SURVEILLANCE IN THE INFORMATION AGE:

TEXT QUANTIFICATION, ANOMALY DETECTION, AND

EMPIRICAL EVALUATION

by

Hsin-Min Lu

A Dissertation Submitted to the Faculty of the

Committee On Business Administration

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN MANAGEMENT

In the Graduate College

THE UNIVERSITY OF ARIZONA

2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

THE UNIVERSITY OF ARIZONA
GRADUDATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Hsin-Min Lu
entitled Surveillance in the Information Age: Text Quantification, Anomaly Detection, and Empirical Evaluation
and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____Date: 4/28/2010
Hsinchen Chen

_____ Date: 4/28/2010
Daniel Zeng

_____ Date: 4/28/2010
Paulo Goes

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____Date: 4/28/2010
Dissertation Direction: Hsinchen Chen

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Hsin-Min Lu

ACKNOWLEDGEMENTS

I am grateful to my advisor, Dr. Hsinchun Chen, for his guidance and encouragement throughout my five-year doctoral study at the University of Arizona. Under his direction, the learning experience in the Artificial Intelligence Lab is invaluable to me. I am thankful for his trust in me and the opportunities he gave me to grow not just as a researcher but also as a person. I am also thankful to my committee members Dr. Daniel Zeng, Dr. Paulo Goes, and Dr. George Jiang, for their guidance and encouragement. I also thank other MIS faculty members for their support during my studies.

My dissertation has been partially supported by National Science Foundation (NSF) program: "A National Center of Excellence for Infectious Disease Informatics" #IIS-0428241 and #IIS-0839990, US Department of Homeland Security (DHS): "DHS Center of Excellence in Border Security and Immigration" #2008-ST-061-BS0002, and Arizona Department of Health Services. My work at the Artificial Intelligence Lab has been supported by many colleagues. I thank Cathy Larson for her encouraging words throughout my tenure in the lab. I would also like to thank my colleagues and friends Yida Chen, Xin Li, Chunju Tseng, Yan Dang, Chun-Neng Huang, Aaron Sun, Ping Yan, Shuo Zeng, Yulei Zhang, David Zimbra, Ximing Yu, Li Fan, Victor Benjamin, Andy Pressman, and David Ware for their support of my studies.

Lastly, I owe my deepest gratitude to my wife, Wan-Hsin Nina Huang, my brothers, Hsin-Chia Lu and Shin-Hou Lu, and my parents, Taiflu Lu and Ai Huang, and the rest of my family. This thesis would not have been possible without their constant support.

DEDICATION

This dissertation is dedicated to my family.

TABLE OF CONTENTS

TABLE OF CONTENTS - *Continued*

TABLE OF CONTENTS - *Continued*

LIST OF ILLUSTRATIONS

LIST OF TABLES

ABSTRACT

Deep penetration of personal computers, data communication networks, and the Internet has created a massive platform for data collection, dissemination, storage, and retrieval. Large amounts of textual data are now available at a very low cost. Valuable information, such as consumer preferences, new product developments, trends, and opportunities, can be found in this large collection of textual data. Growing worldwide competition, new technology development, and the Internet contribute to an increasingly turbulent business environment. Conducting surveillance on this growing collection of textual data could help a business avoid surprises, identify threats and opportunities, and gain competitive advantages.

Current text mining approaches, nonetheless, provide limited support for conducting surveillance using textual data. In this dissertation, I develop novel text quantification approaches to identify useful information in textual data, effective anomaly detection approaches to monitor time series data aggregated based on the text quantification approaches, and empirical evaluation approaches that verify the effectiveness of text mining approaches using external numerical data sources.

In Chapter 2, I present free-text chief complaint classification studies that aim to classify incoming emergency department free-text chief complaints into syndromic categories, a higher level of representation that facilitates syndromic surveillance. Chapter 3 presents a novel detection algorithm based on Markov switching with jumps

models. This surveillance model aims at detecting different types of disease outbreaks based on the time series generated from the chief complaint classification system.

In Chapters 4 and 5, I studied the surveillance issue under the context of business decision making. Chapter 4 presents a novel text-based risk recognition design framework that can be used to monitor the changing business environment. Chapter 5 presents an empirical evaluation study that looks at the interaction between news sentiment and numerical accounting earnings information. Chapter 6 concludes this dissertation by highlighting major research contributions and the relevance to MIS research.

CHAPTER 1. INTRODUCTION

Deep penetration of personal computers, data communication networks, and the Internet has created a massive platform for data collection, dissemination, storage, and retrieval. Every day people engage in numerous online activities, including reading the news and product reviews, commenting on developing events, buying and selling stocks, and widening their social networks. This widespread engagement with online worlds has facilitated the creation of large amounts of textual data.

Organizations can benefit from this massive collection of textual data if effective approaches can be used to quantify it and monitor meaningful changes. Firms can react to potential threats promptly if an information system can automatically conduct surveillance on news reports from a broad range of sources. Decision makers can respond to investment opportunities in a timely manner if effective surveillance approaches can be applied on newswires, news papers, and forum postings.

The concept of surveillance has existed for a long time. A well-known example is the statistical surveillance methods used in factories for production quality monitoring. Necessary steps can be taken once deviations are detected. Direct application of existing surveillance approaches to a growing collection of textual data, nonetheless, is not possible due to the unique characteristics of textual data. The underlying textual data need to be converted to a suitable representation before applying the surveillance approaches. Moreover, new surveillance approaches may need to be developed to achieve satisfactory outcomes. This dissertation has been motivated by the need to develop novel

surveillance frameworks in modern business environment that addresses the issues associated with large collections of textual data. Specifically, my research topics focus on:

- Advancing representation, analysis, and modeling of massive amounts of textual data in significant application areas to provide effective decision support.
- Understanding the information dissemination and digestion processes that involve textual data.

1.1 Research Framework

As shown in Figure 1.1, my research has three main focuses: text quantification, anomaly detection, and empirical evaluation. The first focus, text quantification, deals with the problem of extracting and quantifying valuable information from textual data. While traditional information retrieval (IR) (van Rijsbergen 1979) and information extraction (IE) (Sarawagi 2008) approaches may be useful for rudimentary text quantification tasks, more complicated text quantification problems may require the development of novel approaches that combine advanced statistical machine learning approaches (Bishop 2006) with IR and IE approaches.

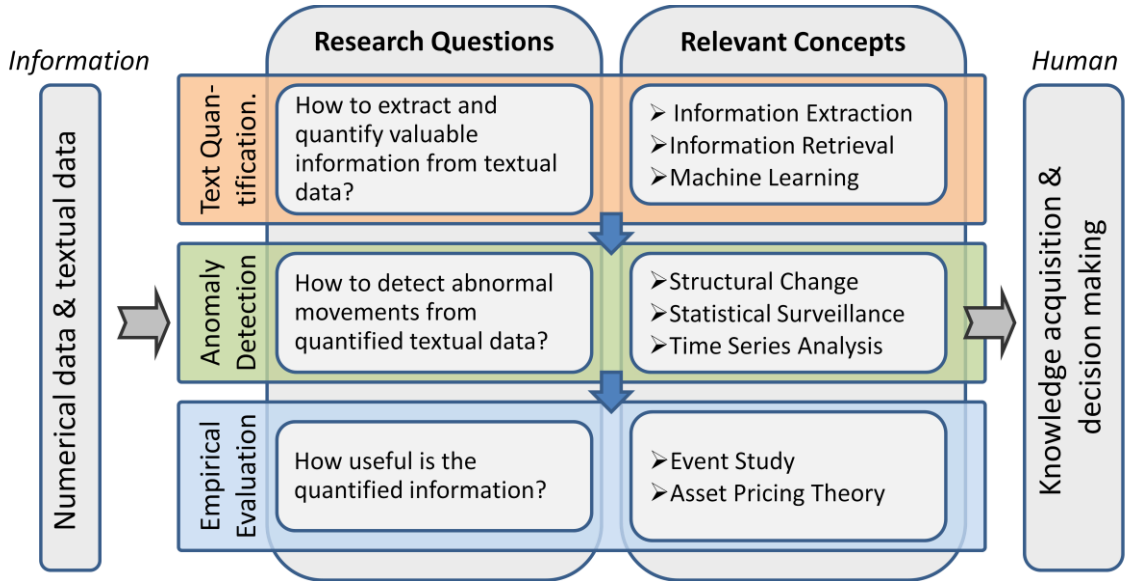


Figure 1.1 Research Framework

The second focus, anomaly detection, addresses the problem of detecting abnormal patterns from quantified textual data in both prospective and retrospective settings. Traditional statistical tests for structure changes focus mostly on retrospective analysis (Chow 1960). Statistical surveillance methods often assume incoming observations to be identically and independently distributed (Montgomery 2005), which is unrealistic for many real-world applications. The development of novel anomaly detection approaches is warranted given the deficiency of existing surveillance approaches.

The third focus, empirical evaluation, aims at studying the importance of the quantified information by triangulating the quantified textual information with external numerical data. The basic idea is that open-source textual data are available to almost everybody at a very low cost. Rational decision makers will utilize the information to

support their own decisions. The reactions of these decision makers then can be used to verify the impact of the quantified textual data.

1.2 Areas of Study

I have applied my framework to two important areas: the development of syndromic surveillance systems and text mining for financial decision making. The goal of syndromic surveillance systems is to identify potential disease outbreaks in a timely manner through the development of novel information technologies. Text mining for financial decision making focuses on leveraging textual data to improve the understanding of various financial issues. As summarized in Figure 1.2, each subsequent Chapter presents a study that focuses on answering one of the three research questions in selected areas.

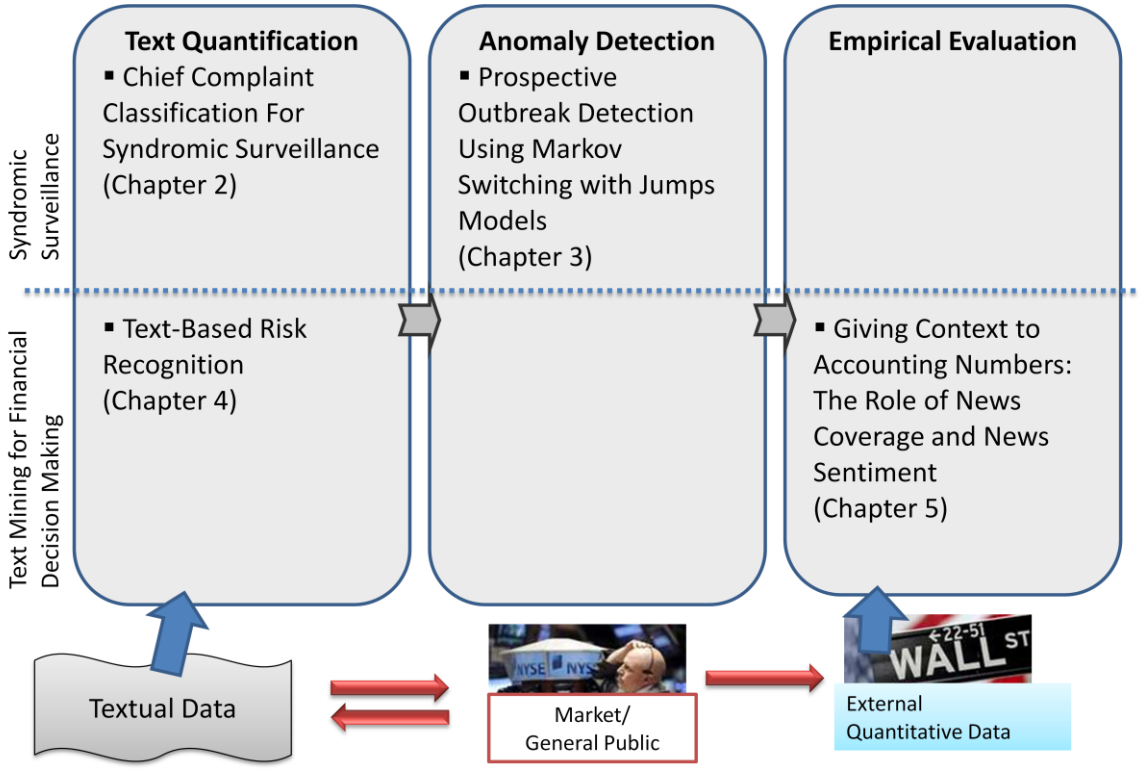


Figure 1.2 Areas of Study

In Chapter 2, I present the text quantification research in the context of syndromic surveillance. To support timely detection of disease outbreaks, syndromic surveillance systems look at emergency department free-text chief complaints (CCs) instead of confirmed cases or diagnostic data. CCs are often taken verbatim as patients describe their problems and are often individually typed as free-text entries, sometimes by trained healthcare personnel and sometimes by hospital staff without medical training. This results in geographic-, facility-, and individual-level differences in synonyms, acronyms, abbreviations, spelling, and truncations of the patients' CCs. The non-standard nature of CCs hinders the subsequent aggregation and analysis. Moreover, medical practitioners in different regions may record CCs in languages other than English. The language differences further deepen the challenge of processing CCs from different regions or countries.

This chapter presents a new CC classification approach that can handle free-text entries in English and Chinese. The basic idea is to first construct the CC classification system for English CCs. CCs recorded in Chinese are handled using a translation layer that maps Chinese to English.

The issue of non-standard English CCs is handled by my novel ontology-enhanced automatic CC classification approach. Exploiting semantic relations in a medical ontology, this approach seeks to address the CC vocabulary variation problem in general

and to meet the specific need for a classification approach capable of handling multiple sets of syndromic categories.

To handle Chinese CCs, a set of 470 Chinese key phrases was extracted from about one million Chinese CC records using statistical methods. Based on the extracted key phrases, the system translates Chinese text into English and classifies the translated CCs into syndromic categories using the CC classification system constructed for English CCs.

Chapter 3 presents the research on anomaly detection under the same context. I developed a novel prospective disease outbreak detection approach based on the Markov switching with jumps model. This approach takes the aggregated syndromic count time series as the input and estimates whether slow-moving changes (such as those caused by infectious disease) and fast-moving changes (such as those caused by special events) have occurred. Most existing detection methods combine a time series filtering procedure followed by a statistical surveillance method. The performance of this "two-step" detection method is hampered by the unrealistic assumption that the training data are outbreak-free. Moreover, existing approaches are sensitive to extreme values, which are common in real-world data sets. My Markov switching with jumps model can address the shortcoming of existing "two-step" methods. A jump component is introduced to absorb sporadic extreme values that may otherwise weaken the ability to detect a slow-moving disease outbreak. My approach outperformed several state-of-the-art detection methods in terms of detection sensitivity using both simulated and real-world data.

The next two chapters present research conducted under the context of text mining for financial decision making. Chapter 4 reports on text quantification research that

focuses on risk-related information. Business documents often contain risk-related information directly relevant to a firm's future opportunity to grow and profit. Timely and comprehensive enterprise risk-related information can help a business make informed decisions and develop suitable plans and strategies. Systematic analysis of risk-related information often involves processing large amounts of documents, which is time consuming and costly. Informed by theories for decision making under uncertainty, I propose a conceptual model for risk recognition in business documents. Three text-based risk measures were proposed to signal risk-related information embedded in textual data. A design framework based on statistical machine learning approaches was developed to generate a class of IT artifacts for the text-based risk recognition problem. I conducted experiments using sentences from the Wall Street Journal to compare the performance of the proposed approaches against baseline models based on opinion mining tools and keyword matching techniques. The results indicate that statistical machine learning approaches can generate effective and parsimonious models that outperform baseline models. Moreover, elastic-net logistic regressions, which simultaneously select variables and estimate models, were identified as the most suitable approach for the underlying tasks. Given the potential benefit of accurately recognizing risk-related information in business documents, the results have important implications for researchers and practitioners who need to analyze risk-related information in large amounts of textual data.

Chapter 5 presents an empirical evaluation that looks at the interaction between sentiment and coverage of news articles and numerical accounting earnings information.

Accounting numbers such as earnings per share are an important source of information that conveys the value of firms. Previous studies on return-earnings relation have confirmed that stock prices react to the information content in accounting numbers. However, other information sources such as financial news may also contain value-relevant information and affect investors' reaction to earnings announcements. I quantify news coverage and news sentiment about S&P 500 constituents in the Wall Street Journal before earnings announcements and model their interaction with the return-earnings relation. My empirical results show a strong corroboration effect for positive news sentiment followed by positive earnings surprises. Moreover, negative news sentiment followed by positive earnings surprise exhibits a surprise effect. The results suggest that investors are sophisticated in considering managers' motivation behind accounting disclosure when exposed to multiple sources of information.

Chapter 6 of this dissertation highlights the major research contributions, relevance to MIS research, and some interesting future directions that are worth pursuing.

CHAPTER 2. FREE-TEXT CHIEF COMPLAINT CLASSIFICATION FOR SYNDROMIC SURVEILLANCE

Syndromic surveillance aims to detect early signs of natural disease outbreaks, bioterrorism attacks, or changes in public health status in a timely manner (Lewis *et al.* 2002). Instead of monitoring confirmed cases or waiting for diagnostic data, syndromic surveillance focuses primarily on pre-diagnostic health-related information in an effort to significantly shorten the time needed to detect unusual events worth further investigation (Mandl *et al.* 2004).

