leq k \leq K$, the longest string that occurs as substrings of at least *k* strings.

Damerau-Levenshtein distance counts the minimum number of operations needed to transform one string into the other to measure the distance between two strings. An operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters. The Damerau–Levenshtein distance differs from

the classical Levenshtein distance by including transpositions among its allowable operations. The classical Levenshtein distance only allows insertion, deletion, and substitution operations.

The Damerau–Levenshtein distance between two strings *a* and *b* is given by $d_{a,b}(|a|,|b|)$ where:

$$d_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j)\ j = 0, \\ min\begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \\ d_{a,b}(i-2,j-2)+1 \end{cases} & if\ i,j > 1\ and\ a_i = b_{j-1}\ and\ a_{i-1} = b_j \\ min\begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & otherwise \end{cases}$$

Jaro and ***Jaro–Winkler*** distance depend on the number and order of the common characters between two strings; it takes into account typical spelling deviations. Jaro is primarily used in the area of record linkage. Jaro–Winkler is an extension of Jaro distance, and it uses a prefix scale, which gives more favorable ratings to strings that match from the beginning for a set prefix length. The higher the Jaro–Winkler distance for two strings is, the more similar the strings are. The Jaro–Winkler distance metric is designed for short strings, such as person names. The score is normalized such that 0 equates to no similarity and 1 is an exact match.

Problem definition for Jaro distance and Jaro–Winkler distance can be described as follows.

The Jaro distance d_j of two given strings s_1 and s_2 is

$$d_j = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & otherwise, \end{cases}$$

where *m* is the number of matching characters and *t* is the number of transpositions.

Jaro–Winkler distance uses a prefix scale *p,* which gives a more generous score to strings that match from the beginning for a set prefix length *l*. Given two strings s_1 and s_2, their Jaro–Winkler distance d_w is:

$$d_w = d_j + (lp(1 - d_j))$$

Where d_j is the Jaro distance for strings s_1 and s_2. *l* is the length of common prefix at the start of the string up to a maximum of four characters. *p* is a constant scaling factor for how much the score is adjusted upwards because having common prefixes *p* should not exceed 0.25, otherwise the distance can become larger than 1. The standard value for this constant in Winkler's work is *p* = 0.1

The ***Needleman-Wunsch*** algorithm is an example of dynamic programming and is used in bioinformatics to align protein or nucleotide sequences. It performs a global alignment to find the best alignment over the entire of two sequences. The algorithm basically divides the full sequence into a series of smaller problems and uses the solutions for the smaller problems to reconstruct a solution to the larger problem. The Needleman–Wunsch algorithm is widely used for optimal global alignment, when the two sequences are of similar length and the global alignment is important.

The ***Smith-Waterman*** algorithm is another example of dynamic programming and performs local sequence alignment. It performs a local alignment to find the best alignment between two strings or nucleotide or protein sequences. It is useful for

dissimilar sequences that are suspected of containing regions of similarity or similar sequence motifs within their larger sequence context. The distinction of the Needleman–Wunsch algorithm is that negative scoring matrix cells are set to zero, which renders the local alignments visible.

Problem definition for Smith-Waterman algorithm can be described as follows.

$$H(i,0) = 0, 0 \leq i \leq m$$

$$H(0,j) = 0, 0 \leq j \leq n$$

$$H(i,j) = max \begin{cases} 0 \\ H(i-1,j-1) + s(a_i, b_j) \quad Match/Mismatch, \\ max_{k \geq 1}\{H(i-k,j) + W_k \quad Deletion \\ max_{l \geq 1}\{H(i,j-l) + W_l \quad Insertion \end{cases}, 1 \leq i \leq m, 1 \leq j \leq n$$

Where *a, b* = String over the Alphabet ∑, *m* = length(*a*), *n* = length(*b*), *s(a, b*) is a similarity function on the alphabet, H(*i, j*) is the maximum similarity score between a suffix of a[1…*i*] and a suffix of b[1…*j*], W_i is the gap scoring scheme.

An n-gram is a sub-sequence of n items from a given sequence of text. The ***n-gram similarity algorithm*** compares the n-gram characters or words in two strings. Text distance is calculated by dividing the number of same n-grams with maximal number of n-grams.

Block Distance is also known as rectilinear distance, boxcar distance, absolute value distance, L_1 distance, city block distance, and Manhattan distance. It computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The block distance between two items is the sum of the differences of their corresponding components. The block distance, d_1, between two vectors *p, q* in an *n*-

dimensional real vector space with fixed Cartesian coordinate system, is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes.

Problem definition for block distance can be described as follows.

$$d_1(p,q) = \|p-q\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

Where (p, q) are vectors

$$p = (p_1, p_2, \dots, p_n) \; and \; q = (q_1, q_2, \dots, q_n)$$

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. Thus, it determines an orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1; two vectors at 90° have a similarity of 0; and two vectors diametrically opposed have a similarity of -1, regardless of their magnitude. Cosine similarity is generally used in positive space, so the outcome is bounded within 0 and 1. One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate.

Cosine similarity can be derived by using the Euclidean dot product formula.

$$a \cdot b \; = \|a\|\|b\| \cos(\theta)$$

Given two vectors of attributes, A and B, the cosine similarity, *cos(θ)*, is represented using a dot product and magnitude as

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Dice's coefficient is defined as twice the number of common terms in the compared strings divided by the total number of terms in both strings. Dice's coefficient

retains sensitivity in more heterogeneous data sets and gives less weight to outliers. Recently it has become popular in computer lexicography for measuring the lexical association score of two given words.

Definition of Dice's coefficient can be described as follows.

$$S_v = \frac{2|A \cdot B|}{|A|^2 + |B|^2}$$

where (A, B) are binary vectors

$$A = (a_1, a_2, \dots, a_n) \; and \; B = (b_1, b_2, \dots, b_n)$$

Euclidean distance or L2 distance is the "ordinary" distance between two points in Euclidean space and can be described as the square root of the sum of squared differences between corresponding elements of the two vectors. It can be described as follows.

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

where p and q are Euclidean vectors.

Jaccard similarity, also known as the Jaccard index, is used for comparing the similarity and diversity of sample sets. It is computed as the number of shared terms over the number of all unique terms in both strings. It can be described as follows.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where $0 \leq \mathrm{J}(A, B) \leq n$ and if *A* and *B* are both empty, we define *J(A,B)* = 1.

Matching coefficient, also known as Simple Matching Coefficient (SMC), is a vector-based approach that simply counts the number of similar terms or dimensions, on which both vectors are non-zero. Given two objects, A and B, each with *n* binary attributes, SMC is defined as follows.

$$SMC = \frac{Number\ of\ Matching\ Attributes}{Number\ of\ Attributes} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

Overlap coefficient, also known as Szymkiewicz-Simpson coefficient, is similar to the Jaccard index but considers two strings a full match if one is a subset of the other. It is defined as the size of the intersection divided by the smaller of the size of the two sets. Overlap coefficient is defined as follows.

$$overlap(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

Corpus-Based Similarity Measures

Corpus-based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. A corpus, which is a large collection of written or spoken text data, is required to compute corpus-based similarity. Figure 15 shows the algorithms for corpus-based similarity measures.

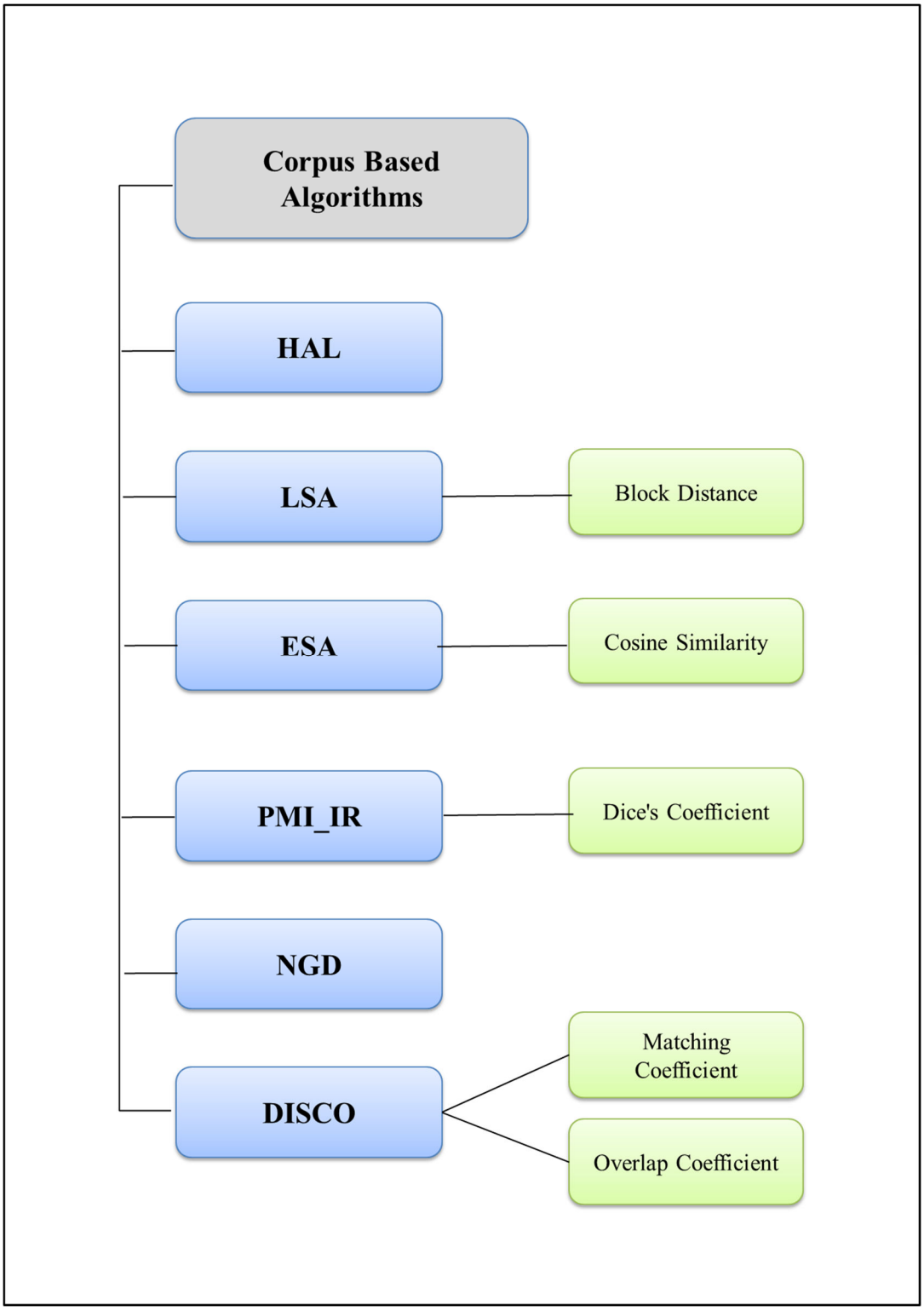

Figure 15. Corpus-Based Similarity Measures (Gomaa and Fahmy 2013)

Hyperspace Analogue to Language (HAL) creates a semantic space from word co-occurrences. The basic premise that HAL relies on is that words with similar meaning repeatedly occur closely (i.e., co-occurrence). For example, in a large corpus of text, one could expect to see the words *mountain*, *valley,* and *river* appear close to each other often. The same might be true for *mouse*, *cat,* and *dog*. HAL creates an *N* by *N* matrix, where *N* is the number of words in its lexicon and each matrix element is the strength of association between the word represented by the column and row. As the text is analyzed, a focus word is placed at the beginning of a 10-word reading frame that records which neighboring words are counted as co-occurring, and the 10-word reading moves incrementally through a corpus of text. Matrix values are accumulated by weighting the co-occurrence inversely proportional to the distance from the focus word; closer neighboring words are thought to reflect more of the focus word's semantics and so are weighted higher. The semantic similarity between two words is given by the cosine of the angle between their vectors.

Latent Semantic Analysis (LSA) is one of the most popular techniques of the corpus-based similarity measure algorithm. It assumes that words are semantically similar if they appear together in the same context. In LSA, a T × D matrix is constructed from a text corpus where T is the number of terms in the corpus and D is the number of documents. With a T × D matrix, a singular value decomposition (SVD) is used to reduce the number of columns while preserving the similarity structure among rows. Words are then compared by taking the cosine of the angle between the two vectors formed by any two rows.

Generalized Latent Semantic Analysis (GLSA) is a framework for computing semantically motivated term and document vectors. GLSA extends the applicability of the idea of the LSA approach, but GLSA focuses on term vectors instead of the dual document-term representation. GLSA requires a measure of semantic association between terms and a method of dimensionality reduction. The GLSA approach can combine any kind of similarity measure on the space of terms with any suitable method of dimensionality reduction. The traditional term document matrix is used in the last step to provide the weights in the linear combination of term vectors.

Explicit Semantic Analysis (ESA) is a vectorial representation of text that uses a document corpus as a knowledge-based measure. It computes the “semantic relatedness” between two arbitrary texts. The Wikipedia-based technique represents terms (or texts) as high-dimensional vectors; each vector entry presents the TF-IDF weight between the term and one Wikipedia article. The semantic relatedness between two terms (or texts) is expressed by the cosine measure between the corresponding vectors. The name "explicit semantic analysis" contrasts with latent semantic analysis (LSA) because the use of a knowledge base makes it possible to assign human-readable labels to the concepts that make up the vector space.

Cross-Language Explicit Semantic Analysis (CL-ESA) is a multilingual generalization of ESA. CL-ESA utilizes a document-aligned multilingual reference collection like Wikipedia to represent a document as a language-independent concept vector. The relatedness of two documents in different languages is assessed by the cosine similarity between the corresponding vector representations.

Pointwise Mutual Information - Information Retrieval (PMI-IR) is a measure of the similarity of pairs of words. It uses a web-based search engine to calculate probabilities. The more often two words co-occur near each other on a web page, the higher is their PMI-IR similarity score. PMI-IR uses Pointwise Mutual Information (PMI) as follows.

$$pmi(X,Y) = log\frac{p(x,y)}{\mathrm{p(x)p(y)}} = log\frac{p(x|y)}{p(x)} = log\frac{p(y|x)}{p(y)}$$

Second-order Co-occurrence Pointwise Mutual Information (SCO-PMI) is a semantic similarity measure using pointwise mutual information to sort lists of important neighbor words of the two target words from a large corpus. SOC-PMI can calculate the similarity between two words that do not co-occur frequently because they co-occur with the same neighboring words. The method considers the words that are common in both lists and aggregates their PMI values (from the opposite list) to calculate the relative semantic similarity.

Normalized Google Distance (NGD) is a semantic similarity measure based on the number of hits from the Google search engine for a given set of keywords. Keywords with the same or similar meanings tend to be "close" in units of Google distance, while words with dissimilar meanings tend to be farther apart.

The Normalized Google Distance between two search terms x and y is as follows:

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)| - \log f(x,y)}{\mathrm{logM} - \min\{\log f(x), \log f(y)\}}$$

where *M* is the total number of web pages searched by Google; *f(x)* and *f(y)* are the number of hits for search terms *x* and *y*; and *f(x, y)* is the number of web pages on which

both *x* and *y* occur. If the two search terms *x* and *y* never occur together on the same web page and they occur on separate web pages, the NGD is infinite. If they always occur on the same page, their NGD is zero.

Extracting Distributionally-related Words Using Co-occurrences (DISCO) is a Java-based application that allows the retrieval of the distributional similarity between arbitrary words and phrases. The Distributional Hypothesis in linguistics is derived from the semantic theory of language usage. The words that are used and occur in the same contexts tend to purport similar meanings. Large text collections are statistically analyzed to get the distributional similarity. When two words are subjected for exact similarity, DISCO simply retrieves their word vectors from the indexed data and computes the similarity according to Lin measure. If the most distributionally similar word is required, DISCO returns the second order word vector for the given word. DISCO has two main similarity measures: DISCO1 and DISCO2. DISCO1 computes the first order similarity between two input words based on their collocation sets. DISCO2 computes the second order similarity between two input words based on their sets of distributionally similar words.

Knowledge-Based Similarity Measures

Knowledge-based similarity is a semantic similarity measure that determines the degree of similarity between words using information derived from semantic networks. WordNet is the most popular semantic network in the area of measuring the knowledge-based similarity between words. WordNet is a huge lexical database of English words. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. Synsets are interlinked by means of conceptual-semantic and lexical relations. Figure 16 shows knowledge-based similarity measures, which can be categorized into two groups: measures of semantic similarity and measures of semantic relatedness.

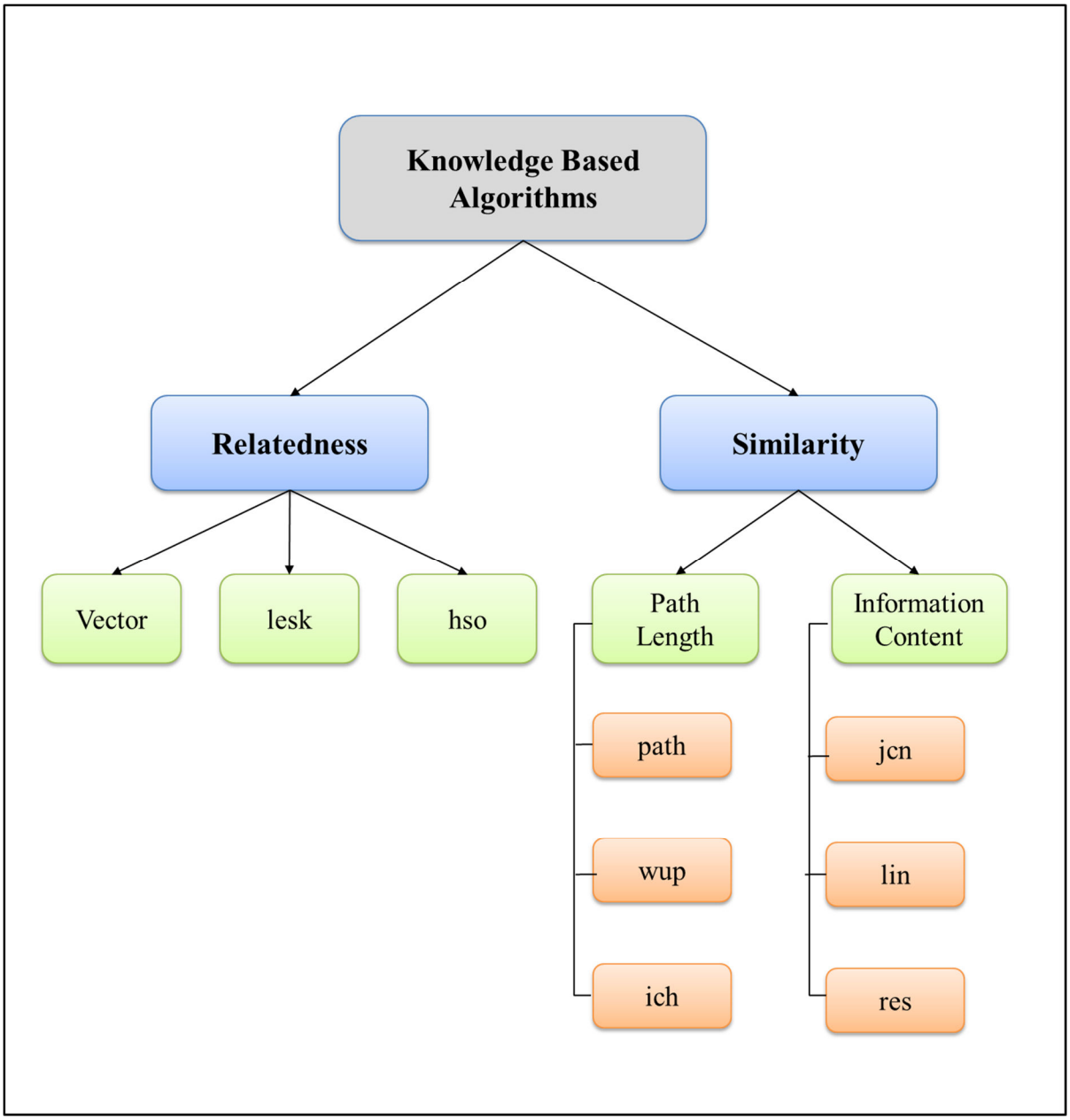

Figure 16. Knowledge-Based Similarity Measures (Gomaa and Fahmy 2013)

Measures of semantic similarity are often based on information regarding “is-a” relations found in a concept hierarchy. It takes two concepts as input and returns a numeric score that measures how much they are alike, based on is-a relationships. For example, common cold and illness are similar in that a common cold is a kind of illness.

However, there are other relations between concepts such as has-part, is-a-kind-of, is-a-specific-example-of, and is-the-opposite-of that existing measures of similarity cannot use since they only account for is-a relations (Pedersen, 2005). This suggests that more general measures of semantic relatedness are needed to take advantage of the increasingly rich ontologies (particularly in the medical domain) that have a wealth of relations beyond is-a (Pedersen, 2005).

There are six measures of semantic similarity; three of them are based on information content and the other three measures use path length.

The ***res*** is a Perl module for computing semantic relatedness of word senses that uses an information content-based measure described by Resnik (1995). The res measure uses the information content of concepts, computed from their frequency of occurrence in a large corpus, to determine the semantic relatedness of word senses.

The ***lin*** (Lin 1998) and ***jcn*** (Jiang and Conrath 1997) measure and augment the information content of the Least Common Subsumer (LCS) with the sum of the information content of concepts A and B themselves. The lin measure scales the information content of the LCS by this sum, while jcn takes the difference of this sum and the information content of the LCS.

Three similarity measures are based on path lengths between concepts: ***lch*** (Leacock & Chodorow 1998), ***wup*** (Wu & Palmer 1994), and ***path***. The lch measure finds the shortest path between two concepts and scales that value by the maximum path length in the "is-a" hierarchy in which they occur. The wup measure finds the path length to the root node from the LCS of the two concepts, which is the most specific concept they share as an ancestor. This value is scaled by the sum of the path lengths from the

individual concepts to the root. The measure path is equal to the inverse of the shortest path length between two concepts.

Furthermore, there are three measures of semantic relatedness. The ***hso*** by Hirst and St-Onge (1998), the ***lesk*** by Lesk (1986), and the ***vector***. The hso measure works by finding lexical chains linking two word senses. There are three classes of relations that are considered: extra-strong, strong, and medium-strong. The maximum relatedness score is 16. The lesk measure works by finding overlaps in the terms of the two synsets. The relatedness score is the sum of the squares of the overlap lengths. The vector measure creates a co–occurrence matrix for each word used in the WordNet glosses from a given corpus and then represents each gloss and concept.

In this section, similarity measurement algorithms were reviewed and categorized according to three approaches: string-based measures, corpus-based measures, and knowledge-based measures. Generally, patient records and clinical trial subject eligibility criteria are not written in grammatically perfect sentences. In most cases, they are written as fragmented sentences or bullet points. Therefore, using corpus-based similarity measures is not a good idea since this requires a corpus, which is a large collection of written or spoken text data. Clinical trial subject eligibility criteria include a large number of medical terms, so cosine similarity from term-based distance measures was selected for the matching process. Also, knowledge-based measures were combined with term-based distance measures by using UMLS semantic networks for semantic feature expansion. In this essay, I adopt a hybrid similarity measure, which combine term-based and knowledge-based distance measures.

4.4. Research Method

Figure 17 shows the steps for matching patient health records with clinical trial clusters and individual clinical trials. The first step of this research was to prepare clinical trial data from ClinicalTrials.gov and patient data from a prior research database. The second step was pre-processing using lemmatization, tokenization, and stop word removal. The next step was expanding the feature set with the custom dictionary and UMLS semantic network. A two-phase matching process was then conducted. Phase I matched patient information with the clusters that were generated in essay 2. Phase II matched patient information and clinical trials within the clusters. Also, patient information was matched with the entire clinical trial data set. Finally, internal and external evaluations were conducted.