Emergency department (ED) triage free-text chief complaints (CCs) are short free-text phrases entered by triage personnel describing reasons for patients' ED visits. Symptoms, diseases, mechanisms of injury, and other medical or non-medical concepts are commonly seen in CCs. ED CCs are a popular data source used by many syndromic surveillance systems because of their timeliness and availability (Espino & Wagner 2001; Ivanov *et al.* 2002; Chapman *et al.* 2004; Chapman *et al.* 2005a; Chapman *et al.* 2005b). CCs are among the first data elements collected for any ED visit and many hospitals increasingly have free-text CCs available in electronic form.

For automatic capture of syndromic surveillance data, free-text CC records need to be systematically classified into syndromic categories according to the symptom-related information they contain. Temporal analysis of classified results then can be used for outbreak detection. In the early stages, many diseases have similar non-specific symptoms. Grouping CCs into syndromic categories or syndromes instead of specific symptoms may provide more informative indication of potential outbreaks (Lee *et al.*

2003; Chapman *et al.* 2005b). In effect, most existing syndromic surveillance systems accept this approach of classifying CCs into syndromic categories (Lombardo *et al.* 2003; Tsui *et al.* 2003; Yan *et al.* 2006). However, major technical challenges remain for automatic CC classification. CCs are often taken verbatim as patients describe their problems and are often individually typed as free text entries, sometimes by trained healthcare personnel and sometimes by hospital staff without medical training. This results in geographic, facility and individual level differences in synonyms, acronyms, abbreviations, spelling, and truncations of the patients' CCs.

The issue concerning the lack of a standard vocabulary for ED CCs can be addressed in different ways. A supervised learning method which learns from the pairs of raw CCs and assigned labels can entirely bypass this problem but requires a large manually-labeled training sample. Other approaches such as medical thesauri, spell checking algorithms, and manually-created synonym lists have also been tried in the past (Day *et al.* 2004; Shapiro 2004; Travers & Haas 2004; Thompson *et al.* 2006). The performance of these approaches, however, to a large degree depends on the CCs used to construct the synonym list or tune the system. These approaches may perform poorly with CCs that are different from those used in the system development and tuning. These existing approaches do not take advantage of the fact that medical terms appearing in CCs can be semantically related. I argue that by exploiting such semantic relations through the help of a medical ontology, the CC vocabulary problem (Travers 2003; Chapman 2006) can be better handled and in turn a more effective CC syndrome classification approach can be developed.

The use of ontologies has been discussed in the context of syndromic surveillance (Crubezy *et al.* 2005). The discussion has focused on the integration of different data sources into a unified problem-solving architecture as opposed to processing specific data sources such as free-text CCs. In this article, I propose an ontology-enhanced method to classify CCs into syndromic categories. At the core of this approach is a new grouping method based on weighted semantic similarity scores (WSSS) (Lu *et al.* 2006). Utilizing the semantic relationships from a medical ontology, this method can be effectively applied to process CC terms not covered by syndrome mapping rules or past CC records with known syndromic category associations. The CC classification subsystems from two syndromic surveillance systems, Early Aberration Reporting System (EARS) and Real-time Outbreak Detection System (RODS), are chosen as the benchmarks for performance comparison.

A reference standard dataset consisting of CCs and validated classification results is of critical importance in assessing the performance of any CC classification system. In my study, such a reference standard dataset with 1,000 records was constructed with help from three domain experts. This dataset was used to evaluate the performance of both my system and the benchmark systems.

The other import issue regarding chief complaint classification systems is the ability to process CCs in different languages. Despite the fact that syndromic surveillance is largely an international effort, existing CC classification systems do not provide adequate support for processing CCs recorded in non-English languages. I reported a multilingual CC classification effort, focusing on CCs recorded in Chinese. The

multilingual CC classification research was motivated to answer the following research questions: (a) How useful Chinese CCs are for syndromic surveillance and (b) Whether an effective cross-lingual approach can be developed leveraging existing English CC classification methods.

CCs from EDs in Taiwan were collected and analyzed in my research Medical practitioners in Taiwan are trained to record CCs in English. However, it is a common practice to record CCs in both Chinese and English. Furthermore, some hospitals record CCs only in Chinese. I systematically investigated the role and validity of Chinese CCs in the syndromic surveillance context. I then developed a system to classify Chinese CCs based on an automated mechanism to map Chinese CCs to English CCs.

The remainder of this chapter is organized as follows. Section 2.1 provides the background of the CC classification problem and cross-lingual information retrieval methods. The next section articulates the research opportunities and objectives. Section 2.3 presents the details of my technical approach for English CC classification. Section 2.4 presents the details of classifying Chinese CCs based on the English CC classification design. I report in Section 2.5 the experiments designed to evaluate my English CC classification approach. Section 2.6 highlights some issues about the reference standard dataset generation and the ontology-enhanced classification approach. Section 2.7 summarizes the experiments to evaluate the Chinese CC classification approach. The experimental results are presented in Section 2.8 Finally, Section 2.9 concludes the paper with a summary of my contributions.

2.1 Research Background

This section describes the input and output of a typical CC syndromic classification system. As part of the research background, I also briefly discuss CC coding schemes, survey various CC classification methods, and review cross-lingual information retrieval techniques. Some of these methods were used as benchmarks to compare with my own approach.

2.1.1 Free-Text Chief Complaints

CCs are the first records generated by triage personnel for ED patients. Examples of terms commonly seen in CCs are: nvd (nausea, vomiting, and diarrhea), fv (fever), fv w/c (fever with cough); sob (shortness of breath); so (ambiguous meaning); poss uti (possible urinary tract infection). The non-standard nature (misspellings, word variations, institution-specific use of expressions, etc.) of free-text CCs hinders their subsequent use in a syndromic surveillance system (Chapman 2006).

An obvious approach to deal with this significant problem is various spell checking algorithms that have been successfully applied in information retrieval research (Zobel & Dart 1996; Navarro 2001). However, previous research reported mixed results for spell checking algorithms such as those based on edit distance or phonetic similarity. For instance, spell checking algorithms provided limited value in CC classification systems based on Bayesian networks (Chapman *et al.* 2005a). On the other hand, combining edit distance and phonetic similarity was reported to increase the sensitivity of a chief complaint classification system (Shapiro 2004). Since acronyms and idiosyncratic

expressions are major sources of variations in free-text CCs, spell checking algorithms are only of limited value in CC processing.

2.1.2 Chief Complaint Coding Schemes and Medical Ontologies

A coding scheme is a set of standardized terminologies onto which chief complaints can be mapped. Coding schemes facilitate information retrieval, aggregation, and analysis. Two kinds of coding schemes are commonly used in public health surveillance research. The first kind is a general-purpose coding scheme, and encompasses examples such as ICD-9 CM, the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and the Unified Medical Language System (UMLS). General-purpose coding schemes usually include clinical terminology covering diseases, clinical findings and procedures. They are designed for consistently indexing, storing, and retrieving clinical data across medical practitioners and care sites.

Large collections of terminologies are usually included in these general-purpose coding schemes. For example, UMLS, developed and distributed by the U.S. National Library of Medicine, contains 2.5 million English terms. Similarly, SNOMED CT and ICD-9 CM contain 750,000 and 20,000 terms, respectively. Terms with the same meaning are usually organized by concepts. Hierarchies are constructed to reveal the relations among concepts. A major component of the UMLS is its Metathesaurus, which combines selected coding schemes including both ICD-9 CM and SNOMED CT. Figure 2.1 shows a subtree that exhibits the relations among “cramp stomach,” “upper abdominal pain,” and “epigastric pain” in the UMLS. The hierarchy in the UMLS is a valuable resource for medical information processing (Rosse *et al.* 1998; Achour *et al.*

2001). For instance, Leroy and Chen (Leroy & Chen 2001) demonstrated that the semantic relations among medical concepts can be used to help patients or medical experts find terms outside of their field of expertise.

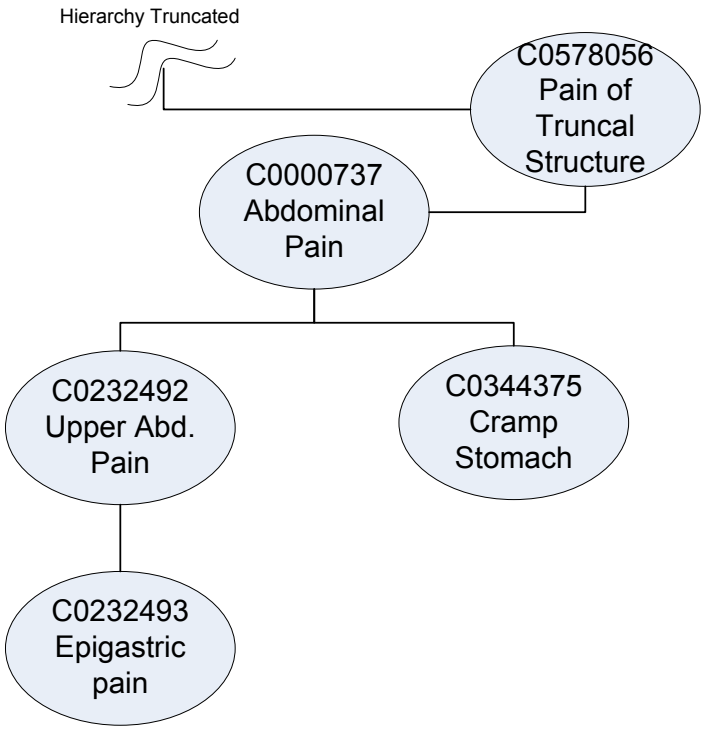


Figure 2.1 An Example of Semantic Hierarchy in the UMLS

The SPECIALIST lexicon, another component of the UMLS, is a general English lexicon that includes many biomedical terms. It can be used to normalize expressions such that the output text strings are in uninflected form without punctuation, genitive markers, and stop words. For example, "treating" and "treated" can be normalized to "treat" by the SPECIALIST Lexicon. The SPECIALIST Lexicon is a valuable tool for medical information processing. For example, Tolle and Chen (Tolle & Chen 2000)

showed that the performance of noun phrasing improved with the addition of the SPECIALIST Lexicon.

Another kind of coding scheme is more domain specific. Reason for Visit Classification (RVC) (Schneider *et al.* 1979) provides such an example in an emergency department care setting. The National Ambulatory Medical Care Survey uses RVC to classify chief complaints into one of the more than 770 standardized entries (McCaig & Nawar 2006). The Canadian Emergency Department Information System (CEDIS) workgroup proposed a coding scheme of 161 entries (Grafstein *et al.* 2003). Similar research (Aronsky *et al.* 2001; Day *et al.* 2004; Thompson *et al.* 2006) created coding schemes that range from 20 to 228 entries.

It should be noted that a small set of standardized codes with proper synonyms/keywords can capture a majority of chief complaint records. For example, it was reported that 67% of chief complaints in testing samples can be processed using 208 keywords which correspond to 20 chief complaint groups (Day *et al.* 2004). However, moving beyond this level of performance requires a disproportional amount of keywords or synonyms. For instance, only 85.7% of training records can be processed using 2557 keywords which correspond to 228 chief complaint groups (Thompson *et al.* 2006).

Choosing a proper coding scheme is crucial in building a flexible and effective chief complaint classification system. The coding scheme is the basic building block of CC classification systems. Coding schemes focusing on ED care settings are usually built by analyzing data collected from the field and thus can be applied relatively easily to process ED CCs. However, it is not clear how much external validity this kind of coding

scheme has. General-purpose coding schemes could provide a lot of useful information, as once the CCs are mapped to them, the existing semantic relations among the entries can be readily exploited to facilitate the syndromic category mapping process. In either case, there are difficulties in connecting the coding schemes and free-text CC records as none of these coding schemes can perfectly fit in CC records collected from different institutions. A possible solution involves using a combination of both types of coding schemes. For instance, a particular general-purpose coding scheme may be chosen and a customized synonym list may be built by analyzing the CCs collected from the EDs in order to bridge the gap between free-text CC and the coding schemes. The Emergency Medical Text Processor (EMT-P) system is an example of this(Travers & Haas 2003, 2004). It uses manually compiled synonym lists and the SPECIALIST lexicon tool to map expressions in CCs into a standardized form. The UMLS Metathesaurus then is used as a dictionary to map CCs to UMLS concepts.

2.1.3 Syndromic Categories

There are two issues related to using syndromic categories in CC classification. First, there is no consensus about a common set of syndromic categories a system should provide (Graham *et al.* 2002). Each syndromic surveillance system may have its own emphasis on the detection targets which determine the most appropriate syndrome groups and syndrome definitions. For instance, Electronic Surveillance System for Early Notification of Community-based Epidemics (ESSENCE) classifies CCs into eight syndromes: gastrointestinal, neurological, rash, respiratory, sepsis, unspecified, death, and others. RODS also has eight syndrome categories: gastrointestinal, constitutional,

respiratory, rash, hemorrhagic, botulinic, neurological, and others. But there is only partial overlap between the two systems' categories. The syndrome coding systems of EARS and the New York City Department of Health and Mental Hygiene use 41 and 9 syndromic categories, respectively.

The existence of variations in syndromic categories implies that, if a CC classification system is designed to be widely used by many institutions, the system must be flexible, in that adding new syndromic categories or recoding from one set of syndrome definitions to another should be relatively straightforward. Most existing systems, however, have limited flexibility to support multiple sets of syndromic categories.

The second issue is related to the reliability of syndrome definitions. Syndrome assignments in the reference standard dataset are assumed to be accurate when calculating the performance of classifiers. However, syndrome assignments created using unreliable syndrome definitions may introduce errors into the reference standard dataset. As a result, additional variations are introduced into the performance measures.

It has been shown that human experts can generate a reliable reference standard dataset using broadly-defined syndromic categories (Chapman *et al.* 2005b). In the study by Chapman, Dowling, and Wagner (Chapman *et al.* 2005b), medical records were reviewed by multiple physicians and the level of agreement between physicians as measured by Cohen's kappa (Cohen 1960) is high for most syndromes. It should be noted, however, that the level of agreement can be low for some syndromes. The reliability of syndrome definitions, therefore, should be carefully examined before an evaluation study.

2.1.4 Existing Automatic CC Classification Methods

There are two main approaches for automated CC syndrome classification: supervised learning and rule-based classification. A summary of selected syndromic surveillance systems that use CCs as one of their data sources and their underlying classification methods can be found in Table 2.1. The supervised learning methods require CC records to be labeled with syndromes before being used for model training. Naive Bayesian (Olszewski 2003; Espino *et al.* 2006) and Bayesian network (Chapman *et al.* 2005a) models are two examples of the supervised learning methods studied. Implementing the learning algorithms is straightforward; however, collecting training records is usually costly and time-consuming. For instance, 28,990 labeled records were used to train the RODS CoCo naive Bayesian classifier (Olszewski 2003; Tsui *et al.* 2003). Bayesian network classifiers require fewer training records and can achieve better performance than the naive Bayesian classifier. However, unlike most supervised learning methods, the training process was not fully automated. The system must interact with human experts to construct the semantic Bayesian network during the training process (Chapman *et al.* 2005a). Another major disadvantage of supervised learning methods is the lack of flexibility and generalizability. Recoding for different syndromic definitions or implementing the CC classification system in an environment which is different from the one where the original labeled training data were collected could be costly.

Table 2.1 Major CC Classification Methods

Methods	Systems	Related Research
Rule-Based Method		

Keyword match, synonym list, mapping rules	DOHMH Syndrome Coding System	Mikosz et. al., 2003
Same as above	EARS	Hutwagner et. al. 2003
Weighted keyword match (vector cosine method), mapping rule	ESSENCE	Sniegoski, 2004
Supervised Learning		
Naïve bayesian	RODS	Olszewski, 2003; Espino et. al., 2006
Bayesian network	N/A	Chapman et. al., 2005

Rule-based classification methods use a completely different approach and do not require labeled training data. Such methods typically have two stages. In the first stage, CC records are translated to an intermediate representation called "symptom groups" by either a symptom grouping table (SGT) lookup or keyword matching. For example, the ESSENCE system treats each CC as a document and each symptom group as a query. Symptom grouping, then, consists of running queries against CCs (Sniegoski 2004).

In the second stage, a set of rules is used to map the intermediate symptom groups to final syndromic categories. For instance, the standard EARS system uses 42 rules for such mappings.

A major advantage of rule-based classification methods is their simplicity. The syndrome classification rules and intermediate SGTs can be constructed using a top-down approach. The "white box" nature of these methods makes system maintenance and fine tuning easy for system designers and users. In addition, these methods are flexible: adding new syndromic categories or changing syndromic definitions can be achieved relatively easily by switching the inference rules.