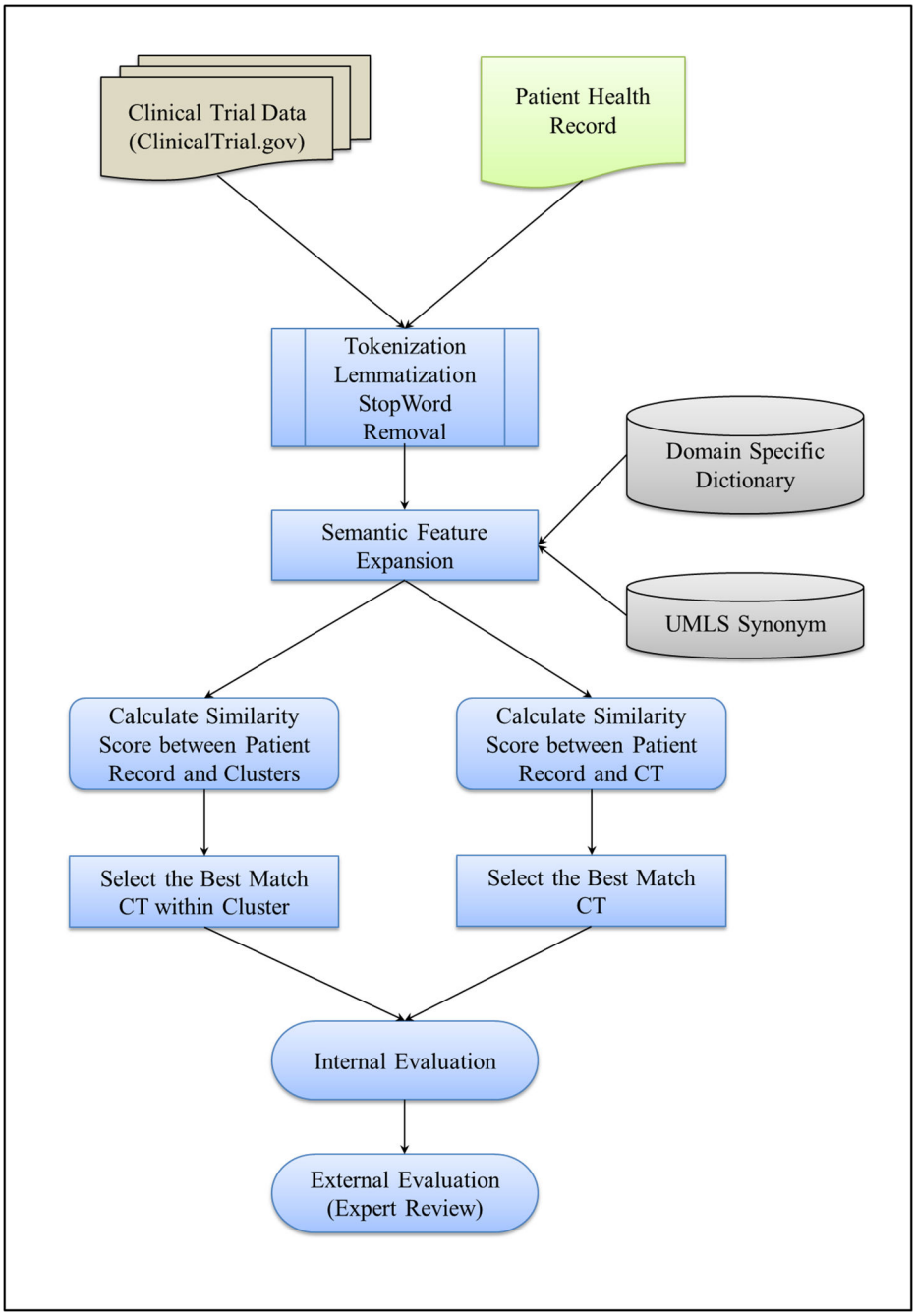

Figure 17. Steps in Automatic Matching of Patient Record and Clinical Trial Clusters / Individual Clinical Trials

4.4.1. Data Set

The patient data were acquired from a large community hospital in a major urban area in the Midwest, where, on average, 150 patients are diagnosed with breast cancer

each year. The original data set was collected for a prior research study (Gaudioso 2010) and has structured data such as demographic information as well as unstructured data such as documents (e.g., pathology, radiology, surgery reports). I collected only unstructured patient text data such as provider notes, biopsy reports, diagnostic workups, personal medical histories, physical exam reports, and surgery reports. All the patient data were de-identified, so the names of the patients were not included. I collected text data for a total of 148 patients, out of which data for 38 patients was excluded because there was not sufficient text data for those patients. Therefore, a data set of 110 patients was used for the matching process.

The database entities in the original study were normalized to secure data consistency. For the uniqueness of patient level records, the lower level data set was integrated into the higher level. The hierarchy of the patient data structure is presented in Figure 18. The lowest level of patient data is “encounter,” which is defined as “An interaction between a patient and healthcare provider(s) for the purpose of providing healthcare service(s) or assessing the health status of a patient” by ANSI-accredited standards developing organization, Health Level Seven International (HL7). The encounter level records were aggregated to Episode, which is defined as "An important event or series of events taking place in the course of continuous events” by Farlex Partner Medical Dictionary (2012). The episode level data was consolidated into case level. The definition of “case” in the medical field is “An instance of disease with its attendant circumstances,” according to the Farlex Partner Medical Dictionary (2012). Finally, all the case records were integrated into the patient level.

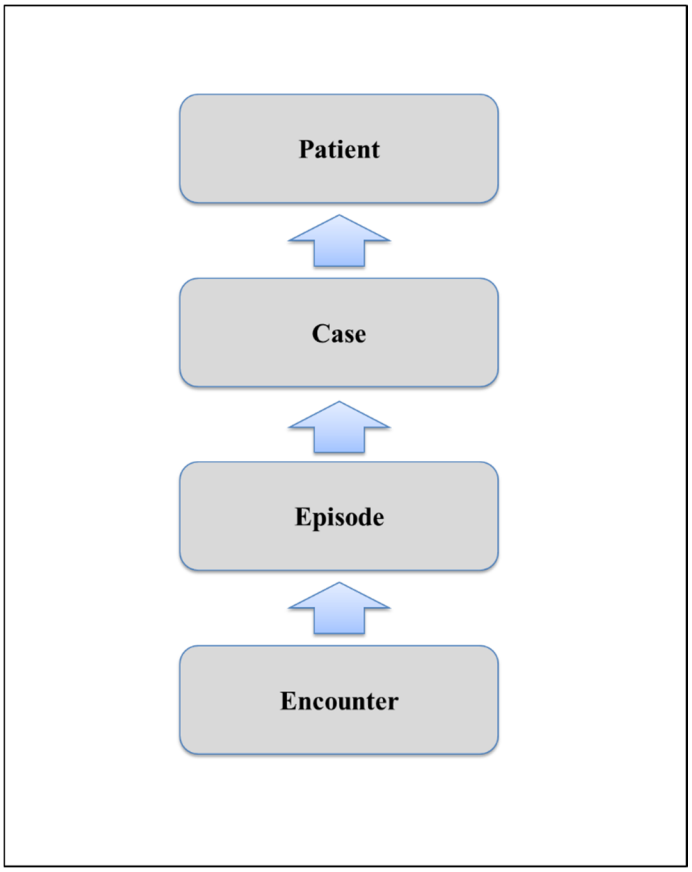

Figure 18. Hierarchy of Patient Data Structure

Figure 19 presents an SQL query statement to integrate a patient record, and Table 27 shows a sample of an integrated patient record.

```
SELECT patient.patient_id, casel.case_id, episode.episode_id, encounter.encounter_id,
       diagnostic_workup.DX_Wkup_result,
       diagnostic_workup.DX_Wkup_result_Desc,
       encounter.provider_note,
       Physical_exam.CBE_Report,
       Surgery.Resection_involved_Margin_Comm,
       Surgery.Orientation_Resection_Sn_Comm,
       Surgery.Reconstructive_BS_Complications_Comm,
       Surgery.Contraindication_BC_Comm,
       Biopsy.complications
FROM patient LEFT JOIN casel ON patient.patient_id = casel.patient_id
       LEFT JOIN episode ON casel.case_id = episode.case_id
       LEFT JOIN encounter ON episode.episode_id = encounter.episode_id
       LEFT JOIN biopsy ON encounter.encounter_id = biopsy.encounter_id
       LEFT JOIN diagnostic_workup ON encounter.encounter_id = diagnostic_workup.encounter_id
       LEFT JOIN Physical_exam ON encounter.encounter_id = Physical_exam.encounter_id
       LEFT JOIN surgery ON encounter.encounter_id = Surgery.encounter_id
```

Figure 19. SQL Query Statement to Integrate Unique Patient Record

Table 27. Sample of Integrated Patient Record

Patient ID	Original Patient Record
1000001	Right breast, partial/simple mastectomy: - Breast tissue with proliferative fibrocystic changes. - Residual areas of lobular cancerization - No definite residual DCIS - Previous biopsy site changes. - Margins free of involvement. -Prognostic factors performed on previous biopsy CS08-12007: ER +(97%), PR+(85%) PREOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. POSTOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. PROCEDURE: Right mastectomy with level I axillary node excision. 1. This is a 50-year-old woman, who has had right mastectomy for recently diagnosed ductal carcinoma in situ of the right breast that also has features of lobular cancerization. She is stage pT1N0M0 and I recommend hormonal therapy with tamoxifen 20 mg daily for 5

Clinical trial data were collected from ClinicalTrials.gov. The search term "breast cancer" was applied to limit clinical trials to only the breast cancer domain. A total of 1,660 breast cancer clinical trials from January 1, 2010, to January 1, 2012, were downloaded as a collection of XML format files. A custom parser was used for removing unnecessary tags, and the clinical trial subject eligibility section was divided based on two opposite criteria: Inclusion and Exclusion. The basic structured information for eligibility criteria, gender, and age range was also maintained by including that information in the data file naming rule.

The clinical trial cluster data came from the second essay. I generated 596 clusters that had more than single instance and labeled each cluster using the most frequent

synonym chunk of semantic features. The same cluster data and labels were used for the matching process in this essay.

4.4.2. Pre-processing

Tokenization and lemmatization were performed for the patient data set with Stanford CoreNLP. The second step in pre-processing was stop word removal. Fox's stop words list with the Apache Lucene framework was applied to sieve out all insignificant words in the data. Table 28 presents a sample of the pre-processed patient record.

Table 28. Sample of Pre-processed Patient Record

Patient ID	Pre-processed Patient Text Data
1000001	right breast partial simple mastectomy breast tissue proliferative fibrocystic change residual area lobular cancerization definite residual dci previous biopsy site change margin free involvement prognostic factor perform previous biopsy cs08-12007 er 97 pr 85 preoperative diagnosis ductal carcinoma situ right breast status post core biopsy intermediate grade cribriform type er pr positive postoperative diagnosis ductal carcinoma situ right breast status post core biopsy intermediate grade cribriform type er pr positive procedure Right mastectomy level axillary node excision 1 50-year old woman right mastectomy recently diagnose ductal carcinoma situ right breast feature lobular cancerization stage pt1n0m0 recommend hormonal therapy tamoxifen 20 mg daily 5

4.4.3. Matching with a Custom Dictionary

First, all trigram combinations from the pre-processed data set were identified; then each trigram term was matched with the custom dictionary. The three unigram tokens in the matched trigram were eliminated from the original data set. After all trigram

matching was completed, all bigram combinations from the modified data set were derived and matched with the custom dictionary. Table 29 shows trigram matching results between patient record 1000013 and the custom dictionary.

Table 29. Trigram Matching with the Custom Dictionary

Patient ID	Pre-processed patient record	Trigram Matching with The Custom Dictionary
1000013	successful ultrasound guided core biopsy highly suspicious palpable mass 12 30 position right breast ribbon shaped clip placement pathology grade iii ***invasive ductal carcinoma*** concordant 2 successful ultrasound guided vacuum assisted biopsy right breast 9 00 position s shaped clip placement pathology grade iii invasive ductal carcinoma concordant 3 Post biopsy digital right mammogram show accurate placement biopsy marking clip separate distance 6.4 cm.yes hematoma 10000132000014300000075000042 1 successful ultrasound guided core biopsy highly suspicious palpable mass 12 30 position right breast ribbon shaped clip placement pathology grade iii invasive ductal carcinoma concordant 2 successful ultrasound guided vacuum assisted biopsy right breast 9 00 position s shaped clip placement pathology grade iii invasive ductal carcinoma	invasive_ductal_carcinoma

Table 30 shows bigram matching results between patient record 1000013 and the custom dictionary.

Table 30. Bigram Matching with the Custom Dictionary

Patient ID	Pre-processed patient record	Bigram Matching with The Custom Dictionary
1000013	successful ultrasound guided core biopsy highly suspicious palpable mass 12 30 position right breast ribbon shaped clip placement pathology grade iii invasive ductal carcinoma concordant 2 successful ultrasound guided vacuum assisted biopsy right breast 9 00 position s shaped clip placement pathology grade iii invasive ductal carcinoma concordant 3 Post biopsy digital right mammogram show accurate placement biopsy marking clip separate distance 6.4 cm.yes hematoma 10000132000014300000075000042 1 successful ultrasound guided ***core biopsy*** highly suspicious palpable mass 12 30 position right breast ribbon shaped clip placement pathology grade iii invasive ductal carcinoma concordant 2 successful ultrasound guided vacuum assisted biopsy right breast 9 00 position s shaped clip placement pathology grade iii invasive ductal carcinoma	core_biopsy

4.4.4. Matching with the UMLS Semantic Network

A patient record is not a wordy document but is a succinct depiction of patient status. Moreover, the contents in a patient record are written by a healthcare provider and the target audience includes healthcare experts, so the patient record usually includes numerous medical terms. For that reason, I expanded the feature set in the patient record and the clinical trial eligibility section with synonymously related terms from the UMLS Semantic Network, based on semantic relatedness.

All bigram and trigram terms that matched with the custom dictionary were processed with the UMLS Semantic Network to find synonyms. Each bigram and trigram term was queried with the UMLS Semantic Network using a custom query statement.

Table 31 shows the UMLS synonym matching results for each trigram and bigram term.