A major problem with the rule-based classification methods is that they cannot handle symptoms that are not included in the SGTs. For example, a rule-based system may have a SGT containing the symptoms ”abdominal pain” and ”stomach ache” which belong to the symptom group “abd_pain.” This system will not be able to handle “epigastric pain” even though “epigastric pain” is closely related to “abdominal pain.” My research is designed to address this vocabulary problem using an ontology-enhanced approach.

2.1.5 Non-English Chief Complaint Classification Methods

Little research has focused on non-English CC classifications. One straightforward extension is adding non-English keywords into existing English CC classification systems. For instance, this approach has been applied to process Spanish CCs in EARS (Hutwagner *et al.* 2003). However, for other languages (such as Oriental languages), it would be difficult to incorporate them in an English-based system.

It is also possible to use International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9) codes instead of free-text CCs to classify ED records. ICD-9 codes are standardized, widely-used, and can be more accurate than CCs in terms of reflecting true patient illness. Wu et al. used ICD-9 codes attached to ED records to classify Chinese CCs into eight syndromic categories (Wu 2005; Wu *et al.* 2008). However, as ICD-9 codes are primarily used for billing purposes, they are not always informative for syndromic surveillance (Fisher *et al.* 1999; Day *et al.* 2004). As such, free-text CCs remain one of the most important data sources for syndromic surveillance (Li & Yang 2006).

2.1.6 Major Cross-Lingual Information Retrieval Approaches

Existing English chief complaint classification methods can be leveraged in a multi-lingual context by incorporating cross-lingual information retrieval (CLIR) methods. Cross-lingual information retrieval (CLIR) uses a query in one language to retrieve documents in different languages (Qin *et al.* 2006). Chinese CCs can be treated as documents in the target language, and an English CC classifier can be considered as a system performing query in English, the source language. There are two basic strategies in CLIR. The first strategy is translating documents in the target language to the source language (the language of original query) and performs information retrieval in the source language. The other strategy is translating queries in the source language to the target language and performs information retrieval in the target language (Chen 2002).

Three major translation approaches are commonly used in CLIR research: machine translation-based approach, corpus-based approach, and dictionary-based approach. The machine translation-based approach (Arnold *et al.* 1994; Sakai 2000) uses existing machine translation techniques to provide automatic text translation. Machine translation packages can be integrated into existing information systems. However, machine translation packages are often hard to customize. Moreover, in the context of syndromic surveillance, free text CCs consist of mostly short phrases or incomplete sentences, which lack the contextual and grammatical structural necessary for machine translation.

The corpus-based approach (Brown 1996; Oard 1996; Li & Yang 2006; Talvensaari *et al.* 2007) analyzes large document collections (parallel or comparable corpora) to

construct a statistical translation model. It has the potential to translate emerging terminologies. However, parallel corpuses are usually very hard to obtain. Existing parallel multilingual corpuses are typically small and cover only a small numbers of subjects.

Dictionary-based approach (Pirkola *et al.* 2001; Aljlayl *et al.* 2002; Daumke *et al.* 2007) uses bilingual dictionaries to translate text. Bilingual dictionary are relatively easy to obtain due to recent significant lexicon development efforts; thus this method can often be implemented more easily. However, multiple definitions of a word may cause translation ambiguity (i.e., word sense ambiguity). Moreover, commonly seen medical and symptom related terminologies are often absent in the multilingual dictionary collection.

2.1.7 Chinese Key Phrase Extraction and Text Segmentation

Chinese sentences are written without word/phrase boundaries explicitly delimited. This creates significant problems for Chinese-based information retrieval and text processing. For example, the precision of an information retrieval system can drop significantly if a query is not processed at the word level (Teahan *et al.* 2001). As such, how to recognize words in written Chinese has been an important research topic. Note that in Chinese, words and phrases are used interchangeably as they refer to a complete and standalone lexicon pattern that contains more than one Chinese character and has independent meanings.

Chinese key phrase extraction and Chinese text segmentation are two related major research questions. Chinese key phrase extraction studies the problem of extracting

important key phrases from a corpus. Chinese text segmentation, on the other hand, focuses on the problem of separating words in a given sentence. These two problems are not completely independent. A text segmentation system can benefit from a good key phrase list and a key phrase extraction system can benefit from good text segmentation results. The major difference between these two problems is that Chinese key phrase extraction usually does not assume the existence of a training dataset. However, it is common to formulate Chinese text segmentation as a supervised learning problem.

2.1.7.1 Chinese Key Phrase Extraction

Similar to the task of constructing multi-word phrases in English, one way to construct the key phrase list is by running through a part-of-speech (POS) tagger and combining characters based on the tagging results. However, because of the lack of word boundaries in Chinese, a Chinese POS tagger needs to either have word segmented before POS tagging or perform word segmentation and POS tagging simultaneously (Ng & Low 2004). Note that under the context of syndromic surveillance, there are few training corpora available to implement this approach.

Another popular Chinese key phrase extraction method relies on statistical evidence that reflects collocations or co-occurrences among Chinese characters. Pointwise Mutual information (Manning & Schütze 1999), a statistical metric used to measure the strength of association between two adjacent characters, is often the basis for such research. The method was used to extract words with two characters (Sproat & Shin 1990) or more (Yang *et al.* 2000). A recent research used this approach to extract significant topics from a text collection of Chinese book and article titles (Wang 2006).

An alternative approach that uses extended mutual information to measure the strength of co-occurrence among lexicon pattern of two or more characters was proposed by Chien (1999) (Chien 1999). All lexicon patterns were checked with respect to the extended mutual information measure and key phrases were extracted without length limitation. This approach often requires more computing resources as a larger pattern candidate space needs to be explored.

2.1.7.2 Chinese Text Segmentation

Existing Chinese text segmentation methods can be broadly classified into two categories: dictionary-based and statistical-based methods. I briefly summarize these methods below.

Dictionary-based approach is the simplest approach to segment Chinese text (Wu & Tseng 1993; Cheng *et al.* 1999). When a large-enough collection of phrases is available, this method can provide reasonable performance using straightforward implementation such as maximum forward match or maximum backward match. However, dictionary-based method has an obvious problem of identifying new words (Goh *et al.* 2005). Thus if there is no suitable dictionary for text collections from a particular field, this method could perform poorly.

Similar to the problem of Chinese key phrase extraction, the collocation information such as n-gram can also be used to perform text segmentation. The compression-based method uses an adaptive language model originally designed for text compression and formulate the text segmentation problem as a hidden Markov model to insert spaces between characters (Teahan *et al.* 2001). Specifically, the Prediction by

Partial Matching (PPM) compression scheme (Cleary & Witten 1984) was studied. This approach learns n-gram from a segmented training dataset. Given a sentence in testing dataset, the segmentation with highest compression is chosen. Experimental results showed good performance when training and testing dataset were from the same corpus. However, performance was significantly worse when training and testing dataset were from different corpus.

One way to alleviate the problem of mismatched training and testing dataset is to make use of a large-enough corpus. The web mining-based segmentation algorithm make use of the n-gram collected by submitting corresponding queries to search engines such as Google and Yahoo (Wang & Yang 2007). After adjusting for the length of words, the combination of words with highest adjusted frequency are chosen as the segmentation result. Experiments showed that this segmentation algorithm outperformed existing state-of-art segmentation methods and were robust to text collections from different geographical areas (Wang & Yang 2007).

2.2 Research Opportunities and Objectives

My review of existing CC classification methods reveals several research opportunities.

1. The UMLS contains meaningful relations between symptoms that could be potentially leveraged in a CC classification system. Knowledge captured in existing CC classification methods, either learned through training samples or acquired directly from human experts, could be enhanced by these

relations. However, most existing research ignores such semantic information.

2. The lack of one common standard for syndromic categories calls for an architecture which can support flexible syndromic categories.
3. The existing term processing and syndromic classification research has resulted in concrete findings and system components that should be leveraged and reused when developing new approaches.
4. Little research has investigate the role of non-English CCs in syndromic surveillance systems.
5. Current syndromic surveillance research provides limited support for non-English CC processing.

Based on these observations, my research is aimed at developing:

1. A novel free-text CC classification approach that can leverage a medical ontology to improve classification effectiveness.
2. A Chinese CC classification system which leverages existing English-based CC classification research.

2.3 An Ontology-Enhanced Chief Complaint Classification Approach

This section reports a new ontology-enhanced CC classification approach that meets the research objective discussed in the previous section. I first discuss its basic design and then discuss its major technical components.

2.3.1 A Rule-based Design

My approach largely follows a rule-based design as opposed to a supervised learning method. As argued before, a rule-based method requires less training data and is flexible in incorporating new syndromic categories. My approach will address its key weakness, i.e. the difficulty associated with handling symptoms not previously encountered, by making use of semantic information contained in the UMLS ontology.

As depicted in Figure 2.2, my syndromic classification approach can be divided into three major stages: CC standardization, symptom grouping, and syndrome classification. Central to my approach is the weighted semantic similarly score (WSSS)-based grouping component that automatically expands the coverage of the symptom grouping table by exploiting the semantic relations between symptoms. In the remainder of this section, I first introduce the symptom grouping table and then discuss the three major stages of my approach in turn.

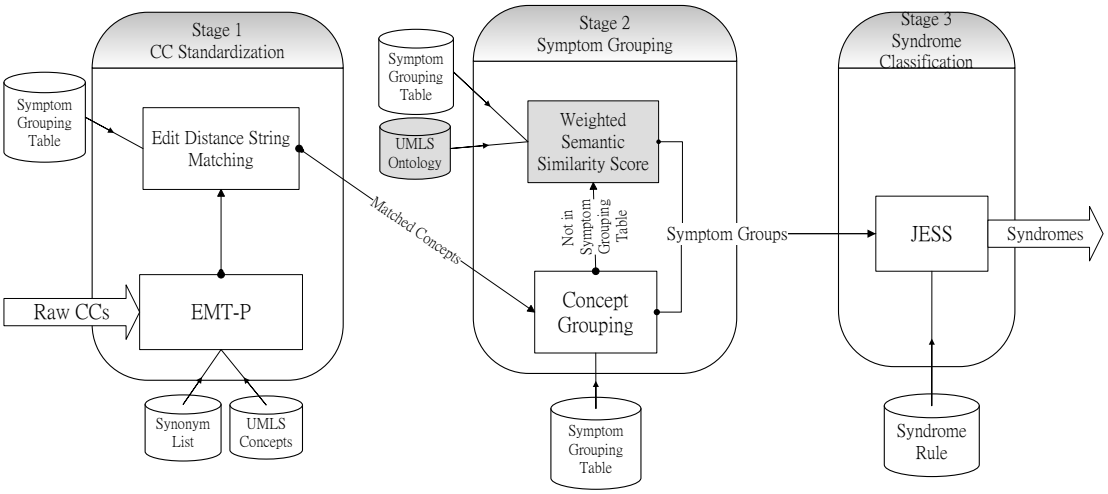


Figure 2.2 System Design for an Ontology-Enhanced Chief Complaint Classification Approach

2.3.2 The Symptom Grouping Table (SGT)

A symptom grouping table records the mapping relations from symptoms to symptom groups. Symptoms to be classified in the same syndromic category are grouped together in a symptom group. For instance, nausea, vomiting, and sickness all point to the same gastrointestinal syndrome and thus are grouped together. Note that the granularity of symptom grouping depends on the final syndrome definitions. For example, if I am interested in respiratory syndrome only, the symptoms apnea, difficulty breathing, gasping, and hemoptysis can all be grouped together. However, if I also consider the hemorrhagic syndrome in addition to the respiratory syndrome, then the original symptom group must be broken down into two: one containing apnea, difficulty breathing, and gasping; and the other containing hemoptysis. Syndrome mapping rules can then be constructed so that the first group is mapped into the respiratory syndrome and the latter into both the hemorrhagic and respiratory syndrome.

Ideally, each and every symptom can only be mapped into one symptom group. Example entries in a SGT can be found in Table 2.2. The symptoms in the SGT are stored in their standardized form following the underlying coding scheme, in my case, UMLS. For example, the second row in Table 2.2 indicates that the symptom "bleeding gums," with a unique id C0017565 in UMLS, belongs to the group "bleeding."

Table 2.2 Selected Records in a Symptom Grouping Table

Symptom Group	Concept Unique ID	Symptom Name
bleeding	C0019080	bleeding
	C0017565	bleeding gums
nvd*	C0151594	bloody diarrhea
	C0011991	diarrhea
	C0027497	nausea

	C0027498	nausea vomit
	C0221423	sickness
	C0277525	stomach flu

*nvd stands for "nausea, vomiting, and diarrhea."

The SGT used in this study contains 61 groups and 392 symptoms (more discussion about the construction of the SGT can be found in "System Benchmarks.") In my study, it is implemented as a relational table with three fields: concept name, concept unique ID, and symptom group. The symptom grouping process identifies symptom groups that match the concept unique IDs from a CC.

2.3.3 Stage One: Chief Complaint Standardization

In Stage One, the acronyms, abbreviations, and truncations in CC records are expanded and normalized using synonym lists, the SPECIALIST lexicon tool, and edit distance string matching. Then the standardized symptoms extracted from CCs are mapped to UMLS concepts. The EMT-P module is capable of expanding acronyms and truncation using synonym lists and the SPECILAIST lexicon tool and is reused as a plug-in module in my system.

The EMT-P, nevertheless, has two shortcomings in this application. First, it cannot handle a simple typographical error such as "sore thorat" or word concatenation such as "sorethroat." Second, the EMT-P does not consider symptoms in the current SGT as more relevant to the CC classification task and may decide to cut CCs into one or more concepts in its own way.

The edit distance string matching module is designed to address the first shortcoming. For each string that cannot be processed by the EMT-P, the similarity

between terms in the SGT and the unrecognized string is calculated. The unrecognized string is deemed as similar to a term in the SGT if each word in the term can find a counterpart in the unrecognized string within a "small" distance and these words appear in the same order as they do in the SGT. Edit distance is considered small if: (a) the distance is zero; or (b) the word (in the SGT term) has more than five characters and the edit distance is one; or (c) two words have the same length, contain more than five characters, and have an edit distance of two. For example, for the unknown string "sore thorat," the term "sore throat" in the SGT is similar to it because "sore" and "throat" can find their counterpart "sore" and "thorat" in the unknown string in the same order as they appear in the SGT (that is, "thorat sore" would not be considered similar to "sore throat").

EMT-P fails to process CC records with concatenated words, those formed by a group of words without any dividing signposts such as spaces or hyphens. For example, EMT-P maps "sore throat" to a UMLS concept successfully but fails to map "sorethroat" to the same concept. I use a simple approach to correct this problem for matching purposes: for each term in the SGT, I produce a concatenated word by linking all words of the term. This concatenated word is then used to match unknown strings.

Finally, as the terms in current SGT were created by domain experts familiar with target CCs of the classification system, the terms in the SGT should have higher priority over those found in the UMLS. (EMT-P does not treat the terms in the SGT differently from the UMLS concepts when determining how an expression should be divided into concepts). In my approach, I added another step searching the EMT-P *output* for terms in the SGT. The benefit of this step is that once part of a CC can be mapped to a term in the

SGT, the grouping and subsequent syndrome classification can be done routinely. As such, the chance of finding any potential match to the current SGT is maximized. For instance, "arm injury" was mapped to one single symptom by EMT-P as EMT-P, as it prefers longer symptoms. In my approach, however, since "injury" appears in the SGT, the same record is standardized into both "arm injury" and "injury."

Given that multiple methods are used to extract concepts in CCs, it is possible that some concepts come from overlapped terms. If multiple matches are found and some terms are contained in other terms, these embedded shorter terms are dropped. For instance, if both "blood" and "blood pressure" are matched to "increased blood pressure sweat," then "blood" is dropped because it is part of the term "blood pressure."

To further illustrate the procedures used in Stage One, I discuss several additional examples. The first example is the free-text CC "DIARRHEA ABD CRAMPING." The EMT-P component is first invoked and identifies this CC as two concepts in the UMLS: abdominal cramp (C0000729) and diarrhea (C0011991). The text strings in the parentheses are the unique concept IDs in the UMLS. The entire free-text CC is successfully mapped to the UMLS concept. The edit distance search and word concatenation search are thus skipped. Terms in the SGT are used to search in "abdominal cramp" and "diarrhea" but no new concepts are found. The final output of step one is two concepts: abdominal cramp and diarrhea.

The second example is the free-text CC "STIFF NECK,UPPER SPINE PAIN." EMT-P identified "stiff neck" (C0151315) as one UMLS concept but marked "upper spine pain" as unidentified. The edit distance and word concatenation searches are

invoked for the latter concept. It turns out that no additional concept is identified. Finally, terms in the SGT are used to search new concepts in "stiff neck" and "upper spine pain" but no new concept is found. The final output has only one concept: stiff neck.

The third example is "SORE THORAT." EMT-P failed to identify any UMLS concept from the input string. The edit distance identifies the string as similar to the concept "sore throat" (C0242429) in the SGT according to the rules described above. The word concatenation search does not find additional concepts. Finally, terms in the SGT are used to search the unmatched EMT-P output "sore thorat" but without a match. The final output in this case is "sore throat."