Table 31. UMLS Synonym Matching Results for Trigram and Bigram

Patient ID	Trigrams and Bigram Found in Custom Dictionary	UMLS Synonym Matching
1000013	invasive_ductal_carcinoma	No Match
	core_biopsy	Biopsy-action BX-Biopsy Biopsy_sampling

4.4.5. Matching patient records with clinical trials within a cluster

There is considerable evidence that information technology could improve the subject recruitment process in clinical research. Dugas et al. (2009) showed that

complete, high-quality, and accurate data can significantly enhance the recruitment process. However, most relevant patient information still remains in an unstructured format (e.g., clinical notes, clinical assessments). The main objective in this essay is to find the best matching trials for a patient and to do this efficiently. Thus, the process starts with matching the patient record with clinical trial information

In this study I selected cosine similarity to compute the matching score between a patient record and a clinical trial cluster as well as between the patient record and each clinical trial, because cosine similarity is a measure of similarity between two vectors and is most commonly used in high-dimensional positive spaces. One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors, as only the non-zero dimensions need to be considered.

The cluster matching process was a two-step process. First, the matching between patient records and clinical trial clusters was conducted, and then each trial within the best matching cluster was also compared with the patient record. In the cluster matching, I included all clusters, including clusters with one trial. One of the main objectives in clustering is to reduce the search space for patient and trial matching. Therefore, to validate the efficiency of clustering, I compared the trial matching results within clusters with the matching results for the entire trial data set. Also, I set the threshold value for the best matched cluster as 0.95. All clusters that scored at or above 0.95 were included for the cluster matching.

Sample results for cluster matching are presented in Table 32. The highest matching score between patient and cluster is 1 and the lowest score is 0.4666. After all the best matching clusters for each patient were identified, I compared the patient records

with the clinical trials within those clusters. Phase I matches the patient records with the cluster information, while Phase II matches patient records with clinical trials within the matched clusters. In Phase II matching, two experiments were also conducted. In the first experiment in Phase II, the matching process was stopped when it found a trial whose similarity score was more or equal to 0.90. In the second experiment, I compared the patient record with all the trials in the best matching clusters. I also examined the matching results with the entire trial data set for the purpose of comparison. Figure 20 presents the matching experiments conducted in this study.

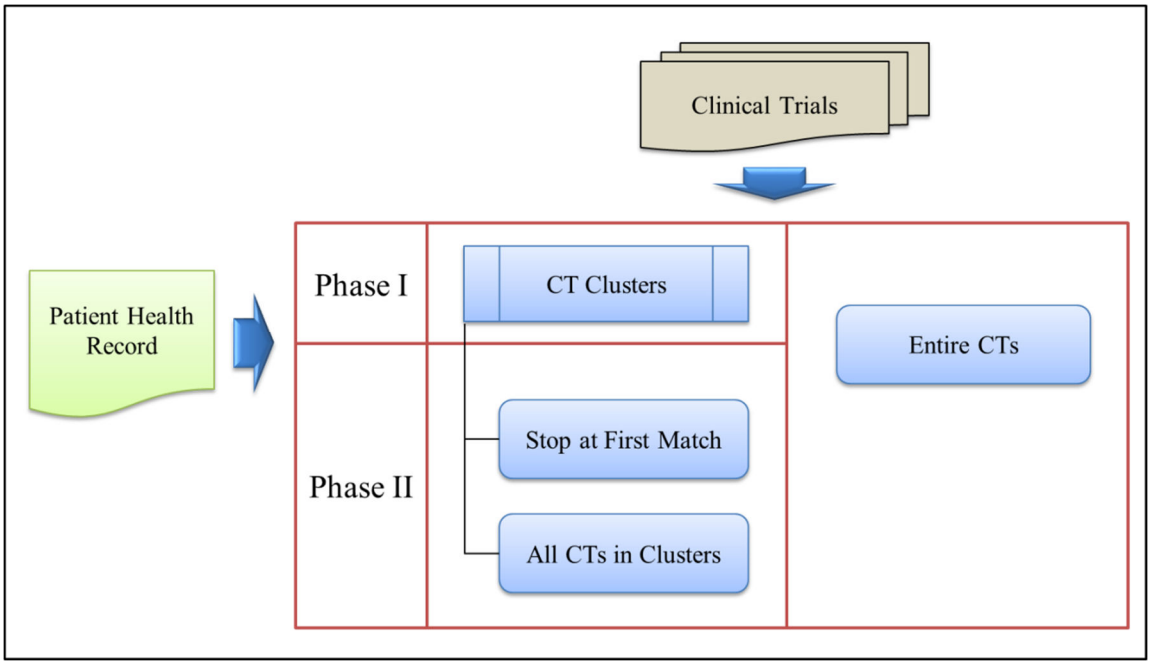

Figure 20. Matching Experiments in Research

In each trial, the subject eligibility criteria were divided into two groups: "Inclusion" and "Exclusion." Inclusion criteria are characteristics that the prospective subjects must have if they are to be included in the study, while exclusion criteria are

those characteristics that disqualify prospective subjects from inclusion in the study. Therefore, in this experiment, I excluded the trials whose exclusion criteria matched with the patient record. For example, the exclusion criteria included terms like “smoking” or “pregnant” because the study participant should not smoke or should not be pregnant. If any feature from the exclusion criteria matched with any feature of the patient record, that match was not included in the final results.

Table 33 shows the results of the trial within cluster matching for a patient. The highest matching score between the patient and trial within the cluster is 0.8101 and the lowest is 0.4730.

4.4.6. Matching patient record with entire clinical trial

Additional experiments were performed with each patient record and the entire set of clinical trials to find the best matches, regardless of clusters. Table 34 shows the sample results of matching between a patient and trials from the entire pool. The highest matching score between patient and trial information is 1 and the lowest matching is 0.1708.

Table 32. Sample of Patient and Cluster Matching Result

Patient ID	Patient_text	Cluster_ID	Cluster_label	Cluster matching score
1000001	Right breast, partial/simple mastectomy: - Breast tissue with proliferative fibrocystic changes. - Residual areas of lobular cancerization - No definite residual DCIS - Previous biopsy site changes. - Margins free of involvement. - Prognostic factors performed on previous biopsy CS08-12007: ER +(97%), PR+(85%) ,PREOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. POSTOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. PROCEDURE: Right mastectomy with level I axillary node excision. ,1. This is a 50-year-old woman, who has had right mastectomy for recently diagnosed ductal carcinoma in situ of the right breast that also has features of lobular cancerization. She is stage pT1N0M0 and I recommend hormonal therapy with tamoxifen 20 mg daily for 5	Inc(24)_Exc(127)	Carcinoma, no subtype|Epithelial tumor, malignant|Carcinoma|Malignant epithelial tumor|Malignant epithelial tumour|Epithelial tumour, malignant||Drug preparation|Drug product|Drug|Medicinal product|General drug type|Pharmaceutical / biologic product|Medicine|Medication|Drug, medicament or biological substance|Drug or medicament	1

Table 33. Sample of Patient and Trial within Cluster Matching Result

Patient ID	Patient_text	Best matching Trial within Cluster		
		CT_ID	CT_text	CT matching score
1000001	Right breast, partial/simple mastectomy: - Breast tissue with proliferative fibrocystic changes. - Residual areas of lobular cancerization - No definite residual DCIS - Previous biopsy site changes. - Margins free of involvement. -Prognostic factors performed on previous biopsy CS08-12007: ER +(97%), PR+(85%) ,PREOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. POSTOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. PROCEDURE: Right mastectomy with level I axillary node excision. ,1.	NCT01183663	Inclusion Criteria: 1. Patients with advanced or metastatic cancer that is refractory to standard therapy, has relapsed after standard therapy, or for which there is no standard therapy available. 2. Patients must be >/= 3 weeks beyond treatment with a cytotoxic chemotherapy regimen, therapeutic radiation, or major surgery. After targeted or biologic therapy there should be 5 half-lives or three weeks, whichever is shorter. Patients may have received palliative localized radiation immediately before or during treatment, providing radiation is not delivered only to the site of disease being treated under this protocol. 3. Eastern Cooperative Oncology Group (ECOG) performance status </= 2 4. Patients must have normal organ and marrow function, defined as absolute neutrophil count >/= 1,000/mL; platelets >/=50,000/mL (unless these abnormalities are due to bone marrow involvement); creatinine clearance >/= 50 ml/min by Cockcroft-Gault formula; total bilirubin </= 2.0; and alanine aminotransferase (ALT)/ serum glutamic pyruvic transaminase(SGPT) </= 5 X ULN (unless patient has liver metastases). 5. All study participants must be registered into the mandatory RevAssist® program, and	0.6410

	This is a 50-year-old woman, who has had right mastectomy for recently diagnosed ductal carcinoma in situ of the right breast that also has features of lobular cancerization. She is stage pT1N0M0 and I recommend hormonal therapy with tamoxifen 20 mg daily for 5		be willing and able to comply with the requirements of RevAssist®. 6. Females of childbearing potential (FCBP) must have a negative serum or urine pregnancy test with a sensitivity of at least 50 mIU/mL within 10 - 14 days prior to and again within 24 hours of prescribing lenalidomide (prescriptions must be filled within 7 days) and must either commit to continued abstinence from intercourse or begin TWO acceptable methods of birth control, one highly effective method and one additional effective method AT THE SAME TIME, at least 28 days before she starts taking lenalidomide. FCBP must also agree to ongoing pregnancy testing. Men must agree to use a latex condom during sexual contact with a FCBP even if they have had a successful vasectomy. 7. Patients must be able to understand and be willing to sign a written informed consent document. 8. Must be >/= 18 years of age. Exclusion Criteria: 1. Any serious medical condition, laboratory abnormality, or psychiatric illness that would prevent the subject from signing the informed consent form. 2. Uncontrolled intercurrent illness, including, but not limited to, uncontrolled infection, uncontrolled asthma, need for hemodialysis, need for ventilatory support. 3. Pregnant or breast feeding females. (Lactating females must agree not to breast feed while taking lenalidomide). 4. Use of any other experimental drug or therapy within 21 days of baseline. 5. Known hypersensitivity to thalidomide. 6. History of hypersensitivity to any component of the formulation. 7. The development of erythema nodosum, if characterized by a desquamating rash while taking thalidomide or	

			similar drugs. 8. Patients unwilling or unable to sign informed consent document. 9. Uncontrolled systemic vascular hypertension (Systolic blood pressure >140 mmHg, diastolic blood pressure > 90 mmHg on medication) for patients treated in the bevacizumab or sorafenib arms. 10. Patients with active deep venous thrombosis or pulmonary embolism or patients receiving anti-coagulation. 11. Patients with clinically significant cardiovascular disease: History of cerebro-vascular accident (CVA) within 6 months; Myocardial infarction or unstable angina within 6 months; Unstable angina pectoris. 12. Uncontrolled intercurrent illness, including, but not limited to, ongoing or active infection requiring parenteral antibiotics on Day 1. 13. Major surgical procedure, open biopsy or significant traumatic injury within 28 days prior to Day 0 of protocol treatment. 14. Patients that are taking CYP3A4 inducers and/or inhibitors, being considered for the temsirolimus arm: If a patient has a history of taking CYP3A4 inducers and/or inhibitors prior to enrollment on the temsirolimus arm, it is strongly recommended that the patient stops the drug and waits at least 5 half-lives of said drug before initiating therapy on the temsirolimus arm.	

Table 34. Sample of Patient and Trial among Entire Trial set Matching Result

Patient ID	Patient_text	Best matching Trial among Entire Trial set		
		CT_ID	CT_text	CT matching score
1000001	Right breast, partial/simple mastectomy: - Breast tissue with proliferative fibrocystic changes. - Residual areas of lobular cancerization - No definite residual DCIS - Previous biopsy site changes. - Margins free of involvement. -Prognostic factors performed on previous biopsy CS08-12007: ER +(97%), PR+(85%) ,PREOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. POSTOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. PROCEDURE: Right mastectomy with level I axillary node excision. ,1. This is a 50-year-old woman, who has had right mastectomy for recently	NCT01757730	Inclusion Criteria: Any participant 18 years or older and are MR safe. Exclusion Criteria: That study participants will be excluded if they have any unapproved metal in their bodies, and that the volunteers are pregnant or possible of becoming pregnant. Also if the participants are claustrophobic.	0.9813

	diagnosed ductal carcinoma in situ of the right breast that also has features of lobular cancerization. She is stage pT1N0M0 and I recommend hormonal therapy with tamoxifen 20 mg daily for 5			

4.5. Results

When I included single-instance and two-instance clusters in the experiment, the best match score between patient and cluster was always 1. Also, all the patients matched with multiple best clusters with score 1, and the number of best matched clusters ranged from 2 to 128. Table 35 shows the score results for patient and cluster matching. All the patients had at least one best match with a cluster, and all of the best matches had a score of 1. Thus, the upper bound and lower bound scores were both 1.

Table 35. Matching Results for Patient and Clinical Trial Clusters

Highest Best Matching Score	1
Lowest Best Matching Score	1
Number of Multiple Matches	110
Range of Multiple Matches	2 to 128

The match results between patient and individual clinical trials within the clusters were obtained through two different experiments. In the first experiment, the matching process was stopped when it found a trial whose match score was more than 0.90; the results from that experiment are presented in Table 36.