2.3.4 Stage Two: Symptom Grouping

In the second stage, each symptom extracted from the previous stage is mapped into an appropriate symptom group. As discussed before, symptom groups are intermediate representations that can enable system modularity, extensibility, and flexibility.

My system uses the SGT to match symptoms to groups. If a corresponding group is located in the SGT, the grouping process terminates. However, it is likely that some symptoms do not directly appear in the SGT. In a traditional rule-based system design, the system simply ignores the unmatched symptoms.

The main technical innovation of my research is the development of an ontology-enhanced approach to process these unmatched symptoms. The basic intuition behind my approach is as follows. Unmatched symptoms may be semantically related to some symptoms in the SGT. If the semantic relations between these symptoms can be exploited,

the system will be able to process the unmatched symptoms using the original SGT. In other words, the ontology can help expand the coverage of SGTs automatically.

At the center of my approach is the Weighted Semantic Similarity Score (WSSS). This score is based on the semantic distance between two concepts, defined as the path distance between them in the UMLS hierarchy. The specific definition of path distance and related computation are described below.

The path distance calculation involves four steps. Since each symptom may have more than one synonym, the first step is to identify all synonyms of the two symptoms between which semantic similarity is to be determined. Next, all ancestor nodes of identified synonyms are located. As the UMLS stores the concept hierarchy in a relational table, locating these ancestor nodes takes only one query which is efficient to execute. Third, the distance between a pair of terms is calculated by comparing their ancestor nodes. After calculating the distances of all possible pairs of synonyms from the two symptoms, the shortest distance is returned as the distance between the two symptoms.

Figure 2.3 provides pesudocode for calculating the path distance between two concepts in the UMLS. For example, “swelling” and “abd. swelling” has a parent-and-child relation in the UMLS; the distance between these two symptoms is one. “Dysphagia” and “bloating” have a common parent “symptoms involving digestive system;” the distance between them is two.

Semantic_Distance(C1, C2)

```
SET shortest_distance = a_large_number
SET syn_set1 = all synonyms of C1
```

```
SET syn_set2 = all synonyms of C2
FOR each syn1 in syn_set1
   FOR each syn2 in syn_set2
         CALCULATE all ancestors of syn1 RETURNING ancestor1
         CALCULATE all ancestors of syn2 RETURNING ancestor2
          CALCULATE the distance of syn1 and syn2 with ancestor1 and ancestor2
              RETURNING distance
          IF distance < shortest_distance THEN
              SET shortest_distance = distance
       ENDIF
   ENDFOR
ENDFOR
RETURN shortest_distance
```

Figure 2.3 Pesudocode for Calculating the Semantic Distance Between two Concepts C1 and C2 in the UMLS

Given a symptom C1 that is not in the SGT, the semantic distances from C1 to all symptoms in the SGT can be calculated. All distances are sorted in ascending order and distances larger than a threshold Z are discarded. The retained distances are then grouped together based on symptom groups. The WSSS measuring the "fitness" between C1 and all candidate symptom groups is then calculated by adding the reciprocals of semantic distance.

Formally, I define d_{ij} as the distance between C1 and symptom j in group i, and S_{Zi} as the set of distances that satisfies threshold Z and belong to group i. Then the WSSS for group i of order Z is defined as

$$w_{Zi} = \sum_{d_{ij} \in S_{Zi}} \frac{1}{d_{ij}}$$

The threshold Z starts at one. The symptom group with the highest score is chosen as the resulting group for the unmatched symptom if only one group meets the threshold requirement. If two or more groups have the same WSSS, Z is increased by one. This process repeats itself until Z is too large to reveal a meaningful relation between the unrecognized symptom and groups in the SGT. Preliminary experiments showed that two symptoms with distances larger than four are usually related in a very weak manner. Thus the above process is repeated until Z is larger than four. It is possible that the WSSS will not result in a match if none of the groups can meet the threshold distance requirement.

I now use several real examples to illustrate the WSSS calculation process and how it is used to determine symptom group assignment. For example, the symptom "gall bladder pain" does not have a direct match in the SGT. By calculating semantic distances, I find that "abdominal pain" is the closest symptom in the SGT. The symptoms "cramp stomach" and "bladder pain" are the next closest. Table 2.3 lists the top ten closest symptoms to "gall bladder pain." Clearly, the symptom group "gi" (gastrointestinal) has the highest score with threshold Z equaling one. As a result, "gall bladder pain" is assigned to group "gi." Another example can be found in Table 2.4. The unknown concept "groin swelling" cannot be matched with any symptom group in the SGT with the distance threshold set to 1; therefore the threshold is extended to two. Seven concepts satisfy the new threshold. The group "swelling" has three concepts with a distance equaling two. Thus the WSSS is 1/2+1/2+1/2=1.5. Group "gi" has two concepts with a distance of two, and has a WSSS of 1/2+1/2=1. The third group, "limbs_pain," has one concept with a distance of two and a WSSS value of 0.5. The last group, "tachycardia,"

also has a WSSS value of 0.5. As the group "swelling" has the highest WSSS, the unknown concept "groin swelling" is assigned to the group "swelling."

Table 2.3 Top 10 SGT Symptoms Closest to "gall bladder pain"

Distance	Symptom	Group
1	abdominal pain	gi
2	bladder pain	gi
2	cramp stomach	gi
2	left sided abdominal pain	gi
2	lower abdominal pain	gi
2	rectal pain	gi
2	right sided abdominal pain	gi
2	stomach ache	gi
2	upper abdominal pain	gi
2	groin pain	limbs_pain
	(list truncated)	

Table 2.4 Top eight SGT symptoms closest to "groin swelling"

Distance	Symptom	Group
2	arm swell	swell
2	groin lump	swell
2	leg swelling	swell
2	abd. pain	gi
2	abd. swelling	gi
2	leg pain	limbs_pain
2	bradycardia	tachycardia
3	abscess	cellulitis leg
	(list truncated)	

One might argue that the closest symptom based on the semantic distance calculation could well serve as the best match for the unmatched symptom under investigation. Based on my computational experience, however, this simplistic design can lead to rather arbitrary results, as typically the unmatched symptom could have multiple closely-related SGT symptoms which suggest different groups. My WSSS-based

approach is designed to mitigate such ambiguous situations. Conceptually, it can be viewed as a weighted voting scheme to determine the best group.

Though the method of grouping un-encountered symptoms using the WSSS is relatively new, it is conceptually similar to the widely used nearest-neighborhood methods (see, for example, (Witten & Frank 2005)). Given a point X to be classified, the nearest-neighborhood method searches k points which are closest to the point X from the training dataset and assigns the classification result by the majority class of the k points. In my approach, I map the expressions in CCs to a concept space constructed by the UMLS and define the distance metric based on the UMLS. The SGT then serves as the training dataset in the nearest-neighborhood method to determine the classification result for symptoms that the system has not encountered before.

2.3.5 Stage Three: Syndrome Classification

In the last stage, the system decides to which syndrome the CC belongs. This is done by mapping the symptom groups obtained from Stage Two to predefined syndromic categories using mapping rules. In my implementation a rule inference engine, JESS (http://herzberg.ca.sandia.gov/jess/), is employed.

As an example, Rule 1 in Figure 2.4 dictates that CCs belonging to symptom groups gastrointestinal (gi); gastrointestinal bleeding (gi_bleed); nausea, vomiting, and diarrhea (nvd); constipation; or jaundice, but not caused by motor vehicle accident (mva), are assigned to gastrointestinal syndrome (GI_CAT). Similarly, Rule 2 in Figure 2.4 dictates that CCs that belong to rash or hemorrhagic rash are classified into rash syndrome.

Rule 1: (defrule s_gi "Gastrointestinal" (and (or (gi) (gi_bleed) (nvd) (constipation) (jaundice)) (not (mva))) => (record GI_CAT))

Rule 2: (defrule s_rash "Rash" (or (rash) (hemorrhagic_rash)) => (record RASH_CAT))

Rule 3: (defrule s_botu "Botulism-like" (or (blurred_vision) (dysphagia) (paralysis)) => (record BOTU_CAT))

Rule 4: (defrule s_hemo "Hemorrhagic" (and (or (bleeding) (hematemesis) (hemoptysis) (hemorrhagic_rash) (gi_bleed)) (and (not (chronic)))) => (record HEMO_CAT))

Figure 2.4 Selected Syndrome Mapping Rules

Note that all symptom groups from the same CC are combined to determine the syndrome classification results. For example, the raw CC "SOB AND NAUSEA" is standardized into "shortness of breath" and "nausea." In Stage Two, "shortness of breath" is grouped into "respiratory" and "nausea" is grouped into "nvd." In the final stage the two groups, "respiratory" and "nvd," are considered simultaneously and the CC is classified into two syndromic categories: respiratory syndrome and gastrointestinal syndrome.

In my implementation, the rule set is stored in a plain text file. This rule set consists of two parts. Rules in the first part encode the main logic behind the mapping from symptom groups to syndromes. Examples from this part can be found in Figure 2.4. There are 17 such rules in total. Rules in the second part dictate the priority of syndrome assignments. For example, one rule in this part states that the "other" syndrome will be dropped if it is not the only syndrome identified. There are 8 rules in the second part.

The rule set can be changed and updated easily and independently when new syndromic categories are needed. For instance, if a new syndrome, "febrile gastrointestinal," needs to be added to existing syndromic categories, the user only needs to add one more rule to the rule file that combines the symptom groups involving the gastrointestinal syndrome and fever symptoms using an "and" operation. This design provides flexibility and extensibility for the system to meet the changing needs of syndromic surveillance.

In a full-fledged system, this simple approach of capturing rules in a plain text file could lead to scalability and maintenance issues as the rule set grows. A more formal, structured approach in dealing with such rules might be needed. However, since the rule set is built upon the symptom groups instead of individual symptoms, the number of rules is not likely to be large. This is another advantage of my symptom group-based approach.

2.4 Classifying Chinese Chief Complaints

It is possible to develop a Chinese CC classification approach from scratch. However, there are significant language processing issues and few comprehensive medical ontologies in languages other than English. Existing Chinese medical terminologies are only related to translations of medicine and disease names, none are designed for syndromic surveillance. Since there are many effective CC classification methods already developed for English CCs, I chose to leverage these methods. The language difference can be bridged by cross-lingual text processing techniques.

There are two major challenges hindering my effort to process Chinese CCs. The first is the lack of a Chinese key phrases list containing common medical phrases

appearing in Chinese CCs. The second is the lack of a Chinese-English translation mechanism for important medical phrases relevant to syndromic surveillance. Although some related general discussions and technical solutions have been discussed in the field of cross-lingual information retrieval, to the best of my knowledge, there are no readily-available solutions that can be directly applied to process Chinese CCs.

Motivated to address these two challenges, I have designed a two-step method that consists of (a) statistical Chinese key phrase extraction based on the concept of mutual information and (b) symptom phrase translation. After translating Chinese CCs to English, the BioPortal CC classification system, which was developed in my prior research for English CCs (Lu *et al.* 2008a), is then used to process translated CCs. In this section, I first discuss the techniques used to process Chinese CCs in order to translate them into English. The following subsections then present the detailed procedure that is used to classify Chinese CCs into syndrome categories.

2.4.1 Chinese Chief Complaint Preprocessing

The goal of Chinese CC preprocessing is to translate Chinese CCs gathered from the field in such a way that the existing, well-tested English CC classifier can be reused. The set of all Chinese CCs (939,024 records in total) was processed using a statistical pattern extraction method based on the concept of mutual information to construct a key phrase list for syndromic surveillance. This key phrase list was then used to perform Chinese word segmentation and Chinese-English translation.

Note that translating Chinese CCs to English is different from typical translation tasks. The Chinese expressions in CCs are in most cases short phrases as opposed to

complete sentences. Moreover, not every word or phrase is informative for syndromic surveillance purposes. As a result, the goal of Chinese CC preprocessing is not to provide verbatim translation of Chinese expressions in CC records. Instead, only information that is useful and relevant to syndromic surveillance should be extracted from the original Chinese CCs.

2.4.1.1 Statistical Chinese Key Phrase Extraction Using Extended Mutual Information

Following the method proposed by Chien (1999) (Chien 1999), I define the Extended Mutual Information (EMI) of a phrase (Chien 1999; Ong & Chen 1999; Lu *et al.* 2008a) as:

$$EMI = \frac{f(c)}{f(a) + f(b) - f(c)} \qquad (1)$$

where f(c) represents the frequency of the pattern c; $c=c_1, c_2,\ldots, c_n$ is the pattern of interest (e.g., 上吐下瀉; vomiting and diarrhea); $a=c_1, c_2,\ldots, c_{n-1}$ and $b=c_2, c_3, \ldots, c_n$ are longest left and right subpatterns of c, i.e. a="上吐下" (a partial word without meaning) and b="吐下瀉" (a partial word without meaning). Based on this measure, EMI will be substantially higher than other random patterns if c is by itself a phrase and its subpatterns a and b appear in the text only because of c. For instance, c="上吐下瀉" may appear in the text 9 times. Its subpatterns a="上吐下" and b="吐下瀉" appear in the text only because they are the subpatterns of c. In this case, I have EMI=9/(9+9-9)=1. Intuitively, stronger co-occurrence indicates a higher chance of being a meaningful phrase. A EMI score of 1 indicates that c should be considered as a complete phrase.

Searching the whole candidate pattern space requires considerable computing power. Fortunately, each Chinese CC record can be treated as a separate document and punctuation marks such as comma and period can be used to further divide the text string. The maximum length of lexicon patterns is thus greatly reduced. As suggested by previous research (Chien 1999; Ong & Chen 1999), I construct a PAT tree (Gonnet *et al.* 1992) from divided text strings and stored the frequency of the semi-infinite strings in corresponding nodes. The PAT tree then could be used to provide an efficient structure of computation. Given a lexicon pattern, the frequency of its subpatterns could be easily retrieved by walking up and down the tree. The EMI measure was calculated solely from the information stored in the PAT tree. Lexicon patterns with EMI higher than a pre-specified threshold were considered as the candidate terms in the Chinese key phrase list.

2.4.1.2 Key Phrase List Construction and Translation

To construct a high quality key phrase list, I used a low threshold to filter the output from the EMI method and manually reviewed 2,533 candidate phrases. All candidate phrases contained at least two Chinese characters. These candidates were sorted in ascending order by phrase length (number of Chinese characters). One of the authors went through the candidates and removed them if (a) the candidate was not a meaningful phrase or (b) the candidate did not contain information relevant to syndromic surveillance or (c) the meaning of the candidate can be caught by the combination of shorter phrases that had been included. Table 2.5 provides a few examples of candidate phrases that were reviewed during the process. It took us about 4 hours to extract four hundred and fifteen symptom-related key phrases from the 2,533 candidate phrases.

Table 2.5 Intermediate Results of Chinese Key Phrase List Construction

Candidate	Included (Yes/No)	Comment
自殺 (suicide)	Yes	
臉部 (face)	Yes	
吸不 (partial phrase, no meaning)	No	Not a phrase
鄰居 (neighbor)	No	Unimportant information
治療 (treatment)	Yes	
被割傷 (trauma)	Yes	
被打現 (partial phrase, no meaning)	No	Not a phrase
被狗咬 (bitten by a dog)	Yes	
被汽車 (partial word, no meaning)	No	Not a phrase

I expanded the key phrase list using a general purpose Chinese-English dictionary of about 220,000 entries (http://www.mandarintools.com/cedict.html). For each candidate Chinese phrase from the Chinese-English dictionary, I included it in my key phrase list if it appeared in my Chinese CC dataset for more than 5 times. Fifty-five additional key phrases were identified. The final symptom key phrase list contains 470 Chinese key phrases.

Three physicians in Taiwan were recruited to translate the extracted Chinese key phrases into English. I provided the physicians with a file listing the Chinese key phrases together with example CCs which contained these phrases. I then reviewed the translations from these physicians to make sure that translations are consistent.

2.4.2 A System Design for Chinese Chief Complaint Processing

Figure 2.5 depicts the design of my Chinese CC classification system. My Chinese CC classification system follows six major stages. Stages 0.1 to 0.3 separate Chinese and English text strings in CCs, perform word segmentation for Chinese text strings, and map symptom-related phrases to English. At the end of Stage 0.3, CC records are in English. In the following three stages (Stages 1-3), the BioPortal CC classifier is invoked (Lu *et al.* 2008a). The terms are mapped into concepts in the UMLS ontology in Stage 1. Related concepts are then gathered and put into symptom groups in Stage 2. In Stage 3, a set of rules are used to map symptom groups to the final syndrome categories. Below I discuss each of these steps in detail.

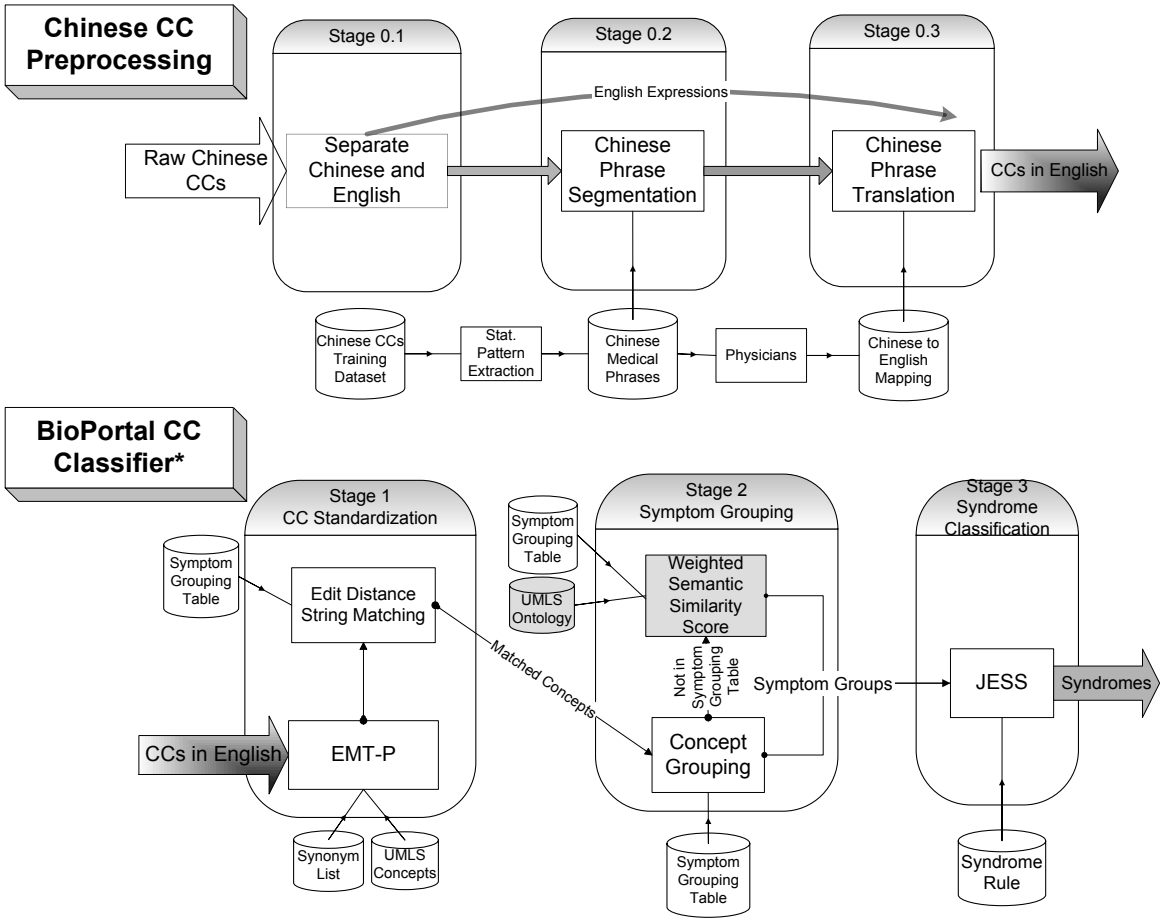


Figure 2.5 Chinese chief complaint classification process

2.4.2.1 Stage 0.1: Separating Chinese and English Expressions

Stage 0.1 separates Chinese from English text strings. Since the BioPortal CC classifier can process English CCs, any existing English text strings are kept. The positions of the Chinese and English strings are also marked for future reference. For example, the chief complaint record “Dyspnea,SOB 早上開始坐骨神經痛 解尿困難” is first divided into two parts: “Dyspnea,SOB,” which will skip subsequent Chinese CC

preprocessing steps; and “早上開始坐骨神經痛 解尿困難,” which will be sent to Stage 0.2 for word segmentation.

2.4.2.2 Stage 0.2: Chinese Expression Segmentation

In this stage, Chinese expressions are segmented using the Chinese symptom key phrase list discussed in the previous section. The longest possible phrases in the phrase list are used for segmentation. For example, the Chinese CC “車輛導致下巴裂傷 (verbatim translation: jaw laceration caused by a car)” is a combination of the following phrases: “車輛 (c car),” “導致 (cause),” “下巴 (jaw),” and “裂傷 (laceration).” They are concatenated without any punctuation marks and thus require segmentation before further processing. Using the key phrase list constructed earlier, the original Chinese CC is segmented as: “[車輛] – [導致] – [下巴] – [裂傷] (verbatim translation: [a car] – [cause] – [jaw] – [laceration]).” Each text string in square brackets is a phrase segmented from the original text. The verbatim translations in square brackets are the meanings of Chinese phrases segmented from the original text string. Although the combination of individual phrase translation does not constitute a complete sentence with a correct grammatical structure, they do carry valuable information about the syndrome associated with the CC.

Note that since my key phrase list is relative small, many proper nouns are not included. As a result, segmentation results may not be accurate if proper nouns are involved. For example, the Chinese CC “左手被吉娃娃咬傷 (verbatim translation: left hand bitten by a Chihuahua)” is segmented as: “[左] – [手] – [被] – [吉] – [娃] – [娃] –

[咬傷] (verbatim translation: [left] – [hand] – [by] – [auspicious] – [baby] – [baby] – [bite]).” In this case, the phrase “吉娃娃 (Chihuahua)” is not correctly segmented. The phrase is segmented as three individual Chinese characters because Chihuahua is not included in the key phrase list. This error, nevertheless, does not prevent us from recognizing the syndrome-related information from the inaccurately segmented result.

2.4.2.3 Stage 0.3: Chinese Phrase Translation

The segmented phrases generated from the previous step are used in Chinese-English symptom mapping. Phrases not recognized are omitted. For example, the segmented Chinese expression “[與] – [人] – [打架], [用] – [鍋] – [鏟] – [打到頭] – [部], [流鼻血]” is mapped to the following English expressions: “[N/A] – [N/A] – [fighting], [N/A] – [N/A] – [N/A] – [head injury] – [N/A], [epistaxis].” “N/A” indicates the term is unavailable in the mapping table. The final translated result thus is “fighting, head injury, epistaxis.”

Note that the translation in this stage only depends on the 470 key phrases extracted using Extended Mutual Information. Compared to the number of commonly-used Chinese characters (about 6,000; see for example, (Wang & Yang 2007)), this key phrase list is fairly small. As shown in Section 6, this key-phrase based translation approach led to good overall syndromic classification performance. This positive finding has practical implications in syndromic surveillance. First, it indicates that triage nurses usually use a relatively small, well defined set of phases to describe symptoms. Second, it is practical

and efficient to develop a standardized vocabulary which can further facilitate the processing, aggregation, and analysis of Chinese CCs.

2.4.2.4 Stages 1-3: English-based Chief Complaint Classification

After substituting the Chinese text strings with the translated English strings in the CCs, I proceed to use the BioPortal CC classifier. There are three major Stages in the BioPortal CC classifier: CC standardization, symptom grouping, and syndrome classification. Detailed steps are discussed in the previous sections.

2.5 Experiment 1: Classifying Chief Complaints Recorded in English

In this section, I report on an experimental study conducted to evaluate the ontology-enhanced CC classification approach with respect to a human generated reference standard dataset (Hripcsak & Wilcox 2002). The evaluation focuses on the usefulness of the WSSS component and the performance difference between the ontology-enhanced system and the supervised learning system. The CC classification subsystems from two syndromic surveillance systems were chosen as benchmarks: Early Aberration Reporting System (EARS) and Real-time Outbreak Detection System (RODS). EARS is a syndromic surveillance system developed by the CDC after the terrorist attacks of September 11, 2001. Major data sources monitored by EARS include free-text CCs, 911 calls, school and business absenteeism, and OTC drug sales. RODS, developed by the University of Pittsburgh, was designed for detection and assessment of disease outbreaks. Similarly, data sources such as free-text CCs and OTC drug sales are monitored. RODS is used by more than 12 states in the US.

I first discuss the performance measures employed in my study and the statistical procedure used to test the performance differences between my approach and the two benchmarks. I then discuss the research test bed, reference standard dataset, and syndromic definitions used in this study. I then report my experimental findings.

2.5.1 Performance Criteria

Sensitivity, specificity, and positive predictive value (PPV) have all been used extensively in previous research (Ivanov *et al.* 2002; Olszewski 2003; Chapman *et al.* 2005a). In addition I also consider the F measure and F2 measure (van Rijsbergen 1979; Pakhomov *et al.* 2006). The F and F2 measures, commonly used in the information retrieval literature, combine PPV and sensitivity to provide a single integrated measure to evaluate the overall performance of a given approach. The goal of a syndromic surveillance system is to detect disease outbreaks while minimizing the false alarm rate. This corresponds to high level of PPV and specificity. Excessive false alarms could happen if these two measures do not reach the desired levels. Sensitivity summarizes the portion of positive cases that can be captured by the classification system and can be linked to higher detection power. However, higher sensitivity could lead to lowered PPV and specificity and increases the false alarm rate. The F measure family is one way of characterizing the trade-off between detection power and the false alarm rate. In this family of measure, the F measure is the harmonic mean of sensitivity and PPV and thus can be interpreted as a measure that considers sensitivity and PPV equally important. The F2 measure assigns sensitivity twice as much weight as PPV and can be interpreted as a

measure that is biased toward sensitivity. It should be noted that specificity is not included in the F measure and F2 measure calculations.

McNemar's test (McNeman 1947; Chapman & Haug 1999; Agresti 2002) is useful in determining whether two systems have the same level of accuracy. When considering only the positive cases of the same syndrome in the reference standard dataset, McNemar's test can also provide a statistical test for sensitivity comparison. A similar technique applies for specificity.

Unfortunately, this technique cannot be applied to PPV. Unlike sensitivity, the denominators of PPV, which are the positively classified CCs from the two systems, only partially overlap in most cases. Moreover, paired or independent comparisons are not applicable due to violated assumptions. The F measure and F2 measure, which encompass PPV, also lack proper statistical tests.

To overcome these difficulties, I apply the bootstrapping method for statistical inference on PPV, sensitivity, specificity, F measure, and F2 measure. Bootstrapping (Efron 1979) is a general-purpose re-sampling technique for assessing statistical accuracy. The basic idea behind bootstrapping is to use the empirical distribution function obtained through the sample on hand to generate bootstrapping samples that in turn provide the sampling distribution of the statistics of interest.

Before proceeding with the detailed bootstrapping procedure used in this study, a clarification is in order. In the statistical learning literature, bootstrapping usually involves both system training and testing (see, for example, (Hastie *et al.* 2001)). For instance, the bootstrapping method developed by Efron (Efron 1983) estimates the

system error rate by the weighted average of the training error using bootstrap sample and the testing error using instances not in the training sample. However, in a setting where training a benchmark system is not practical or where the training process is not fully automated, bootstrapping for both training and testing would be inappropriate.

In this study, my system and the benchmark systems are evaluated using the same reference standard dataset. The point estimator of various performance criteria can be calculated. A confidence interval of performance difference is required for statistical inference. The general bootstrapping procedure (Efron 1979; Efron & Tibshirani 1986; Hinckley 1988) can produce the confidence intervals for all performance criteria of interest. More specifically, I am interested in testing the null hypothesis $H_0 : g_{BioPortal} - g_{Benchmark} \leq 0$ against the alternative hypothesis $H_1 : g_{BioPortal} - g_{Benchmark} > 0$, where $g_{BioPortal}$ is the performance of my system (referred to as BioPortal) under a particular criterion, and $g_{Benchmark}$ is the performance of one of the benchmark systems under the same criterion.

This problem is equivalent to testing whether the performance difference $d = g_{BioPortal} - g_{Benchmark}$ is smaller than or equal to zero. Since this is a one-sided test, the hypothesis is rejected at $1-\alpha$ confidence level if the $1-2\alpha$ level confidence interval is all positive. I define a bootstrap sample as a random sample with replacements from the original reference standard dataset of the same sample size as the original reference standard dataset. Then for bootstrap sample i, i=1, 2, …, B, I calculate the performance difference d_i. The $1-2\alpha$ level confidence interval of d is then the interval covering the α

percentile and $1-\alpha$ percentile of $\{d_i\}$, i=1,2,…,B. A step by step procedure can be found in Figure 2.6.

1. Set the counter i = 1 and the total number of bootstrap samples B=2500*.
2. From the testing dataset of size n, draw a random sample with replacement of size n.
3. Calculate the performance of my system $g_{BioPortal,i}$ and the benchmark system $g_{Benchmark,i}$ using the sample from the previous step.
4. Calculate the difference $d_i = g_{BioPortal,i} - g_{Benchmark,i}$.
5. Increase i by one.
6. If $i \le B$, repeat Steps 2-5.
7. The $1-2\alpha$ level confidence interval is the interval covering the α percentile and $1-\alpha$ percentile of $\{d_i\}$.
8. The null hypothesis $H_0 : g_{BioPortal} - g_{Benchmark} \le 0$ is rejected at confidence level $1-\alpha$ if the $1-2\alpha$ level confidence interval is all positive.

* See the discussion in the Performance Criteria section for the guideline of choosing B.

Figure 2.6 Bootstrapping Procedure for Performance Comparison

An important control parameter of my bootstrapping procedure is the total number of bootstrap samples, B. When building confidence intervals using the bootstrapping method, B is typically set at larger than one thousand (Efron & Tibshirani 1986). Through computational experiments, I also observed that the evaluation results become stable when B is set to a number larger than one thousand. In my study, I have chosen a conservative setting for B, 2,500. Since the bootstrapping method as discussed above is generally applicable to evaluating system performance along all performance criteria used in my study, the subsequent analyses are mainly based on bootstrapping.

The bootstrapping method discussed above has not been used widely in previous public health surveillance studies. As a comparison framework, it can be applied to any similar research involving performance comparison between two systems. To ensure correct inference, the only assumption needed is the independence of the records in the reference standard dataset. It is my intended contribution to advocate this type of method for more rigorous studies of performance comparison between different surveillance methods.

2.5.2 Research Testbed

The CC records used in this study were provided by the Phoenix Metropolitan Hospital through the Arizona Department of Health Services. The training dataset contains 2,256 CC records covering an interval of 11 days. The string length of records varies from one to thirty-two characters. The testing dataset is a random sample of one thousand records from July 2005 to November 2005, excluding the time interval when training data was collected. As the focus of this study is on improving the effectiveness of a CC classifier using a medical ontology, I am more interested in how the performance differs on distinct records as opposed to providing an unbiased estimation of classification performance. Therefore, duplicated chief complaint strings were removed before performing the random sampling.

The training dataset was used during the system development process and also for system tuning. The testing dataset was used to generate the reference standard dataset for system performance evaluation.

2.5.3 Syndromic Definitions and Reference Standard Dataset

Eleven syndromes were chosen by the Arizona Department of Health Service for evaluation: botulism, constitutional, gastrointestinal, hemorrhagic, neurological, rash, respiratory, upper respiratory, lower respiratory, fever, and other (the syndromic definitions used in this study can be found in the appendix). "Other" is a miscellaneous category for CCs that do not fit into any of the other syndromes. One chief complaint could be assigned to more than one syndrome. If upper respiratory or lower respiratory is assigned, it automatically implies respiratory syndrome (but the reverse is not true).

To ensure that the syndrome definitions used were comparable to the benchmarks, text descriptions for each syndromes were compiled based on those used by the RODS Laboratory (Chapman *et al.* 2005a). A mapping table of syndrome assigned by EARS, another benchmark system, to syndromes used by my system was constructed based on the descriptions. All syndromes except constitutional were successfully linked to the EARS syndromes. Table 2.6 lists the mapping from the EARS and RODS syndromes to those used in this study. More detailed discussion about the benchmark systems can be found in the next section.

Table 2.6 Syndrome Mapping Between the BioPortal System and the Benchmark Systems

Bioportal	EARS	RODS
Botulism-like	s_botulism	Botulism-like
Constitutional	N/A	Constitutional
Gastrointestinal	s_gastrointestinal, s_gicat	Gastrointestinal
Hemorrhagic	s_hemorrhagic	Hemorrhagic
Neurological	s_neurons, s_neurological	Neurological
Rash	s_rashcat	Rash

Respiratory	Upper Respiratory, Lower Respiratory	Respiratory
Upper Respiratory	s_upperresp, s_sb_upper_respiratory	N/A
Lower Respiratory	s_lowerresp, s_sb_lower_respiratory	N/A
Fever	s_fever, s_febrile	N/A

To the best of my knowledge, there are no publicly available CC datasets labeled with syndrome definitions. Thus, the reference standard dataset for the system evaluation had to be constructed for this study. Three experts, including two physicians and one nurse in Phoenix, Arizona, were given the syndrome definitions and the testing set of one thousand chief complaints. They were asked to assign CCs to syndromes independently. After collecting the assignments from the experts, majority voting was used to determine the final syndrome assignment of each CC.

Out of these one thousand records, majority voting could not determine the syndrome assignment for eighteen CCs that all three experts labeled differently. In these cases, a fourth expert, an emergency department physician, helped determine the final assignment. Another 85 CCs that had syndrome assignments mixed with "other" were also reviewed by this fourth expert. The final reference standard dataset contains 148 CCs that have been assigned to more than one syndrome. On average, one CC is assigned to 1.18 syndromes.