Table 36. Matching Results for Patient and Trial within Best Matched Cluster (Stop at First Match)

Highest Best Matching Score	0.9862
Lowest Best Matching Score	0. 9289
Average Best Matching Score	0.9632

In the second experiment, I compared the patient record with all the trials in the best clusters. The results from the second approach are presented in Table 37. There are several trial studies that scored 1 because the description of eligibility criteria for those trials was extremely short. The cosine similarity measure only considers orientation, not magnitude, so very short documents could have raised the level of noise in the experiment. To address this shortcoming, all the matches that scored 1 were removed from the results.

Table 37. Matching Results for Patient and Trial within Best Matched Cluster (All Trials)

Highest Best Matching Score	0.9931
Lowest Best Matching Score	0.9493
Average Best Matching Score	0.9845

The matching results between patient and the entire trial set are also presented in Table 38. This matching took around seven times longer than the one involving patient and all trials in the best clusters. Computationally, it incurred higher costs but produced similar results as those from patient and trial within the best matched cluster. Table 38 shows the match results between patient and single clinical trial in the entire trial set.

Table 38 Matching Results between Patient and Entire Trial

Highest Best Matching Score	0.9931
Lowest Best Matching Score	0.9493
Average Best Matching Score	0.9845
Number of Multiple Matches	22
Range of Multiple Matches	2

The efficiency of the matching process was evaluated by measuring the matching algorithm computing time. The main objective of clustering trials in the second essay was to reduce the search space and lower the computational costs for finding the best trial for a patient. To evaluate the efficiency of the clustering approach, I investigated the computing time for the three matching experiments. The system specification for this research is presented in Table 39.

Table 39. Research System Specification

OS Name	Microsoft Windows Server 2008 R2 Standard, 64 bit
OS Version	6.1.7601 Service Pack 1 Build 7601
Processor	Intel(R) Xeon(R) CPU E5440 @ 2.83GHz, 2826 Mhz, 1 Core(s), 1 Logical Processor(s)
BIOS Version/Date	American Megatrends Inc. 080002, 5/5/2008 SMBIOS Version 2.3
Total Physical Memory	12.0 GB
Available Physical Memory	9.12 GB
Total Virtual Memory	24.0 GB
Available Virtual Memory	18.0 GB
Page File Space	12.0 GB
Disk Size	270.99 GB (290,977,505,280 bytes)
Program Language	java version "1.6.0_45"
Integrated Development Environment	Eclipse IDE for Java Developers Version: Juno Service Release 2 Build id: 20130225-0426
Database	MySQL 5.5.30

Table 40 shows the computing times for the matching process of patient and the first matched trial within the best matched cluster.

Table 40 Computing Time for the Matching Process of Patient and Trial within Best Matched Clusters (Stop at First Match)

Longest Computing Time	0.3397 sec
Shortest Computing Time	0.0528 sec
Average Computing Time	0.0867 sec

Table 41 shows the computing time for the matching process of patient and all trials within the best matched clusters.

Table 41 Computing Time for the Matching Process of Patient and Trial within Best Matched Clusters (All trial)

Longest Running Time	0.8298 sec
Shortest Running Time	0.1246 sec
Average Running Time	0.3356 sec

Table 42 shows the computing time for the matching process of patient and the entire trial set.

Table 42 Computing Time for the Matching Process of Patient and Entire Trial Set

Longest Running Time	3.531461 sec
Shortest Running Time	2.070947 sec
Average Running Time	2.2516 sec

In order to analyze the differences among three group means and variation among and between groups, I conducted the analysis of variance (ANOVA) test. Table 43 shows a summary of the three experimental groups, and Table 44 presents the results of the ANOVA test.

Table 43. Summary of Three Experiment Groups for Patient and Trial Matching

Groups	Counts	Sum	Average	Variance
Patient and Trial within Best Matched Clusters (Stop at First Match)	110	9.6327	0.0867	0.0036
Patient and Trial within Best Matched Clusters (All trials)	110	37.5165	0.3379	0.0238
Patient and Entire Trial Set	110	250.3816	2.2556	0.04488

Table 44 shows that the p-value is less than 0.05, indicating that the mean values of computing times for the three experimental groups were significantly different.

Table 44. Results of ANOVA Test for Three Experiment Groups

Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	312.4595	2	156.2297	6477.9563	0.00	3.0230
Within Groups	7.95865	330	0.02411			
Total	320.4182	332				

The mean differences among the three groups were statistically different, and it can be interpreted that the clustering approach in the patient and clinical trial matching can significantly expedite the clinical trial subject recruitment process.

A two-tail pairwise t-test was conducted to find differences among the groups. Table 45 shows results of pairwise t-test. All the p-values were less than 0.5, and statistically significant differences existed among the three groups.

Table 45. Results of Pairwise t-test (Two tail)

Groups	Pairwise t-test (two tail)
Patient and Trial within Best Matched Clusters (Stop at First Match)	0.0001
Patient and Trial within Best Matched Clusters (All trials)	0.0001
Patient and Entire Trial Set	0.0001

The quality of the matching was evaluated in this study using psycholinguistic evaluation. This evaluation approach is usually used for assessing the quality of semantic similarity measures. The psycholinguistic approach compares the computational approaches with human judgements. The correlation between the computational approach and human assessment is used as an evaluation measure to judge the quality of the similarity measure. The matching results were evaluated internally by researchers involved in the study and were then reviewed by an external medical expert who was a medical doctor as well as a PhD in management science. The internal researchers preliminarily tested the quality of the matching results and then the external medical expert assessed the quality of a sample of five final matching results. Table 46 shows the results of expert evaluation. The expert review reported one 'Very Good' and four 'Average' ratings. The expert provided comments as part of the evaluation. The comments on average ratings explained why he didn't mark those matches 'Very Good'

or ‘Good’. All the ‘Average’ rated matches were because of significant missing information in patient data that is required for a good match. For example, one of the expert’s comments was that “*Match on Confirmed diagnosis and ER, PR, Her status, and ax LNs. However, we are missing data on menopausal status and performance status…*”

The results from three experiments were discussed and showed the proposed two step matching results provided statistically improved performance. The matching results could be used for patient recruitment, estimation of clinical trial feasibility, and helping terminal disease patients.

Table 46. Results of Expert Evaluation for 5 Sample Matches

Value in Likert Scale	Count
Very Good	1
Good	-
Average	4
Poor	-
Very Poor	-
Total	5

4.6. Discussion

To the best of our knowledge, no attempt has so far been made to build an entire automatic matching process for clinical trial and patient information using state-of-the-art NLP and text mining algorithms. Also, this research is the first study that adopts the semantic-based feature expansion technique, which can improve clinical trial text analysis performance. Based on prior studies, the n-gram feature induction approach yielded more accurate outcomes for machine learning-based text analysis. I tried to capture the n-gram medical terms using the domain-specific custom dictionary, which, in clinical trial research, is the first attempt at applying the n-gram feature induction approach. Previous research on clinical trials failed to grasp the characteristics of the two opposite criteria in the eligibility section. In this research, we divided the subject eligibility section into "Inclusion Criteria" and "Exclusion Criteria" section to reflect the impact of each set of criteria precisely. Finally, we matched patient data with clinical trial clusters, under which similar alternative clinical trials were grouped. The results of the matching reduced healthcare practitioners' search space for clinical trials and significantly enhanced their patients' participation opportunity in trials.

I have presented a feasibility study for an NLP and text mining-based approach to matching patient records with clinical trials. Using a real-world patient data set, we described various framework and algorithms to address issues in the automatic patient recruitment process.

This study contributes to both research and practice. The study contributes to research by proposing a framework and providing algorithms based on semantic feature

expansion. Moreover, the algorithms and framework from this research could be used for different types of diseases and patient groups.

This study is not without limitations. In its current scope, it has limited generalizability. I only focused on the breast cancer domain with a limited set of patient records. Furthermore, this study adopted the cosine similarity measure in the matching process. However, there are several similarity measures that have been used in other research domains, such as information retrieval and computer science. Emerging similarity measure algorithms could be evaluated in the future. Moreover, by the nature of cosine similarity, semantics of documents were not considered in this research. Negation expression in clinical trial and patient text could not be captured.

There are several ways in which future research could strengthen the results of this study. As a further extension of our work, future researchers could conduct a field study involving a real hospital environment. Future studies could investigate the proposed model in the context of different types of disease. Semantic analysis could also be included in future research.

About 85% of people with cancer were either unaware or unsure that participation in clinical trials was an option, although about 75% of them said they would have been willing to enroll had they known it was possible. However, the clinical trial subject matching process is labor intensive and error prone. Our research would streamline the entire matching process and provide effective support to terminal disease patients.

CHAPTER 5

Conclusion and Future Directions

"It is a very sad thing that nowadays there is so little useless information."

Oscar Wilde

5.1. Introduction

It has been extensively recognized that recruitment of an adequate number of participants is essential for success of a clinical trial. Several studies have found that low participation in clinical trials is a significant issue resulting in inadequate statistical analyses of outcomes, premature closure of trials, longer trial duration, and higher costs of medical treatment.

In the field of oncology, fewer than 3% of potentially eligible patients enroll in clinical trials, and patient enrollment for clinical trials is as low as 2% of patient recruitment goals. Furthermore, more than 75% of participants are not even aware that trials exist, even though surveys have shown that a majority of people would be open to participating in these studies if they knew about them.

Extensive literature has been written about barriers to clinical trial participation, and one of the salient barriers for potential participation is participation of physicians. The participation of physicians is necessary to the success of clinical trial subject recruitment because they serve a critical role in helping their patients access trials. However, they do not have enough time to identify eligible study subjects efficiently, or they simply forget to offer and enroll patients in possible trials.

Therefore, it is necessary to develop new technologies and automatic tools that can process large text data into useful information and knowledge intelligently. NLP and text mining is a technique that can combine traditional data analysis methods with complex algorithms to deal with large amounts of text data. Additionally, text mining is a complex process that can extract the unknown and valuable modes or rules from mass data.

This three-essay dissertation attempts to contribute a solution to clinical trial subject recruitment problem. This study aims to provide an automatic matching framework for patient text information and clinical trial subject eligibility description. To achieve the main objective, I created a domain-specific custom dictionary as a lexical resource in essay 1, generated clinical trial clusters for the breast cancer domain in essay 2, and proposed a two-step automatic matching process in essay3.

One of the most time-consuming and high labor cost tasks in text mining research is the creation, compilation, and customization of the necessary lexicons. The first essay attempted to build a domain-specific lexicon focusing on breast cancer and showed the semi-automated dictionary building process. The evaluations for the breast cancer domain-specific dictionary shows that even though the coverage of a domain-specific dictionary is slightly less than the UMLS Metahthesaurus, the efficiency is more than 30-fold higher than UMLS resources

This second essay grouped and summarized clinical trial subject eligibility using the clustering approach. This essay also showed the framework for clustering clinical trial and labeling process. The findings from the second essay suggest that the clustering

approach could help practicing physicians reduce the search space of potential clinical trials.

The last essay proposed an entire automatic matching process for clinical trial and patient information using state-of-the-art NLP and text mining algorithms. This study contributes to both research and practice. The study contributes to research by providing algorithms and a framework based on semantic feature expansion. Moreover, the findings in this research, such as algorithms and the framework on which they are based, could be used for different types of diseases and patient groups.

5.2. Limitations

No claim is made as to the completeness of this research study. For the first essay, the coverage rate of the custom dictionary is relatively low because the data set included not only noun but also verb, adverb, and adjective words. Second, the custom dictionary included limited online sources. Thus, if a more comprehensive resource is included in future research, it will result in better performance.

The second essay focused only on the context of the breast cancer domain, which may represent lack of generalizability. While agglomerative hierarchical clustering with cosine distance was adopted to cluster clinical trials, other clustering algorithms and distance measures need to be compared.

The scope of the last essay has limited generalizability. I focused only on the breast cancer domain with limited patient records. The last essay also adopts cosine similarity to measure in the matching process. However, there are several other similarity measures that can be used for other research domains, such as information retrieval and

computer science. Emerging similarity measure algorithms should be evaluated. Semantic analysis of documents were not considered in this research. Negation expressions in data were not captured.

5.3. Future Directions

As described in the conclusion section for each individual essay, there is always room for enhancement and extension of the algorithms used in these essays. Future research for the first essay could be evaluation with only a noun word data set, which could increase the coverage rate of custom dictionary. The custom dictionary included limited online sources. Therefore, if a more comprehensive resource is included in future research, it will result in better performance.

There are several ways in which future research could strengthen the results of the second essay. First, future studies could investigate the proposed clustering framework in the context of different kinds of diseases to extend generalizability. Second, different approaches for clustering and document similarity metrics could be used. For example, Latent Dirichlet allocation (LDA), latent semantic indexing, independent component analysis, probabilistic latent semantic indexing, non-negative matrix factorization, and Gamma-Poisson distribution techniques are used in bioinformatics research. These new techniques could be applied in future research.