The prevalence of the 11 syndromes can be found in Figure 2.7. Similar to previous research (Olszewski 2003; Chapman *et al.* 2005a), the "other" category has the highest

prevalence. Syndromes such as respiratory, gastrointestinal, and neurological have a prevalence of about 10%. Botulism has the lowest prevalence, at 0.6%.

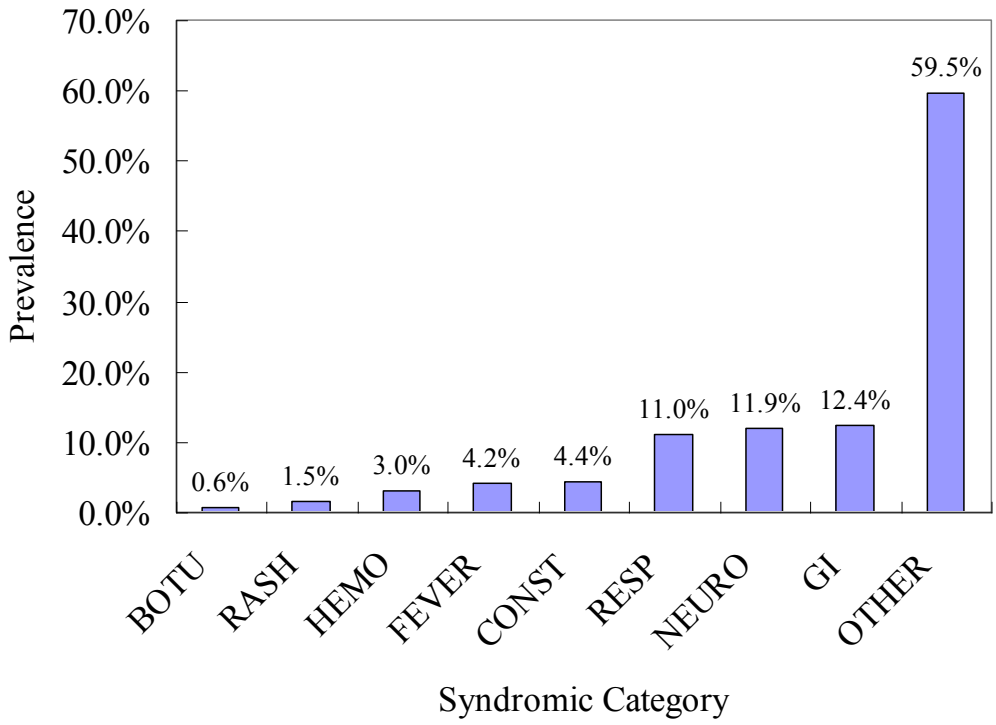


Figure 2.7 Syndrome Prevalence

The kappa statistic is calculated using the assignment from the first three experts. The overall agreement is good (kappa=0.71). Table 2.7 summarizes the kappa statistic of each syndromic category. Some syndromic categories such as botulism, constitutional, and lower respiratory syndromes have low agreement, while the fever and neurological syndromes have moderate agreement (according to the standard proposed by (Fleiss 1981, p.218)). The low kappa value for botulism syndrome may be due to its low prevalence. It is very difficult to have a large kappa value for a rare syndrome because a few disagreements can strongly influence the kappa value. In order to have reliable estimation

of system performance, only syndromes with excellent agreement (kappa higher than 0.75) were used in my evaluation study.

Table 2.7 Kappa Statistics of Each Syndromic Category

Syndrome	Kappa
Botulism-like	0.22
Constitutional	0.24
Lower Respiratory	0.38
Fever	0.46
Neurological	0.64
Other	0.74
Upper Respiratory	0.77
Respiratory	0.80
Hemorrhagic	0.81
Rash	0.82
Gastrointestinal	0.85
Overall	0.71

2.5.4 System Benchmarks

The CC classification subsystems of RODS and EARS serve as the benchmarks to compare against my ontology-enhanced approach. RODS uses its own CoCo naïve Bayesian classifier and is treated as a black-box CC classification method for the evaluation (Olszewski 2003). It is referred to as the CoCo naïve Bayesian classifier (CoCoNBC) in subsequent discussion. The CC classification subsystem of EARS, on the other hand, is a rule-based classification system that shares some common architectural design elements with ours. It is referred to as EARS CC classification subsystem (ECCCS). ECCCS uses a symptom table to map raw chief complaints into groups and a

set of rules to assign syndromes. In effect, the symptom table from ECCCS was used to construct the initial SGT of my system as follows. For symptoms listed in the ECCCS symptom table, EMT-P was used to standardize them into UMLS concepts. I took care to merge any redundant groups. For example, symptoms in "Poisoning" are very similar to those in "CO Poisoning." In another example, there is no clear distinction between "Death" and "Unexplained Death" and they were merged together. The final SGT in my approach contains 61 groups and 392 symptoms. To ensure a fair comparison, the rule set from ECCCS was amended based on my syndrome definitions to construct the initial rule set for my system.

The setting using the SGT and rule set adapted from ECCCS is referred to as ECCCS in BioPortal. The BioPortal project is an infectious disease informatics project with funding support from the National Science Foundation and other federal state agencies. The reported research is part of this project (Chang *et al.* 2005; Hu *et al.* 2005; Yan *et al.* 2006). ECCCS in BioPortal and ECCCS share the common symptom grouping table and compatible syndrome rules. As such, I can examine the effect of the WSSS component in isolation and fairness. Comparing ECCCS in BioPortal to CoCoNBC can help us evaluate whether an ontology-enhanced approach can achieve performance comparable to that of the naïve Bayesian method.

2.5.5 Performance Comparisons

Table 2.8 summarizes the results of the comparison between ECCCS in BioPortal and ECCCS by syndromic categories. The second to the fifth columns of Table 2.8 list the true positive (TP) cases, false negative (FN) cases, true negative (TN) cases, and false

positive (FP) cases in each syndromic category. The sixth through the eighth columns list the PPV, sensitivity, and specificity measures. Comparing the TP and FN cases across syndromes, I find that the WSSS component raises the number of TP cases and reduces the number of FN cases. For example, the WSSS component increased the TP cases by 15 for the respiratory syndrome. At the same time, the FN cases decreased by the same amount. Raising the TP cases comes at the cost of an increased number of the FP cases. The increase of FP cases is different across syndromes. For instance, the respiratory syndrome only has 3 additional FP cases while the number of TP cases was increased by 15. On the other hand, for gastrointestinal syndrome, there are 20 additional FP cases while the number of TP cases is increased by 14.

Table 2.8 Performance Comparison Between ECCCS in BioPortal and ECCCS

ECCCS in BioPortal									
Syndrome	TP	FN	TN	FP	PPV	Sensitivity	Specificity	F	F2
GI	104	20	850	26	0.8000	**0.8387*** **	0.9703	**0.8189**	**0.8254***
HEMO	19	11	967	3	0.8636	**0.6333*** **	0.9969	**0.7308****	**0.6951*** **
RASH	10	5	976	9	0.5263	**0.6667****	0.9909	**0.5882**	**0.6122**
RESP	90	20	879	11	0.8911	**0.8182*** **	0.9876	**0.8531*** **	**0.8411*** **
URESP	36	7	935	22	**0.6207**	**0.8372*** **	0.977	**0.7129****	**0.7500*** **
ECCCS									
Syndrome	TP	FN	TN	FP	PPV	Sensitivity	Specificity	F	F2
GI	90	34	870	6	**0.9375*** **	0.7258	**0.9932*** **	0.8182	0.7849
HEMO	10	20	970	0	**1.0000***	0.3333	**1.0000***	0.5000	0.4286
RASH	7	8	982	3	**0.7000***	0.4667	**0.9970*** **	0.5600	0.5250
RESP	75	35	882	8	**0.9036**	0.6818	**0.9910**	0.7772	0.7426
URESP	27	16	938	19	0.5870	0.6279	**0.9801**	0.6067	0.6136

From the discussion above, it is not surprising to observe that the WSSS component has opposite effects on PPV and sensitivity. The PPV of ECCCS in BioPortal is lower in most syndromic categories except in the upper respiratory syndrome. The difference is significant in the gastrointestinal syndrome (p-value < 1%), the hemorrhagic syndrome (p-value < 10%), and the rash syndrome (p-value < 10%). On the other hand, the sensitivity of ECCCS in BioPortal is significantly higher in all syndromes under consideration. The p-values are less than 1% in the gastrointestinal syndrome, the hemorrhagic syndrome, the respiratory syndrome, and the upper respiratory syndrome; and are less than 5% in the rash syndrome. ECCCS in BioPortal also has lower specificity. But since specificities in both systems are very high (all larger than 97.03%), the difference is not substantial. When considering PPV and sensitivity together, ECCCS in BioPortal has higher F measures and F2 measures in all syndromes. The differences are significant in the hemorrhagic syndrome, the respiratory syndrome, and the upper respiratory syndrome for both the F measure and F2 measure (p-value < 5%), and significant in the gastrointestinal syndrome for the F2 measure (p-value < 10%). Comparing the significance level of the F measure and the F2 measure, I find that the F2 measure is significant in gastrointestinal syndrome (p-value < 10%) while the F measure is not significant. Similarly, the F2 measure is significant in the upper respiratory syndrome at the 1% level while the F measure is only significant at the 5% level. The differences in statistical significance levels reflect the fact that the F2 measure emphasizes sensitivity over PPV. To summarize, ECCCS in BioPortal achieves higher sensitivity but lower PPV. But in terms of the F measure and F2 measure, ECCCS in

BioPortal outperforms ECCCS. Since the major difference between ECCCS and ECCCS in BioPortal is whether the WSSS grouping is used, I conclude that adding the WSSS component to a rule-based system increases its sensitivity and F and F2 measures at the expense of lowered PPV.

Table 2.9 summarizes the comparison between ECCCS in BioPortal and CoCoNBC. ECCCS in BioPortal has more TP and FP cases in most syndromes. The only exception is the hemorrhagic syndrome. CoCoNBC has one more TP case than that of ECCCS in BioPortal. ECCCS in BioPortal has lower PPV in three out of four syndromes. The difference, however, is only significant for the gastrointestinal syndrome. ECCCS in BioPortal has significantly higher sensitivity in most syndromes including the gastrointestinal syndrome, the rash syndrome, and the respiratory syndrome. CoCoNBC delivers higher sensitivity in the hemorrhagic syndrome. Given that the numbers of TP and FP cases in the hemorrhagic syndrome have only small differences between these two classifiers, it is not surprising that the statistical tests find no significant difference. I also observe that both systems have fairly high specificity. ECCCS in BioPortal has higher F measure and F2 measure in the gastrointestinal syndrome, the rash syndrome, and the respiratory syndrome but not in the hemorrhagic syndrome. The differences are significant in the gastrointestinal syndrome (p-value < 5%) and the respiratory syndrome (p-value < 1%). Note that the F2 measure is significant at the 1% level in the gastrointestinal syndrome but the F measure is only significant at the 5% level. This difference, again, reflects the fact that the F2 measure puts more weight on sensitivity.

Although the differences are significant only for half of the syndromes, these syndromes cover more than 80% of the CCs under consideration.

Table 2.9 Performance Comparison Between ECCCS in BioPortal and CoCoNBC

ECCCS in BioPortal									
Syndrome	TP	FN	TN	FP	PPV	Sensitivity	Specificity	F	F2
GI	104	20	850	26	0.8000	**0.8387*** **	0.9703	**0.8189** **	**0.8254*** **
HEMO	19	11	967	3	0.8636	0.6333	0.9969	0.7308	0.6951
RASH	10	5	976	9	0.5263	**0.6667***	0.9909	**0.5882**	**0.6122**
RESP	90	20	879	11	**0.8911**	**0.8182*** **	0.9876	**0.8531*** **	**0.8411*** **
URESP	36	7	935	22	0.6207	0.8372	0.977	0.7129	0.7500
CoCoNBC									
Syndrome	TP	FN	TN	FP	PPV	Sensitivity	Specificity	F	F2
GI	80	44	867	9	**0.8989** **	0.6452	**0.9897*** **	0.7512	0.7122
HEMO	20	10	968	2	**0.9091**	**0.6667**	**0.9979**	0.7692	**0.7317**
RASH	7	8	980	5	**0.5833**	0.4667	**0.9949***	0.5185	0.5000
RESP	65	45	881	9	0.8784	0.5909	**0.9899**	0.7065	0.6633
URESP	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

2.6 Experiment 1: Discussion

This section discusses issues related to the reference standard dataset generation and the WSSS component. I then summarize the significance and limitations of my study and point out future work.

2.6.1 Discrepant Kappas among Syndromic Categories

The kappa statistics of the eleven syndromic categories vary substantially. As briefly discussed before, rare syndromes such as botulism may have difficulty achieving

a high kappa. However, the rash syndrome has moderate prevalence (1.5%) but a high kappa (0.82). Compared to the study of Chapman et. al. (Chapman *et al.* 2005a), I find similarities and differences. The respiratory, hemorrhagic, and gastrointestinal syndromes share similar high levels of agreement in both studies. The rash syndrome has an excellent level of agreement in my results (kappa=0.82), but the lowest level of agreement in (Chapman *et al.* 2005a) (kappa=0.23). The botulism and constitutional syndromes have low levels of agreement in my research but have moderate levels of agreement in (Chapman *et al.* 2005a). Fever has a moderate level of agreement in my results but a high level of agreement in (Chapman *et al.* 2005a). My kappa statistic for the neurological syndrome is about 15% lower than that reported in (Chapman *et al.* 2005a).

This comparison shows that it is not uncommon to have different agreement levels across different syndrome categories. One possible explanation is that the experts' different work experiences or specialty concentrations may lead to different interpretations of the chief complaints (and other information in ED reports in the study of (Chapman *et al.* 2005a)) and thus result in different syndrome assignments.

Although a detailed analysis about why syndromic categories such as botulism and constitutional have low levels of agreement and whether the syndromic definitions can generate a reliable reference standard dataset is beyond the scope of this article, a few examples may shed some light on this important topic. The chief complaint "NAUSEA WEAKNESS NOT EATING" was assigned to the constitutional and gastrointestinal syndromes by expert one, gastrointestinal by expert two, and constitutional by expert

three. “HA VOMITING” was assigned to the neurological and gastrointestinal syndromes by expert one, neurological by expert two, and constitutional by expert three. The above examples indicate some possible reasons for low levels of agreement. First, the definition of some syndromes may not be clear. In some cases, even if the definitions are clear, the experts may have a difficult time fully understanding and consistently following these definitions. More research is needed to understand this issue further.

2.6.2 The Effect of the WSSS Component

As summarized in the previous section, the WSSS component is able to increase the number of TP and FP cases simultaneously. The resulting sensitivity is higher while PPV and specificity decrease. The increase in sensitivity means that the classification system can single out the desired signal better. At the same time, additional noise is introduced into the classification results because of higher PPV. Practically, for syndromes with low prevalence such as the rash syndrome, improving sensitivity should be the first priority to avoid delay in outbreak detection. Alternatively, if the syndrome has moderate or high prevalence, then the trade-off between sensitivity and PPV becomes less clear-cut. In cases where the classification system has moderate sensitivity but very high PPV, increased sensitivity and decreased PPV may benefit the subsequent statistical detection task by increasing the signal level higher than the noise it introduced. However, it is possible that this kind of adjustment make the detection task more difficult. If the relative importance between the detection ability of a surveillance system and the cost of having a false alarm can be determined, a weighting scheme which reflects the relative importance can be used to customize the measure from the F measure family. This measure then can

be used to determine whether the trade-off between sensitivity and PPV is beneficial for the surveillance system. It should be noted that the decision to adapt the WSSS method is determined on a syndrome-by-syndrome basis. That is, the method may be applied to only some syndromes in a CC classifier while other syndromes are classified using the original method.

As noted in the "Research Test Bed," the reference standard dataset contains only distinct CC strings. Evaluating the BioPortal CC classification system with this reference standard dataset can tell us how the WSSS component extends the knowledge of a CC classifier. It is also interesting to know the performance impact of the WSSS component on the reference standard dataset that contains duplicated records (i.e. a random sample without duplicated records removed). I recalculated the performance of the ECCCS in BioPortal and ECCCS using the new reference standard dataset that contains duplicated records. The basic pattern of increased sensitivity and decreased PPV are the same. However, ECCCS in BioPortal has significantly lower F measure and F2 measure in the gastrointestinal syndrome. Looking into individual records, I find that the WSSS misclassified high frequency CCS "flank pain," "left flank pain," "right flank pain," and "kidney pain" into the gastrointestinal syndrome. These false positives substantially reduced PPV of ECCCS in BioPortal. These cases, as a result, should receive higher priority in error analysis.

The gastrointestinal syndrome does not have good performance using the original reference standard dataset either. Among all syndromes, the gastrointestinal syndrome had the largest increment in FP cases using ECCCS as a benchmark. The number of FP

cases (26) was also substantially higher than that of CoCoNBC (9). I thus select the gastrointestinal syndrome as the focus of error analysis.