As a further extension of the third essay, researchers could conduct a field study involving a real hospital environment. Also, future studies could investigate the proposed model in the context of different types of diseases. To capture negation expressions,

semantic analysis could be included in future direction. Our research would streamline the entire matching process and provide effective support to terminal disease patients.

REFERENCE

Albrecht, T. L., Eggly, S. S., Gleason, M. E. J., Harper, F. W. K., Foster, T. S., Peterson, A. M., . . . Ruckdeschel, J. C. (2008). Influence of Clinical Communication on Patients' Decision Making on Participation in Clinical Trials. Journal of Clinical Oncology, 26(16), 2666-2673. doi: 10.1200/jco.2007.14.8114

Amine, A., Elberrichi, Z., Simonet, M., & Malki, M. (2008). *WordNet-Based and N-Grams-Based Document Clustering: A Comparative Study*. Paper presented at the Third International Conference on Broadband Communications, Information Technology & Biomedical Applications.

Apache Lucene from https://lucene.apache.org/

Apostolova, E., Lytinen, S., & Raicu, D. (2008). *Named Entity Recognition in Acute Inflammatory Response Studies*. Paper presented at the Swarmfest 2008.

Ashery, R. S., & McAuliffe, W. E. (1992). Implementation issues and techniques in randomized trials of outpatients psychosocial treatments for drug abusers recruitment of subjects. *American Journal of Drug & Alcohol Abuse, 18*(3), 305-329.

Bloehdorn, S., & Hotho, A. (2006). *Boosting for text classification with semantic features*. Paper presented at the Proceedings of the 6th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis, Seattle, WA.

Boland, M. R., Miotto, R., Gao, J., & Weng, C. (2013). Feasibility of feature-based indexing, clustering, and search of clinical trials. A case study of breast cancer trials from ClinicalTrials.gov. *Methods of Information in Medicine, 52*(5), 382-394. doi: 10.3414/me12-01-0092

Bollier, D., Communications, & Program, S. (2010). *The Promise and Peril of Big Data*: Aspen Institute, Communications and Society Program.

Botsis, T., Buttolph, T., Nguyen, M. D., Winiecki, S., Woo, E. J., & Ball, R. (2012). Vaccine adverse event text mining system for extracting features from vaccine safety reports. *Journal of the American Medical Informatics Association, 19*(6), 1011-1018. doi: 10.1136/amiajnl-2012-000881

Breitfeld, P. P., Weisburd, M., Overhage, J. M., Sledge, G., & Tierney, W. M. (1999). *Pilot Study of a Point-of-use Decision Support Tool for Cancer Clinical Trials Eligibility* (Vol. 6).

Campbell, D. A., & Johnson, S. B. (1999). *A Technique for Semantic Classification of Unknown Words Using UMLS Resources*. Paper presented at the AMIA Annual Symposium Proceedings 1999.

Cavnar, W. B., & Trenkle, J. M. (1994). *N-Gram-Based Text Categorization*. Paper presented at the 3rd Annual Symposium on Document Analysis and Information Retrieval.

Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D., & Siani-Rose, M. A. (2004). A Knowledge-Based Clustering Algorithm Driven by Gene Ontology. *Journal of Biopharmaceutical Statistics, 14*(3), 687-700. doi: 10.1081/bip-200025659

Chung, E.-K. (2009). A Semantic-Based Feature Expansion Approach for Improving the Effectiveness of Text Categorization by Using WordNet. *Journal of the Korean Society for information Management, 26*(3), 261-278.

Cohen, T., & Widdows, D. (2009). Methodological Review: Empirical distributional semantics: Methods and biomedical applications. *J. of Biomedical Informatics, 42*(2), 390-405. doi: 10.1016/j.jbi.2009.02.002

Corbett, P., Batchelor, C., & Teufel, S. (2007). Annotation of Chemical Named Entities. *Biological, translational, and clinical language processing*, 57 - 64.

Corley, C., & Mihalcea, R. (2005). *Measuring the semantic similarity of texts*. Paper presented at the Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Ann Arbor, Michigan.

Damashek, M. (1995). Gauging Similarity with n-grams: Language-Independent Categorization of Text. *Science 267*(5199), 843-849.

de Buenaga Rodriguez, M., Gómez-Hidalgo, J. M., & Diaz-Agudo, B. (1997). *Using WordNet to Complement Training Information in Text Categorization.* Paper presented at the Recent Advances in Natural Language Processing {II}: Selected Papers from the Second International Conference on Recent Advances in Natural Language Processing (RANLP 1997), March 25-27, 1997, Stanford, CA, USA.

Dunham, M. H. (2003). *Data Mining Introduction and Advanced Topics*. New Jersey: Pearson Education Inc.

Embi, P. J., Jain, A., Clark, J., Bizjack, S., Hornung, R., & Harris, C. (2005). EFfect of a clinical trial alert system on physician participation in trial recruitment. *Archives of Internal Medicine, 165*(19), 2272-2277. doi: 10.1001/archinte.165.19.2272

Fink, E., Kokku, P. K., Nikiforou, S., Hall, L. O., Goldgof, D. B., & Krischer, J. P. Selection of patients for clinical trials: an interactive web-based system. *Artificial Intelligence in Medicine, 31*(3), 241-254. doi: 10.1016/j.artmed.2004.01.017

Fisher, S., & Roark, B. (2007). *Feature expansion for query-focused supervised sentence ranking.* Paper presented at the Document Understanding Conference 2007.

Fox, C. (1989). A stop list for general text. *SIGIR Forum, 24*(1-2), 19-21. doi: 10.1145/378881.378888

Frank, G. (2004). Current challenges in clinical trial patient recruitment and enrollment. *SoCRA Source, 2*, 30-38.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*: Kluwer Academic Publishers.

Gomaa, W. H., & Fahmy, A. A. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications, 68*(13), 13-18. doi: 10.5120/11638-7118

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*: Kluwer Academic Publishers.

Hao, T., Rusanov, A., Boland, M. R., & Weng, C. (2014). Clustering clinical trials with similar eligibility criteria features. *Journal of Biomedical Informatics*(0). doi: http://dx.doi.org/10.1016/j.jbi.2014.01.009

Harris Interactive. (2001a). Misconceptions And Lack of Awareness Greatly Reduce Recruitment For Cancer Clinical Trials. *Health Care News, 1*(3).

Harris Interactive. (2001b). A Survey on Clinical Trial Barriers, *Health Care News*. Retrieved from http://harrisinteractive.com/about/healthnews/HI_HealthCareNews2001Vol_iss3.pdf

Hillner, B. E. (2004). Barriers to Clinical Trial Enrollment: Are State Mandates the Solution? *Journal of the National Cancer Institute, 96*(14), 1048-1049. doi: 10.1093/jnci/djh225

Institute of Medicine. (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*: National Academy Press.

Jensen, L., & Martinez, T. (2000). *Improving Text Classification by Using Conceptual and Contextual Features.* Paper presented at the Workshop on Text Mining at the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00}.

Jensen, L. S., & Martinez, T. (2000). Improving Text Classification by Using Conceptual and Contextual Features.

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

Johnson, S. B. (1999). *A Semantic Lexicon for Medical Language Processing*. Paper presented at the Journal of the American Medical Informatics Association.

Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science, 38*(6), 420-442.

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., & Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association, 18*(5), 601-606. doi: 10.1136/amiajnl-2011-000163

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

Johnson, S. B. (1999). *A Semantic Lexicon for Medical Language Processing*. Paper presented at the Journal of the American Medical Informatics Association.

Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science, 38*(6), 420-442.

Jonnalagadda, S., Cohen, T., Wu, S., & Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *J. of Biomedical Informatics, 45*(1), 129-140. doi: 10.1016/j.jbi.2011.10.007

Jonnalagadda, S., Cohen, T., Wu, S., Liu, H., & Gonzalez, G. (2013a). Evaluating the Use of Empirically Constructed Lexical Resources for Named Entity Recognition (pp. 23-33): Association for Computational Linguistics.

Jonnalagadda, S., Cohen, T., Wu, S., Liu, H., & Gonzalez, G. (2013b). Using Empirically Constructed Lexical Resources for Named Entity Recognition. *Biomedical Informatics Insights*(3738-BII-Using-Empirically-Constructed-Lexical-Resources-for-Named-Entity-Recog.pdf), 17-27. doi: 10.4137/bii.s11664

Kehagias, A., Petridis, V., Kaburlasos, V., & Fragkou, P. (2003). A Comparison of Word- and Sense-Based Text Categorization Using Several Classification Algorithms. *Journal of Intelligent Information Systems, 21*(3), 227-247. doi: 10.1023/a:1025554732352

Khan, A., Baharudin, B., & Khan, K. (2010, 15-17 June 2010). *Semantic based features selection and weighting method for text classification.* Paper presented at the Information Technology (ITSim), 2010 International Symposium in.

Kim, J. (2003). GENIA corpus-semantically annotated corpus for bio-textmining. *Bioinformatics, 19*(Suppl 1), i180 - i182.

Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y., & Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, 70 - 75.

Kohler, M. (2008). *Unified Medical Language System for Information Extraction*: VDM Publishing.

Leacock, C., Miller, G. A., & Chodorow, M. (1998). Using corpus statistics and WordNet relations for sense identification. *Comput. Linguist., 24*(1), 147-165.

Lin, D. (1998). *Automatic retrieval and clustering of similar words*. Paper presented at the Proceedings of the 17th international conference on Computational linguistics - Volume 2, Montreal, Quebec, Canada.

Liu, H., Wu, S. T., Li, D., Jonnalagadda, S., Sohn, S., Wagholikar, K., . . . Chute, C. G. (2012). *Towards a semantic lexicon for clinical natural language processing.* Paper presented at the AMIA Annual Symposium Proceedings 2012.

Liu, Y., Loh, H. T., & Lu, W. F. (2008). Deriving Taxonomy from Documents at Sentence Level *Emerging Technologies of Text Mining: Techniques and Applications* (pp. 99-119): IGI Global.

Luo, Z., Duffy, R., Johnson, S., & Weng, C. (2010). Corpus-based Approach to Creating a Semantic Lexicon for Clinical Research Eligibility Criteria from UMLS. *AMIA Summits Transl Sci Proc, 2010*, 26-30.

Luo, Z., Johnson, S. B., Lai, A. M., & Weng, C. (2011). Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. *AMIA Annu Symp Proc, 2011*, 843-852.

Luo, Z., Johnson, S. B., & Weng, C. (2010). *Semi-Automatically Inducing Semantic Classes of Clinical Research Eligibility Criteria Using UMLS and Hierarchical Clustering* (Vol. 2010).

Luo, Z., Miotto, R., & Weng, C. (2013). A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. *J Biomed Inform, 46*(1), 33-39. doi: 10.1016/j.jbi.2012.07.006

Luo, Z., Yetisgen-Yildiz, M., & Weng, C. (2011). Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. *J Biomed Inform, 44*(6), 927-935. doi: 10.1016/j.jbi.2011.06.001

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press.

Mansuy, T. N., & Hilderman, R. J. (2006). *Evaluating WordNet Features in Text Classification Models.* Paper presented at the FLAIRS Conference.

Nai-Lung, T., Wible, D., & Chin-Hwa, K. (2003, 26-29 Oct. 2003). *Feature expansion for word sense disambiguation.* Paper presented at the 2003 International Conference on Natural Language Processing and Knowledge Engineering,.

National Cancer Institute. (2001). *Cancer Clinical Trials The Basic Workbook*: National Cancer Institute.

National Cancer Institute. (2001). *Cancer Clinical Trials The In-Depth Program*: National Cancer Institute.

National Cancer Institute. (2013). Fact Sheets: Clinical Trials.

National Cancer Institute. (2013 Dec). NCI Dictionary of Cancer Terms, from http://www.cancer.gov/dictionary

National Institutes of Health. from http://www.nhlbi.nih.gov/health/health-topics/topics/clinicaltrials/

National Institutes of Health. The Need for Awareness of Clinical Research, from http://www.nih.gov/health/clinicaltrials/providers/awareness.htm

National Institutes of Health, N. C. I. (1997). *Results from Quarterly Omnibus Survey: Clinical Trials Questions-April 22*. Bethesda, MD.

National Library of Medicine. from http://www.nlm.nih.gov/medlineplus/tutorials/clinicaltrials/

National Library of Medicine. (2009). *UMLS Reference Manual*

National Research Council. (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: The National Academies Press.