The WSSS component utilizes semantic information from a medical ontology for symptoms grouping purposes. While the WSSS grouping results coincided with the assignments of human experts most of the time, it is possible that the WSSS assigns the wrong group to an unseen symptom. For example, “left flank pain” was assigned to group “gi” and subsequently classified to the gastrointestinal syndrome because it is very close to the symptom “abdominal pain” in the UMLS (distance = 2). But “left flank pain” was not considered part of the gastrointestinal syndrome in the reference standard dataset. “Vaginal pain” was also classified into group ”gi” because its closest neighbor in the UMLS is “abdominal pain” (distance = 1). The right mapping for “Vaginal pain” was actually “other” in the reference standard dataset.

The above examples indicate that, in certain cases, the UMLS ontology is not suitable for the purpose of syndromic surveillance. Detailed error analysis should be able to provide a more complete picture about the potential factors that affect the performance of the ontology method and shed light on the direction of future performance improvement.

2.7 Experiment 2: Classifying Chief Complaint Recorded in Chinese

This section reports an evaluation study. To the best of my knowledge, there is no publicly available CC classification system for Asian languages. Therefore, there is no existing system that can be directly used as a benchmark in my evaluation study. Instead of conducting a system-level evaluation study, I compare the core component of my

Chinese CC preprocessing approach against other Chinese-English mapping methods (i.e. bilingual dictionary translation and machine translation methods) (Sakai 2000; Daumke *et al.* 2007) in terms of the final syndromic classification performance. In addition to assessing the efficacy of my approach, this comparative study can provide insights about the unique characteristics of the multilingual CC classification problem and provide directions for future improvements.

In this section, I first summarize the syndrome definitions and the gold standard dataset used in this study. The translation methods used as benchmarks are described next. Finally, the empirical findings are presented with examples that illustrate the difference between these translation methods.

2.7.1 Syndromic Definitions and the Gold Standard

I used eight syndrome categories chosen by five local collaborating physicians: constitutional, gastrointestinal, rash, respiratory, upper respiratory, lower respiratory, fever, and other. "Other" was a miscellaneous category for CCs that did not fit into any of the rest syndromes. One chief complaint could be assigned to more than one syndrome. For example, if the upper respiratory or lower respiratory was assigned, the respiratory syndrome automatically applied as well. These categories were similar to those reported in previous studies (Gesteland *et al.* 2003; Chapman *et al.* 2005a; Lu *et al.* 2008a).

To the best of my knowledge, there is no publicly available dataset with labeled Chinese CCs. Therefore gold standard for system evaluation had to be constructed for this study. The gold standard dataset was a random sample of 1884 CC records from the MK Hospital in Taiwan. Three experts including two physicians and one nurse in Taiwan

were given the syndrome definitions and the set of 1884 testing CCs. They were asked to assign CCs to syndromes independently. After collecting the assignments from the experts, a majority rule was used to determine the final syndrome assignments of each CC. On average, one CC was assigned to 1.44 syndromes. According to the final gold standard, gastrointestinal syndrome had the highest prevalence of 31.28%. About 20% of CCs contained fever syndrome. The prevalence of constitutional and respiratory syndromes is about 15%.

Kappa statistic was calculated to determine the assignment agreement among the three experts. The overall agreement was good (kappa=0.83). All syndromic categories had kappa higher than 0.85 except for the constitutional syndrome, which had kappa of 0.56. Only syndromes with excellent agreement (kappa higher than 0.75) were used in the evaluation study (Fleiss 1981, p. 218).

2.7.2 Performance Benchmarks: Bilingual Dictionary and Google Translation

Several alternative approaches could provide Chinese-English translations. Translations using a bilingual dictionary provided a simple and reasonable performance baseline. For terms with more than one translation in the bilingual dictionary, the first translation was used. A popular and publicly available Chinese-English dictionary was used to provide translations in this setting (http://www.mandarintools.com/cedict.html). There are about 220,000 entries in the collection. This setting is referred to as Bilingual Dictionary translation.

Machine translation is often more sophisticated. I adopted the machine translation method as another benchmark for my evaluation. I used Google Language Tools to

provide the translations (http://www.google.com/language_tools?hl=EN). According to a recent machine translation evaluation study conducted by the National Institute of Standards and Technology (NIST) in 2006, the machine translation system developed by Google was one of the best systems among 46 participants for Chinese-English translation (NIST 2006). As such, the web-accessible Google machine translation system provided an excellent professional benchmark. After collecting translations from Google Language Tools, the same BioPortal CC classifier was used to provide syndrome classification results. This setting is referred to as Google Translation in the subsequent section.

For my approach, I used an extended mutual information measure to construct a key phrase list for Chinese-English mapping. My approach is referred to as Mutual Information-based Mapping (MIM).

2.7.3 Performance Comparison

In my study, system performance was measured using widely-used metrics, including sensitivity (recall), specificity, positive predictive value (PPV or precision), F measure, and F2 measure (van Rijsbergen 1979; Ivanov *et al.* 2002; Olszewski 2003; Chapman *et al.* 2005a; Pakhomov *et al.* 2006). The performance of all methods under consideration was measured using the same gold standard. McNemar's test (McNemar 1947; Agresti 2002) could be applied for accuracy and sensitivity comparison. However, McNemar's test could not be used to compare PPV, F measure, and F2 measure. Standard paired and independent comparisons were not applicable in this situation as their assumptions did not hold. I thus applied a bootstrapping method to calculate the

confidence intervals of the performance differences for all measures so that the experimental results could be interpreted in terms of formal hypothesis testing (Lu *et al.* 2008a).

2.8 Experiment 2: Results and Discussion

Performance comparison results between MIM and Google Translation can be found in Table 2.10. The second column of Table IV lists the positive cases in each syndromic category. The third through the 7th columns list the performance in terms of PPV, sensitivity, specificity, F measure, and F2 measure. In most syndromic categories, the MIM method generates PPV, sensitivity, specificity, F measure and F2 measure higher than 0.9. Rash syndrome has the worst performance with F measure of 0.82. The fever syndrome has the best performance with F measure of 0.97.

Table 2.10 Performance comparison for MIM and Google Translation

Mutual Information-based Mapping (MIM)						
Syndrome	TP+FN	PPV	Sensitivity	Specificity	F	F2
GI	592	0.97***	0.97***	0.98***	0.97***	0.97***
RASH	45	0.87**	0.77	0.99**	0.82***	0.80***
RESP	331	0.89***	0.96***	0.97**	0.93***	0.94***
URESP	132	0.86***	0.91**	0.98***	0.88***	0.89***
LRESP	272	0.93	0.98***	0.98	0.95***	0.96***
FEVER	413	0.99**	0.96	0.99**	0.97	0.97
Google Translation						
Syndrome	TP+FN	PPV	Sensitivity	Specificity	F	F2

GI	592	0.91	0.90	0.96	0.91	0.91
RASH	45	0.76	0.73	0.99	0.75	0.74
RESP	331	0.84	0.83	0.96	0.83	0.83
URESP	132	0.70	0.83	0.97	0.76	0.78
LRESP	272	0.96**	0.80	0.99***	0.87	0.84
FEVER	413	0.98	0.96	0.99	0.97	0.97

* p-value < 0.1 ** p-value < 0.05 *** p-value < 0.01
Statistical test is based on 3,000 bootstrappings.

Compared to Google Translation, the MIM method has significantly higher PPV, sensitivity, and specificity in most syndromic categories. Given the significant differences in PPV and sensitivity, it is not surprising to find that the MIM method has significantly higher F measure and F2 measure than those of the Google Translation, as these two measures are the functions of PPV and sensitivity. It is interesting to note that MIM has significantly higher F measure and F2 measure in all syndromic categories except the fever syndrome. MIM and Google Translation have almost the same performance for the fever syndrome. A review of translation results in this syndromic category shows that one keyword ("fever") can cover more than 90% of all true positive cases. As a result, providing good translation for this category is relatively easier than that of other categories. Overall the experimental results indicate that the MIM method provides better syndrome classification performance comparing to processing Chinese CCs using the Google machine translation system.

Table 2.11 summarizes performance comparison between MIM and Bilingual Dictionary. In general, MIM performs much better than Bilingual Dictionary in terms of PPV, sensitivity specificity, F and F2 measures. Most of the performance difference is significant at a 99% confidence level. Note that Bilingual Dictionary has zero sensitivity in fever syndrome. The reason behind the low performance is because fever was translated to "have a high temperature" by the definition of the bilingual dictionary. The BioPortal CC classifier failed to recognize the phrase as related to fever syndrome. A review of individual translated CC records indicated that there was a gap between the terms covered by the bilingual dictionary and the terms that were commonly seen in my Chinese CC dataset.

Table 2.11 Performance comparison for MIM and Bilingual Dictionary

Mutual Information-based Mapping (MIM)						
Syndrome	TP+FN	PPV	Sensitivity	Specificity	F	F2
GI	592	0.97***	0.97***	0.98***	0.97***	0.97***
RASH	45	0.87***	0.77	0.99***	0.82***	0.80***
RESP	331	0.89	0.96***	0.97	0.93***	0.94***
URESP	132	0.86***	0.91***	0.98*	0.88***	0.89***
LRESP	272	0.93	0.98***	0.98	0.95**	0.96***
FEVER	413	0.99	0.96***	0.99	0.97	0.97
Bilingual Dictionary						
Syndrome	TP+FN	PPV	Sensitivity	Specificity	F	F2
GI	592	0.36	0.36	0.70	0.36	0.36
RASH	45	0.54	0.77	0.98	0.64	0.68

RESP	331	0.88	0.79	0.97	0.83	0.82
URESP	132	0.43	0.16	0.98	0.24	0.20
LRESP	272	0.95**	0.90	0.99**	0.93	0.92
FEVER	413	NA	0.00	1.00**	NA	NA

2.8.1 Examples

A few examples may help us understand the performance difference among these translation methods. Table 2.12 provides an example of the input and output of the syndromic classification system. The raw Chinese CC "全身酸痛 喉嚨痛今早始 (verbatim translation: whole body soreness and sore throat. began this morning)" has two important keywords: soreness and sore throat. The MIM method caught both keywords. Google translated the CC as "general soreness sore throat this morning before," which was accurate. The translation result from Bilingual Dictionary, nevertheless, failed to provide any meaningful information for syndromic surveillance. As mentioned above, the major reason behind the poor translation results of Bilingual Dictionary was the lack of medically related terminologies in the dictionary collection.

Table 2.12 Example 1: Raw Chinese CC, Translations and Classification Results

Raw Chinese CC: 全身酸痛 喉嚨痛今早始 (verbatim translation: whole body soreness and sore throat. began this morning).			
Translation Method	Translation Outcome	Syndrome Outcome	Gold Standard
MIM	soreness, sore throat	UPPER RESP, RESP	CONST, RESP, UPPER RESP
Bilingual	ache, today early begin	UNKNOWN	

Dictionary			
Google Translation	general soreness sore throat this morning before.	UPPER RESP, RESP	

Another example can be found in Table 2.13. The raw Chinese CC "吐 晚上開始 (verbatim translation: vomiting. began this evening)" contains symptoms related to gastrointestinal syndrome. The MIM method did a better job by giving the translation "vomiting." Google translated it as "spit at the beginning," which is incorrect. Surprisingly, the translation of Bilingual Dictionary was very similar to that of Google. The poor performance of Google may be due to the concise nature of CCs. There is no context for the machine translation system to disambiguate "吐" as vomiting instead of spit.

Table 2.13 Example 2: Raw Chinese CC, Translations and Classification Results

Raw Chinese CC: 吐 晚上開始 (verbatim translation: vomiting. began this evening).			
Translation Method	Translation Outcome	Syndrome Outcome	Gold Standard
MIM	vomiting	GI	GI
Bilingual Dictionary	to spit , in the evening begin	UNKNOWN	
Google Translations	spit at the beginning	UNKNOWN	

Finally, in Table 2.14, the Chinese CC "昨天開始發燒 喘 (verbatim translation: fever and dyspnea. began yesterday)" is related to fever and respiratory syndrome. The

MIM method gave a correct translation while the Bilingual Dictionary translated "喘 (gasping)" and "發燒 (fever)" as "to gasp" and "have a high temperature." "to gasp" is recognized by the BioPortal CC classifier as related to respiratory syndrome. But "have a high temperature" could not be linked to fever syndrome in subsequent processing. The Bilingual Dictionary indeed had "have a fever" as its second translation. However, there was no simple way to decide when other translations instead of the first one should be used ex ante. Google Language Tool provided the correct translation for fever but gave "surge" as the translation for "喘 (gasping)." The translation for "喘 (gasping)" was wrong and I could not find any relation between the translated term "surge" and the original Chinese expression. A possible explanation is that the training dataset for Google translation system did not include documents in medical context and thus it has problem providing high quality medical translation.

Table 2.14 Example 3: Raw Chinese CC, Translations and Classification Results

Raw Chinese CC: 昨天開始發燒 喘 (verbatim translation: fever and dyspnea. began yesterday)			
Translation Method	Translation Outcome	Syndrome Outcome	Gold Standard
MIM	fever , dyspnea	RESP, LRESP, FEVER, CONST	RESP, LRESP, FEVER
Bilingual Dictionary	yesterday begin have a high temperature , to gasp	LRESP, RESP	
Google Translation	surge began yesterday fever	FEVER, CONST	

The above examples help confirm the discussion about the shortcomings of bilingual dictionary and machine translation approaches for multilingual syndromic classification in my literature review. Bilingual dictionaries often lack terminologies that are commonly seen in Chinese CCs. Machine translation performs better but may provide translations that are meaningless in medical context. The proposed MIM method constructs terminologies bottom-up using a statistical pattern extracting method thus can provide the best translation results for Chinese CCs.

2.9 Contributions

In this study I developed and evaluated chief complaint classification approaches that can handle CCs recorded in English and Chinese. The English CC classification approach can cope with multiple sets of syndrome definitions. At the core of this approach is the UMLS-based weighted semantic similarity score (WSSS) grouping method that is capable of automatically assigning previously un-encountered symptoms to appropriate symptom groups. An evaluation study shows that this approach can achieve a higher sensitivity, F measure, and F2 measure, when compared to the CC classification subsystem of EARS that has the same symptom grouping table and syndrome rules. This approach also outperforms RODS' CoCo naitve Bayesian classifier for syndrome categories that cover most CCs under consideration. As a side result, I also applied a bootstrapping-based statistical testing procedure to compare the performance of different methods. This procedure can be applied to compare sensitivity, specificity, positive predictive value, F measure, and F2 measure as long as the systems under

consideration share a common reference standard dataset in which the independent assumption among records is reasonable.

It is clear from the experimental results that the proposed approach can potentially improve the effectiveness of a CC classification system. My results indicate that semantic information captured in medical ontologies can be effectively leveraged to expand the coverage of the symptom grouping table automatically without extra knowledge acquisition efforts. The specific technical approach developed, the WSSS component, can be seen as a booster for an existing rule-based CC classification system.

I have also developed a Chinese CC classification approach by leveraging a Chinese-English translation module and the existing English CC classification approach. I used a statistical pattern extraction method based on the mutual information to extract important phrases from Chinese CCs and constructed mappings to English. The UMLS-based CC classifier, which was designed to process CCs in English, was used to process translated CCs. I compared the syndrome classification performance of the proposed translation method with those using the machine translation system provided by the Google Language Tool and a bilingual dictionary. Compared to Google Translation, my approach delivered significantly higher PPV, sensitivity, specificity, F measure, and F2 measure for most syndromic categories. I found similar results in the comparison between my approach and the translations provided by the bilingual dictionary.

The observed superior performance of my proposed Chinese-English mapping approach indicates that the 470 key phrases extracted from about one million Chinese CCs could cover common triage usage. I believe that with a more comprehensive study

of Chinese CC records, a set of standardized vocabulary could be constructed and my approach can be adopted in real-world applications. I do caution that languages are constantly evolving. Periodic reviews of extracted key phrases would be necessary to ensure inclusion of new phases.

CHAPTER 3. PROSPECTIVE INFECTIOUS DISEASE OUTBREAK DETECTION USING MARKOV SWITCHING MODELS

Detecting and controlling infectious disease outbreaks have long been a major concern in public health (Hu *et al.* 2007). Recent efforts in building syndromic surveillance systems have included increasing the timeliness of the data collection process by incorporating novel data sources such as emergency department (ED) chief complaints (CCs) and over-the-counter (OTC) health product sales (Niiranen *et al.* 2008). Research shows that these data sources contain valuable information that reflects current public health status (Espino & Wagner 2001; Ivanov *et al.* 2002; Chapman *et al.* 2005a; Chapman *et al.* 2005b). However, the noise caused by routine behavior patterns, seasonality, special events, and various other factors is blended with the disease outbreak signals. As a result, disease outbreak detection using the time series from syndromic surveillance systems is a challenging task.

In a typical syndromic surveillance system (Lombardo *et al.* 2003; Espino *et al.* 2004; Yan *et al.* 2008) the data are classified and aggregated to generate univariate or multivariate time series at a daily frequency. An example of a univariate time series is the daily ED visits associated with a particular syndrome (for example, the respiratory syndrome). An example of a multivariate time series is the number of daily visits with a particular syndrome from multiple EDs. If geographic information such as the ZIP code is available, the multivariate time series in these examples would be the daily counts of patients with a particular syndrome from the ZIP code areas near an ED.