Nguyen, A. N., Lawley, M. J., Hansen, D. P., Bowman, R. V., Clarke, B. E., Duhig, E. E., & Colquist, S. (2010). Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association, 17*(4), 440-445. doi: 10.1136/jamia.2010.003707

Nussenblatt, R. B., & Meinert, C. L. (2010). The status of clinical trials: cause for concern. *J Transl Med, 8*, 65. doi: 10.1186/1479-5876-8-65

Ohno-Machado, L. (2011). Electronic health records and computer-based clinical decision support: are we there yet? *Journal of the American Medical Informatics Association, 18*(2), 109. doi: 10.1136/amiajnl-2011-000141

Patel, C., Cimino, J., Dolby, J., Fokoue, A., Kalyanpur, A., Kershenbaum, A., . . . Srinivas, K. (2007). Matching Patient Records to Clinical Trials Using Ontologies (I. R. Division, Trans.) *IBM Research Report*.

Patel, C., Gomadam, K., Khan, S., & Garg, V. (2010). TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records. *Web Semantics: Science, Services and Agents on the World Wide Web, 8*(4), 342-347. doi: http://dx.doi.org/10.1016/j.websem.2010.08.004

Patel, C., Gomadam, K., Khan, S., & Garg, V. (2012). *TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records* (Vol. 8): Elsevier.

Patrick, J. D., Nguyen, D. H. M., Wang, Y., & Li, M. (2011). A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association, 18*(5), 574-579. doi: 10.1136/amiajnl-2011-000302

Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics, 40*(3), 288-299. doi: 10.1016/j.jbi.2006.06.004

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). *WordNet::Similarity: measuring the relatedness of concepts*. Paper presented at the Demonstration Papers at HLT-NAACL 2004, Boston, Massachusetts.

Penberthy, L., Brown, R., Puma, F., & Dahman, B. (2010). Automated matching software for clinical trials eligibility: Measuring efficiency and flexibility. *Contemporary Clinical Trials, 31*(3), 207-217. doi: 10.1016/j.cct.2010.03.005

Peto R, P. M., Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer 34*(6), 585-612.

Pratt, W., & Fagan, L. (2000). The usefulness of dynamically categorizing search results. *J Am Med Inform Assoc, 7*(6), 605-617. doi: citeulike-article-id:455128

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., & Jimeno, A. (2008). Text processing through Web services: Calling Whatizit. *Bioinformatics, 24*(2), 296 - 298.

Rebholz-Schuhmann, D., Kirsch, H., Gaudan, S., Arregui, M., & Nenadic, G. (2006). Annotation and Disambiguation of Semantic Types in Biomedical Text: a Cascaded Approach to Named Entity Recognition. *Workshop on Multi-Dimensional Markup in NLP, EACL*.

Remus, R., & Rill, S. (2013). Data-Driven vs. Dictionary-Based Word n-Gram Feature Induction for Sentiment Analysis *Language Processing and Knowledge in the Web*. Berlin Heidelberg: Springer

Riloff, E. (1996). An empirical study of automated dictionary construction for information extraction in three domains. *Artif. Intell., 85*(1-2), 101-134. doi: 10.1016/0004-3702(95)00123-9

Roberts, K., & Harabagiu, S. M. (2011). A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association, 18*(5), 568-573. doi: 10.1136/amiajnl-2011-000152

Rosenbloom, S. T., Denny, J. C., Xu, H., Lorenzi, N., Stead, W. W., & Johnson, K. B. (2011). Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association, 18*(2), 181-186. doi: 10.1136/jamia.2010.007237

Ross, J., Tu, S., Carini, S., & Sim, I. (2010). Analysis of eligibility criteria complexity in clinical trials. *AMIA Summits Transl Sci Proc, 2010*, 46-50.

Rosso, P., Ferretti, E., Jiménez, D., & Vidal, V. (2004). *Text Categorization and Information Retrieval Using Wordnet Senses.* Paper presented at the In Proceedings of the 2nd Global Wordnet Conference (GWC'04).

Sahami, M. (1999). *Using Machine Learning To Improve Information Access.* Ph.D, Stanford University.

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Saruladha, K., Aghila, G., & Raj, S. (2010, 9-11 Feb. 2010). *A Survey of Semantic Similarity Methods for Ontology Based Information Retrieval.* Paper presented at the Machine Learning and Computing (ICMLC), 2010 Second International Conference on.

Scott, S., & Matwin, S. (1998, August). *Text Classification Using WordNet Hypernyms.* Paper presented at the Workshop on usage of WordNet in NLP Systems (COLING-ACL '98).

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys 34*(1), 1-47.

Séroussi, B., & Bouaud, J. (2003). Using OncoDoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials. *Artif. Intell. Med., 29*(1-2), 153-167. doi: 10.1016/s0933-3657(03)00040-x

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal, 30*(1).

Sibbald, B., & Roland, M. (1998). Understanding controlled trials: Why are randomised controlled trials important? *BMJ, 316*(7126), 201. doi: 10.1136/bmj.316.7126.201

Spilker, B., & Cramer, J. A. (1992). *Patient Recruitment in Clinical Trials*: Raven Press

Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association, 17*(6), 646-651.

Stanford CoreNLP. from http://nlp.stanford.edu/software/corenlp.shtml

Stenner, S. P., Johnson, K. B., & Denny, J. C. (2012). PASTE: patient-centered SMS text tagging in a medication management system. *Journal of the American Medical Informatics Association, 19*(3), 368-374. doi: 10.1136/amiajnl-2011-000484

Torii, M., Wagholikar, K., & Liu, H. (2011). Using machine learning for concept extraction on clinical documents from multiple data sources. *Journal of the American Medical Informatics Association, 18*(5), 580-587. doi: 10.1136/amiajnl-2011-000155

Tsai, R., Wu, S., Chou, W., Lin, Y., He, D., Hsiang, J., Hsu, W. (2006). Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics, 7*, 92.

Tu, S. W., Peleg, M., Carini, S., Bobak, M., Ross, J., Rubin, D., & Sim, I. (2011). A practical method for transforming free-text eligibility criteria into computable criteria. *Journal of Biomedical Informatics, 44*(2), 239-250. doi: 10.1016/j.jbi.2010.09.007

UC Davis Cancer Center study. Understanding Cancer Patients' Needs, Concerns Is Key to Improving Clinical Trial Participation, from http://www.ucdmc.ucdavis.edu/publish/news/newsroom/2628

Wilbur, W., Hazard, G., Divita, G., Mork, J., Aronson, A., & Browne, A. (1999). *Analysis of biomedical text for chemical names: a comparison of three methods.* Paper presented at the AMIA Annual Symposium Proceedings 1999.

Xu, Y., Hong, K., Tsujii, J., & Chang, E. I.-C. (2012). Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association, 19*(5), 824-832. doi: 10.1136/amiajnl-2011-000776

Xu, Y., Tsujii, J., & Chang, E. I.-C. (2012). Named entity recognition of follow-up and time information in 20 000 radiology reports. *Journal of the American Medical Informatics Association, 19*(5), 792-799. doi: 10.1136/amiajnl-2012-000812

APPENDIX: 5 Sample Matching Results between Patient and Clinical Trials (Essay III)

Patient ID	Patient_text	CT_text	Matched Terms	Expanded Matched Terms	Matching evaluation	Comments (Optional)
1000001	Right breast, partial/simple mastectomy: - Breast tissue with proliferative fibrocystic changes. - Residual areas of lobular cancerization - No definite residual DCIS - Previous biopsy site changes. - Margins free of involvement. -Prognostic factors performed on previous biopsy CS08-12007: ER +(97%), PR+(85%) ,PREOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. POSTOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. PROCEDURE: Right mastectomy with level I axillary node excision. ,1. This is a 00-year-old woman, who has had right mastectomy for recently diagnosed ductal carcinoma in situ of the right breast that also has features of lobular cancerization. She is stage pT1N0M0 and I recommend hormonal therapy with tamoxifen 20 mg daily for 5	Inclusion Criteria: 1. Pathologically confirmed ductal carcinoma in situ of the breast or early invasive breast cancer defined as pathologic stage Tis, T1, or T2, N0, N1mic, or N1a (pathologic staging of the axilla is required for all patients with invasive disease but is not required for patients with DCIS only). 2. Treatment with breast conserving surgery. 3. Final surgical margins must be negative, defined as no evidence for ductal carcinoma in situ or invasive breast cancer touching the inked surgical margin. If the invasive or in situ breast cancer approaches within less than 1 mm of the final surgical margin, then a reexcision is strongly encouraged. Lobular carcinoma in situ at the final surgical margin will be disregarded. 4. Age 40 years or older. This age cutoff is justified because breast cancers in women under the age of 40 are known to have a significantly higher risk of IBTR presumably due to underlying biologic differences. 5. Female sex. 6. Attending radiation oncologist declares intention to treat the whole breast only and that a third radiation field to treat regional lymph nodes is not planned (radiation of the undissected level I/II axilla with high tangents is allowed). 7. If the patient has a history of a prior non-breast cancer, all treatment for this cancer must have been completed prior to study registration and the patient must have no evidence of disease for this prior non-breast	ductal carcinoma in situ right breast T1 level I/II axilla DCIS ductal carcinoma	Duct_adenocarcinoma Duct_carcinoma Duct_cell_carcinoma Axillary_fossa Axilla_structure Axilla Armpit Structure_of_axillary_fossa Axillary_region Axillary Axillary_region_structure	Very Good	

		cancer. 8. Patients must be enrolled on the trial within 12 weeks of the later of two dates: the final breast conserving surgical procedure or administration of the last cycle of cytotoxic chemotherapy. Exclusion Criteria: 1. Pathologic or clinical evidence for a stage T3 or T4 breast cancer. 2. Pathologic evidence for involvement of 4 or more axillary lymph nodes, or imaging evidence of involvement of infraclavicular, supraclavicular, or internal mammary lymph nodes. 3. Clinical or pathologic evidence for distant metastases. 4. Any prior diagnosis of invasive or ductal carcinoma in situ breast cancer in either breast. 5. Current diagnosis of bilateral breast cancer. 6. History of therapeutic irradiation to the breast, lower neck, mediastinum or other area in which there could potentially be overlap with the affected breast. 7. Patients not fluent in English or Spanish. (The Informed Consent will be available in these two languages) 8. Patient is pregnant.				
2000024	Subtle nodularity in the central subareolar region of the left breastSubtle nodularity in the central subareolar region of the left breast is identified as mildly prominent ductal elements. There is no suspicious finding within the left breast. 2. Post lumpectomy change on the right with an interval change in the mammographic appearance of the right breast with an 8 mm poorly defined zone of nodularity seen in the 12 o'clock position within the right breast with accompanying calcifications. Treated with lumpectomy followed by radiotherapy. Ultrasound only questionably demonstrates a subtle zone of altered echotexture in this region. A discrete palpable lump is not identified on physical	Inclusion Criteria: 1. Women with a histological diagnosis of breast cancer experiencing edema in the ipsilateral arm such that there is a minimum 10% and maximum 40% increase in arm volume over the unaffected arm (mild to moderate lymphedema). 2. Patients must have completed all primary and adjuvant treatments (surgery, chemotherapy, radiotherapy) prior to randomization. 3. Patients must have their own fitted compression garment for daytime maintenance. 4. No past or current use of a night-time compression system for maintenance. Those patients who have trialed a night-time compression system in the past year must observe a six-month washout period before	radotherapy edema in the ipsilateral arm	Radiation_therapy Plesiotherapy_radiation Therapeutic_radiology Radiation_oncology Oedema Dropsy Hydrops Edematous Interstitial_edema Interstitial_oedema	Average	patient has History of rather than experiencing edema - Questionable imaging findings

	examination. History of edema in the ipsilateral arm SCREENING TO DIAGNOSTIC MAMMOGRAPHY AND BILATERAL BREAST ULTRASOUND: The patient presented for screening mammography. The breasts were imaged in the craniocaudal and MLO projections. Review of these images demonstrated an interval change with the appearance of a subtle zone of asymmetric density within the 12 o'clock position within the right breast with current.	entering the trial. Exclusion Criteria: 1. Clinical or radiological evidence of active disease, either local or metastatic. 2. History of contralateral breast cancer and axillary surgery. 3. Serious non-malignant disease, such as renal or cardiac failure, which would preclude daily treatment and follow-up. 4. Patients for whom compression is contraindicated. 5. Psychiatric or addictive disorders which preclude obtaining informed consent or adherence to the protocol. 6. Unable to comply with the protocol, measurement and follow-up schedule.		Oedematous Edema_-_lesion Oedema_-_lesion Edema_-_symptom Oedema_-_symptom		
2000001	Radical mastectomy : - Breast tissue with proliferative fibrocystic changes. - Residual areas of lobular cancerization - Positive axillary lymph nodes. - Margins free of involvement. -Prognostic factors performed on previous biopsy CS08-12007: ER(-) / PR(-), PREOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. POSTOPERATIVE DIAGNOSIS: Ductal carcinoma in situ, right breast, status post core biopsy, intermediate grade, cribriform type ER-PR positive. PROCEDURE: Right mastectomy. This is a 00-year-old woman, who has had right mastectomy for recently diagnosed ductal carcinoma in situ of the right breast that also has features of lobular cancerization.	Inclusion Criteria: 1. Patient must accept the modified radical mastectomy 2. Patients with histologically confirmed ER(-) PR(-) and HER-2(-) 3. Positive axillary lymph nodes;negative axillary lymph node with age < 35 years or Ⅲ grade or intravascular cancer embolus. 4. Age between 18 years to 65 years 5. Able to give informed consent 6. Patients with an Eastern Cooperative Oncology Group (ECOG) performance score of 0 or 1. 7. Not pregnant, and on appropriate birth control if of child-bearing potential. 8. Adequate bone marrow reserve with ANC > 1000 and platelets > 100,000. 9. Adequate renal function with serum creatinine < 2.0. 10. Adequate hepatic reserve with serum bilirubin < 2.0, AST/ALT < 2X the upper limit of normal, and alkaline phosphatase < 5X the upper limit of normal. Serum bilirubin > 2.0 is acceptable in the setting of known Gilbert's syndrome. 11. No active major medical or psychosocial problems that could be complicated by study participation. Exclusion Criteria: 1. received neo-adjuvant therapy	Radical mastectomy Positive axillary lymph nodes ER(-) PR(-)	Mammectomy Excision_of_breast_tissue Axillary_fossa Axilla_structure Axilla Armpit Structure_of_axillary_fossa Axillary_region Axillary Axillary_region_structure Structure_of_lymph_node Lymph_node_structure Lymph_node	Average	There is concept match on mastectomy, ER, PR, and positive LN, however the patient had mastectomy already and CT requires that patient accepts mastectomy. Missing performance status

		2. Cardiac dysfunction documented by an ejection fraction less than the lower limit of the facility normal by multi-gated acquisition (MUGA) scan, or 45% by echocardiogram. -The rate of Disease recurrence 3. Uncontrolled medical problems. 4. Evidence of active acute or chronic infection. 5. Pregnant or breast feeding. 6. Hepatic, renal, or bone marrow dysfunction as detailed above.		Lymph_gland Lymphatic_gland		
2000002	Follow-up with surgical consultation: Postoperative changes and postradiation changes left breast. No ,Invasive, moderately differentiated, ductal carcinoma, mBR Grade II Negative for lymphovascular space invasion. - Microcalcification within tumor. ,Infiltrating lobular carcinoma, mBR Grade I, 1 cm, in a random section from the lower outer quadrant (prognostic factors pending), located approximately 2 cm from the recent biopsy site. 2.Microscopic focus of residual infiltrating duct carcinoma, mBR Grade II, 0.1 cm, adjacent to biopsy cavity (previous stereotactic biopsy, CS-08-10468, showed 0.8 cm tumor). Right axillary lymph nodes metastasis. - Prognostic factors performed on previous biopsy ER 100%, PR 92%, Her-2/neu 2+ (not amplified by SISH). 3.Biopsy-related changes with patchy adjacent atypical duct hyperplasia and fibrocystic change with associated microcalcifications. 4. One benign intramammary lymph node. 5. Skin and nipple negative for malignancy. 6. Margins of resection negative for atypia and ma	Inclusion Criteria: 1. The participant has histopathologically-confirmed primary breast cancer in Japanese. 2. The participant is aged 20 years or older when informed consent is obtained 3. The participant has estrogen receptor (ER)-positive tumor cells and/or progesterone receptor (PgR)-positive primary tumor. And HER-2 is negative. 4. The participant has breast cancer in the clinical stages of T1-T3, N-any and M0 by TNM classification (the seventh edition, proposed by UICC in 2009). (No distant metastasis to lung, liver and bone should be confirmed on the image-based diagnosis at study enrollment. The image taken within 12 weeks prior to study enrollment is also available for the diagnosis.) The number of axillary lymph node metastasis is not limited. 5. Any operative procedure for breast cancer is acceptable. In principle, after breast-conserving surgery, the participant will receive postoperative radiation to the conserving breast. 6. Neoadjuvant chemotherapy and adjuvant chemotherapy prior to study enrollment are acceptable. (It is advisable the same kind of chemotherapy is performed at each site.) 7. The participant has a history of regular menstrual periods within 12 weeks prior to study enrollment, or the participant has	ER PR HER-2 axillary lymph nodes metastasis operative procedure	Axillary_fossa Axilla_structure Axilla Armpit Structure_of_axillary_fossa Axillary_region Axillary Axillary_region_structure Structure_of_lymph_node Lymph_node_structure Lymph_node Lymph_gland Lymphatic_gland Operative_procedure Surgery Surgical	Average	Match on Confirmed diagnosis and ER, PR, Her status, and ax LNs. However, we are missing data on menopausal status and performance status. Patient has post radiation changes suggesting that she received RT.

		FSH of less than 40 mIU/mL and E2 of 10 pg/mL or more measured within 12 weeks prior to study enrollment. The participant has not had a chemical menopause (i.e., FSH of less than 40 mIU/mL and E2 of 10 pg/mL or more) within 12 weeks after completing adjuvant chemotherapy. 8. The participant is in a condition to receive study drug and Tamoxifen (TAM) within 12 weeks after surgery or after adjuvant chemotherapy prior to study enrollment. Adjuvant chemotherapy prior to study is required to have been completed at the time of study enrollment. 9. The participant has ECOG performance status of grades 0 or 1 at the time of study enrollment. 10. The participant meets the following criteria of hepatic, renal and bone marrow functions on the laboratory test results at screening: - Hepatic function: AST (GOT) ≤ 3.0 times the upper limit of normal (ULN) ALT (GPT) ≤ 3.0 times the ULN - Renal function: serum creatinine level < 1.5 times the ULN - Bone marrow function : white blood cell count ≥ 3,000/mm3 platelet count ≥ 100,000/μL hemoglobin ≥ 10.0g/dL 11. The participant agrees to use a non-hormonal method of contraception through the study period. Exclusion Criteria: 1. The participant has received neoadjuvant or adjuvant hormonal therapy for the latest breast cancer surgery. 2. The participant has received bilateral oophorectomy and bilateral ovarian irradiation. 3. The participant has inflammatory breast cancer or bilateral breast cancer. 4. The participant has non-invasive ductal carcinoma. 5. The participant has multiple primary cancers, or a history of carcinoma in other organs. 6. The participant is pregnant or breast-feeding.		Operation Surgical_procedure Operative_procedures Operations_by_method		

		7. The participant has a history of hypersensitivity to synthetic LH-RH, LH-RH derivative, TAM, TAM analogue (antiestrogen) or any component of the study drug. 8. The participant has a history of, or has been diagnosed with thromboembolism including myocardial infarction, cerebral infarction, venous thrombosis, and pulmonary embolism, or cardiac failure. 9. Patients whose QTcF interval exceeded 460 msec on the 12-lead electrocardiogram at screening.				
2000134	,Document Type: Surg Path Final Report Document Date: 2010 Document Status: Auth (Verified) Performed by/Author: XXXX RT on 2010 Verified By: XXXX MD on 2010 Encounter info: 0000000000,, COL, Outpatient, 2010 - 2010 * Final Report * Specimen: (Verified) A U/S core bx left breast 14g B U/S core bx left breast 14g C U/S SUROS left breast Clinical Information: (Verified) A) U/S core biopsy left breast, 14g, location ? 1:30 lateral. Size 0.8 cm. Left breast mass. Rad diff dx: Favor invasive CA. B) U/S core biopsy left breast, 14g, location ? 1:30 medial. Size 1.1 cm. Left breast mass. Rad diff dx: Invasive CA strongly favored. C) U/S SUROS left breast, 9g vacuum assisted. Location ? 3:00, size 6 mm. Note is made that the patient has undergone a prior right-sided lumpectomy. in 1999 Left breast mass. Rad diff dx: Favor invasive CA vs FCC with fibrosis. Post radiation changes left breast Invasive ductal carcinoma in situ with lobular features. Gross Description: (Verified) Specimen A: Specimen received fresh and placed in formalin (on 2010	DISEASE CHARACTERISTICS: - Female patients newly diagnosed with breast carcinoma including ductal carcinoma in situ (DCIS) - Stage 0-IIIA disease - Status post-lumpectomy, -quadrantectomy, or -mastectomy - Plan to receive adjuvant radiation to the whole breast or chest wall and/or regional lymph nodes - No sites that cannot send blood/urine specimens to Wake Forest by overnight (next day) express shipping PATIENT CHARACTERISTICS: - *This stratum is closed as of April 25, 2012. - No patients who do not understand English and are unable to complete form with assistance PRIOR CONCURRENT THERAPY: - Total dose > 40 Gy, dose per fraction > 1.8 - 2.0 Gy, use of 2D, 3D-conformal, or intensity-modulated radiation therapy (IMRT) treatment techniques allowed; a daily fraction of 2.7 Gy to the whole breast is suggested for hypofractionated regimens - Concurrent and sequential boost techniques are allowed for both standard and hypofractionated regimens - Adjuvant hormonal therapy will be allowed prior to, during, and/or after radiotherapy (RT) at the discretion of a medical oncologist - Targeted therapies,	lumpectomy radiation ductal carcinoma in situ	Tylectomy Tylectomy_of_breast Lumpectomy_of_breast Breast_lumpectomy Excision_of_breast_lump Excision_of_lesion_of_breast Radiation_therapy Plesiotherapy_radiation Therapeutic_radiology Radiation_oncology Duct_adenocarcinoma Duct_carcinoma Duct_cell_carcinoma	Average	Matches on Histology (?invasive), lumpectomy, and radiation. However, though not completely clear, pat has post radiation changes suggesting htat she already receave RT

		such as Herceptin, will be allowed prior to, during, and/or after RT at the discretion of the medical oncologist - No prior radiation to the involved breast or chest wall - No concurrent chemotherapy - No patients who underwent breast reconstruction following mastectomy - Placement of tissue expanders and implants are not allowed - No patients who have undergone MammoSite® or any other form of brachytherapy as well as those who will be treated with skin-sparing IMRT - Patients may not be concurrently enrolled in a protocol that involves treatment of the skin, i.e., applying lotions/moisturizers - Protocols that do not involve treatment of the skin are allowed		Ductal_carcinoma		

CURRICULUM VITAE

EUISUNG JUNG

University of Wisconsin-Milwaukee

EDUCATION

August 2015 ***Ph.D in Information Technology Management***
Minor : Computer Science
University of Wisconsin-Milwaukee
Sheldon B. Lubar School of Business

2002 ***Master of Business Administration (concentration in MIS)***
Kyungpook National University, Daegu, Korea.
Title of Thesis : *Relationship between Internet-based B2B system and BPI*

2000 ***Bachelor of Business Administration***
Kyungpook National University, Daegu, Korea.

DISSERTATION TITLE

Three Essays on Enhancing Clinical Trial Subject Recruitment Using Natural Language Processing and Text Mining

JOURNAL PUBLICATIONS

- Kwak, D.-H., Kizzier, D., Zo, H., and **Jung, E.** 2012. "Cross-Cultural Investigation of Security Knowledge Process," *International Journal of Business Information Systems* (10:1), pp. 1-19.
- Kwak, D.-H., Kizzier, D., Zo, H., and **Jung, E.** 2011. "Understanding Security Knowledge and National Culture: A Comparative Investigation between Korea and the U.S," *Asia Pacific Journal of Information Systems*, (21:3), pp. 51-69.
- **Euisung Jung**, Changkyo Suh, , "Effect of internet-based B2B system on BPI", *Economics and Business Papers*, 32(1), 101-119, The Institute of Economics and Business Research Kyungpook National University, 2004.08.23

TEACHING EXPERIENCE

2014 - Present **Assistant Professor**
College of Business and Innovation, The University of Toledo
- BUAD3050 Information Technology Management
- INFS3780 / 6930 Enterprise Resource Planning

2014 Spring **Adjunct Faculty**
Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee
- BUS ADM 532 Web Development for Open Business Systems

PROFESSIONAL ACTIVITIES

Reviewer
- Journal of Information Technology Theory and Application (JITTA) 2013
- International Conference on Information Systems (ICIS), 2012, 2013, 2014
- Americas Conference on Information Systems (AMCIS), 2010, 2011, 2012, 2013
- Hawaii International Conferences on System Sciences (HICSS), 2011

HONORS AND AWARDS

- Sheldon B. Lubar Doctoral Scholarship, University of Wisconsin–Milwaukee (2013)
- Business Advisory Council Doctoral Scholarship, University of Wisconsin – Milwaukee (2012)
- Graduate Student Travel Award, University of Wisconsin – Milwaukee (2010, 2011, 2012, 2013)
- Innovative Research Award, National Research Council for Economics, Humanities and Social Sciences (2007)
- Award for meritorious service, President of Korea Environment Institute, 2007
- Outstanding Research Scholarship, Kyungpook National University (2000)
- Winner of the semifinals, National University Armature Tennis Tournament (1998)
- Exemplary soldier award, Commander of Republic of Korea Army 22 Division (1996)
- Academic Excellent Scholarship, Kyungpook National University, (1993)