Most time series outbreak detection methods follow a two-step procedure (Reis & Mandl 2003a; Reis & Mandl 2003b; Reis *et al.* 2003; Takeuchi & Yamanishi 2006). In the first step, a baseline model describing the "normal pattern" is estimated using the training data that usually contain a historical time series without outbreaks. The baseline model then is used to predict future time series values. In the second step, statistical surveillance methods such as the Shewhart control chart (Shewhart 1939; Montgomery 2005) or the Cumulated SUM (CUSUM) (Page 1954) method then take the prediction error (observed value minus predicted value) as the input, and output alert scores. Higher alert scores are usually associated with a higher risk of having outbreaks. When the alert scores exceed a predefined threshold, the alarm is triggered.

Two main problems exist for current detection methods. First, the two-step procedure is based on the assumption that there are no outbreaks in the training data. When a real-world dataset is used for training, the assumption is very hard to verify. Moreover, a full investigation of disease outbreaks during the data collection period is usually too expensive to conduct.

The validity of the detection results may be seriously impaired if it cannot be verified that the training data are outbreak-free. The estimated parameters of the baseline model may be biased by outbreak-related observations. Subsequent prediction and outbreak detection, as a result, may be negatively affected. The problem can seriously reduce the practical value of the outbreak detection method.

Second, existing time series detection methods also lack the ability to handle sporadic extreme values. Special events such as holidays and the media coverage of a

particular disease may cause spikes that are not associated with disease outbreaks (CDC 2006). These extreme values usually last for a very short time (often just one or two days) and do not affect subsequent time series values. Anomalies related to real disease outbreaks, on the other hand, usually show a prolonged upward drift. The magnitude of disease-related drift is usually much smaller compared to the sporadic spikes caused by special events. Many outbreak detection algorithms take advantage of these characteristics and accumulate the errors so that small increases can be detected effectively (Reis & Mandl 2003b; Reis *et al.* 2003; Buckeridge *et al.* 2005). The accumulation process, nevertheless, is susceptible to the presence of extreme values.

The deficiencies of current outbreak detection methods motivate my efforts to develop novel algorithms that can address these shortcomings. To deal with the problem of having outbreak-related observations in training data, a flexible statistical model must be used so that the model can adjust itself automatically when outbreak-related observations exist. In econometrics and time series literature, this is usually refereed to as the problem of modeling endogenous structural changes (Chu *et al.* 1996; Clements & Hendry 2006).

A natural way of modeling structure changes in a time series is introducing additional hidden state variables which control the underlying time series dynamics. The Markov switching models originally proposed by Hamilton are one popular model of this kind (Hamilton 1989). This family of models includes a hidden state variable that may have a different value in each period. It takes values of either 0 or 1 that correspond to different conditional means, variances, and autocorrelations of the time series. The

hidden state evolves following a first-order Markov process. That is, the current hidden state depends only on its historical values from the last period.

This hidden state method can be easily extended to handle extreme values. An additional hidden state can be included to model the presence of sporadic extreme values. With this additional hidden state, the model can distinguish between "normal" and "extreme" observations. That is, if a spike appears without signs that the sudden increase can be associated with drifts either before or after it, then the model can, based on the statistical evidence, assign the sudden increase as an extreme value instead of an outbreak. The negative effect of extreme values on outbreak detection can thus be reduced.

The main contribution of this paper is to present a prospective outbreak detection method that is robust to pre-existing outbreaks and extreme values. Prospective outbreak detection, as opposed to retrospective detection in which the entire set of the observations is available to the detection algorithms, assumes that only observations made prior to the time of the detection are available to the detection algorithms. Retrospective detection is useful primarily for offline analysis of historical data, whereas prospective detection is intended for use in monitoring incoming public health data streams in an online fashion. I utilized the Markov switching model that includes three hidden state variables in each period. The first hidden state variable models the disease outbreak state and the second hidden state variable models the presence of extreme value. If the extreme value exists, the third hidden state variable represents the size of the extreme value. I demonstrate that my approach outperforms several existing state-of-art outbreak detection algorithms using both simulated and real-world time series data.

This paper is organized as follows. Section 2 briefly introduces current outbreak detection methods and the Markov switching models. Section 3 presents my outbreak detection method. An evaluation study that uses both simulated and real-world data is summarized in Section 4. I conclude my paper in Section 5.

3.1 Background

Current time series outbreak detection methods mostly follow a two-step procedure: a base-line time series estimation step followed by a statistical surveillance step (Reis & Mandl 2003a; Reis *et al.* 2003; Takeuchi & Yamanishi 2006). I review these two major steps in this section.

Markov switching models, which belong to a broader class of statistical models that make use of hidden state variables, are also reviewed. I present the typical model settings and the estimation approaches.

3.1.1 Time Series Modeling

The first step in traditional outbreak detection methods is to develop a model that can describe the normal time series patterns. The most widely used model is the Autoregressive Integrated Moving Average (ARIMA) models of Box and Jenkins (Box & Jenkins 1970). The model setting can be described by three parameters: (p,d,q). The parameter p refers to the length of historical time series values that can affect current observations. The second parameter d specifies how many difference operations are required to make the time series stationary. The third parameter q specifies the length of historical error terms that can affect current observations. In a typical setting that does not

involve seasonal fluctuation, the observed time series is usually assumed to be stationary, that is, $d = 0$. Specifically, an ARIMA(p,0,q) model can be written as:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} \cdots + b_q \varepsilon_{t-q}$$

where y_t is the observed time series and ϵ_t is the error term. To ensure that the model "learns" the normal time series pattern, the data used for model estimation should be outbreak free. Given p and q, the parameter values (a_0, a_1, ..., b_q) can be estimated using likelihood maximization (Greene 2000). However, different model settings that correspond to different values of p and q may affect prediction accuracy. The values of p and q are usually determined by model selection criteria that take both goodness of fit and model complexity into consideration. Commonly used model selection criteria include Akaike information criterion (AIC) (Akaike 1970, 1973) and Bayesian information criterion (BIC) (Schwarz 1978). Note that the model selection criteria are closely related to the "cross-validation" evaluation approach (Bishop 2006) commonly used by the machine learning community (White 2006). In fact, cross-validation is asymptotically equivalent to AIC (Shao 1997).

Other modeling techniques such as the generalized linear model using Poisson distribution (Jackson *et al.* 2007), expectation-variance model (Wieland *et al.* 2007), and the Wavelet Model (Zhang *et al.*) have been evaluated in previous studies.

For the purpose of detecting outbreaks, there are two issues warranting further discussion: the modeling of the day-of-week and seasonal effects.

3.1.1.1 Day-of-Week Effect

The syndromic surveillance time series usually exhibits strong day-of-week effects. For example, there are usually more ED visits during the weekends than during the weekdays (Reis & Mandl 2003b). The variation among different day-of-weeks is usually assumed to be fixed. As an illustrative example, an ARIMA(1,0,0) model with a fixed day-of-week effect can be written as

$$y_t = w_1 d_{t,1} + w_2 d_{t,2} + \cdots + w_6 d_{t,6} + a_0 + a_1 y_{t-1} + \varepsilon_t$$

where $d_{t,i}$=0,1, *and* i = 1,2,3,4,5,6 are dummy variables indicating a particular day-of-week. For example $d_{t,1}$ = 1 if day t is a Monday and 0 otherwise. Note that I need only 6 dummy variables for 7 day-of-weeks because of the existence of the constant term a_0.

3.1.1.2 Seasonal Effect

Similar to the day-of-week effect that refers to a weekly cyclic pattern, the seasonal effect refers to a yearly cyclic pattern. Tri-geometric functions are commonly used to model deterministic seasonal fluctuation. This technique is usually referred to as the Serfling model (Serfling 1963; Brillman *et al.* 2005) which can be written as:

$$y_t = a_0 + b_1 \cos\left(\frac{2\pi t}{365.25}\right) + b_2 \sin\left(\frac{2\pi t}{365.25}\right) + w_1 d_{t,1} + w_2 d_{t,2} + \cdots + w_6 d_{t,6} + \varepsilon_t$$

Note that both day-of-week and seasonality are included in the model. The model can be refined by including more tri-geometric functions that correspond to semi-annual and even quarterly cyclic patterns. However, it has the obvious problem of assuming the same seasonal peaks and troughs across the whole monitoring period (Brillman *et al.* 2005). My preliminary experiments show that the Serfling model fits the observed

syndromic time series poorly especially when the seasonality is strong. The Serfling model assumes a particular shape of the time series that may not be empirically valid.

Other modeling techniques allow more flexible seasonal fluctuation across years. One possibility is to use the Holt-Winters exponential smoothing to model seasonality (Winters 1960; Holt 2004). An empirical study showed that, in the context of syndromic surveillance, Holt-Winter exponential smoothing outperformed the Serfling model in terms of prediction accuracy (Burkom *et al.* 2007).

The concept of the seasonal random walk (Hamilton 1994) can be applied to model the seasonal effect. The basic idea is that the same day-of-year should have the same expected value. Reis and his colleagues estimated the expected value using the trimmed-mean of historical time series value with the same day-of-year in an 8-year window (Reis & Mandl 2003b; Reis *et al.* 2003). The seasonal effect can then be filtered out by subtracting the observed value from the day-of-year expectation.

3.1.2 Statistical Surveillance Methods

For outbreak detection purposes, the prediction errors from the time series modeling step are further processed using statistical surveillance methods. Various statistical surveillance methods such as the Shewhart control Chart (Shewhart 1939), Cumulated Sum (CUSUM) (Page 1954), Exponential Weighted Moving Average (EWMA) (Montgomery 2005), Shiryaew-Roberts method (Shiryaev 1963; Roberts 1966) and the likelihood ratio methods (Frisen & De Mare 1991) can be applied for disease outbreak detection. However, most syndromic surveillance studies use the Shewhart control chart,

CUSUM, EWMA and their variations. My review focused mainly on these three methods. More detailed reviews can be found elsewhere (Sonesson & Book 2003).

The Shewhart control chart checks the t-value of the prediction errors period by period. It performs the best if large, isolated outbreaks are involved. However, since disease outbreaks often exhibit only small deviations in their early stages, the Shewhart control chart may not be the best choice for my purposes.

The CUSUM method minimizes the maximum value of the conditional expected delay "when the outcome before outbreak is the worst possible" (Moustakides 1986). It uses a recursive formula to accumulate the prediction errors:

$$C_t^+ = \max\left[0, e_t - K + C_{t-1}^+\right]$$

where e_t is the prediction error from the time series model and K is a predefined constant that is commonly referred to as the allowance. The alarm is triggered if C_t^+ exceeds a predefined threshold.

The EWMA method can be seen as a linear approximation of the likelihood ratio method (Frisen & De Mare 1991; Frisen 2003). The alert score is computed by accumulating forecasting errors with exponentially decaying weights. Similar to the CUSUM method, higher outbreak scores are usually associated with a higher risk of having an outbreak. The threshold can be determined from theoretical analysis or empirical studies (Chandrasekaran *et al.* 1995; Steiner 1999).

Some syndromic surveillance studies use a moving average scheme to accumulate forecasting errors (Reis & Mandl 2003b; Reis *et al.* 2003). Their studies have showed

that a linear increasing weighting schemes performed best in terms of outbreak detection ability.

3.1.3 Performance Measures

The most commonly used performance measure in statistical surveillance literature is the Average Run Length (ARL). ARL^0 denotes the expected run length until the first false alarm, and ARL^1 denotes the expected run length until an alarm when the process is out of control at the start of the surveillance (Chandrasekaran *et al.* 1995; Steiner 1999; Sonesson 2003; Sonesson & Book 2003).

These measures, nevertheless, are less intuitive under the context of disease outbreak detection. Most disease outbreak detection studies use per day sensitivity and false alarm rate (Reis *et al.* 2003; Jackson *et al.* 2007; Wieland *et al.* 2007). Sensitivity is the probability of having alarms on outbreak days. False alarm rate is the probability of having alarms on non-outbreak days.

3.1.4 Extreme Values in Syndromic Surveillance Time Series

Current surveillance methods are very sensitive to extreme values. The main reason is because the statistical surveillance methods accumulate the forecasting errors and there are no simple methods that can be used to filter out the extreme values. Burkom (Burkom 2007) proposed using a "reset" rule to bring down the alert scores when extreme values are known to be causing the elevated scores. However, it is not clear how to establish effective reset rules.

Common reasons behind the extreme values include holidays, media coverage, and special events (CDC 2006). However, existing studies have not offered help for handling the negative effects caused by the extreme values. Previous studies have used holiday dummies to absorb the holiday effects (Wieland *et al.* 2007). This technique, nevertheless, imposes an unrealistic assumption that all holidays have the same effect on the time series.

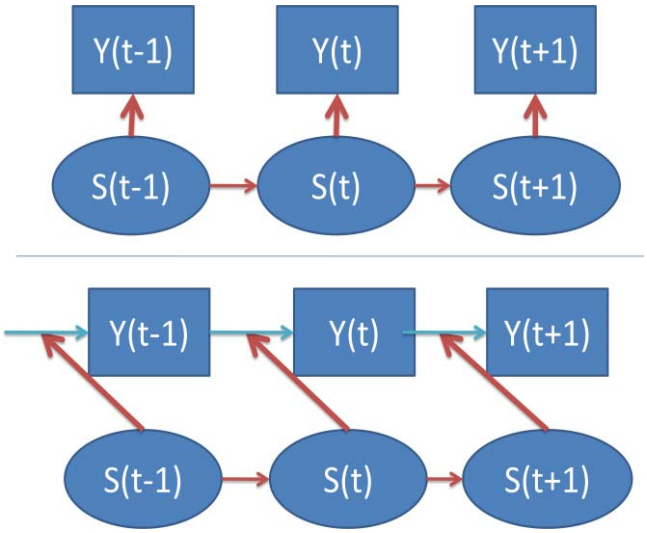
3.1.5 The Markov Switching Model

The Markov switching model belongs to the family of state-space models. A state-space model is a statistical model with hidden state variables controlling observable random variables. There are two types of equations in this model: the measurement equations and the transition equations (Kim & Nelson 1999). The measurement equation defines how hidden states affect the observable random variables. The transition equation, on the other hand, defines how the state variables evolve over time.

When the state variable is discrete, the state-space model is usually called the hidden Markov model (Baum & Petrie 1966; Baum & Egon 1967) or the Markov switching model (Hamilton 1989) depending on the choice of the measurement equation. The measurement equation in the hidden Markov model is usually formulated so that the observable random variables at period t only depend on the hidden state variables at the same period.

The Markov switching model addresses the weakness of the hidden Markov model by including lagged observations. The observable random variables in the Markov switching model depend on their historical values as well as the hidden state variables.

This setting makes the Markov switching model more suitable for time series related problems. Figure 3.1 illustrates the dependency difference between the Markov switching model and the hidden Markov model.



*The rectangles are observable random variables and the circles are hidden state variables. Arrows indicate the dependencies among variables.

Figure 3.1 Markov Switching Models (Upper Panel) and Hidden Markov Models (Lower Panel)

Strat and Carrat (Strat & Carrat 1999) applied the state-space model for disease outbreak detection. They used a two-state hidden Markov model on a weekly influenza-like illness (ILI) incidence and showed that the hidden Markov model clearly differentiated between epidemic and non-epidemic rates. However, as they pointed out in the conclusion, "the validity of the hypothesis that ILI incidence rates are independent conditional on the state is questionable." They also pointed out that autoregressive terms should be included for better performance. I am unaware of prior studies on applying Markov switching models for outbreak detection.

Most applications of the Markov switching models fall in the field of economics and finance. Notable examples are identifying macroeconomics business cycle (Hamilton 1989) and modeling changing interest rates regimes (Dahlquist & Gray 2000). A simple Markov switching model can be written as

$$y_t = a_{0,0} + a_{0,1}s_t + (a_{1,0} + a_{1,1}s_t)y_{t-1} + e_t \quad (1)$$
$$p(s_t = j \mid s_{t-1} = i) = p_{ij} \quad (2)$$
$$s_t \in \{0,1\} \quad (3)$$
$$e_t \sim N(0, \sigma^2) \quad (4)$$

Equation 1 defines how the hidden state variable s_t controls the dynamics of the observable random variable y_t. At an non-outbreak period ($s_t = 0$), y_t is determined by a drift term $a_{0,0}$ and the autoregressive parameter $a_{1,0}$. If an outbreak occurs ($s_t = 1$), the drift term increases to $a_{0,0} + a_{0,1}$ and the autoregressive parameter increases to $a_{1,0} + a_{1,1}$ (assuming $a_{0,1} \geq 0$ and $a_{1,1} \geq 0$). Equation 2 indicates that the hidden states evolve following a Markov process with transition probability p_{ij}.

Note that if I have a time series of T period, there are 4 parameters and T hidden state variables in Equation 1, together with 2 variables for transition probability in Equation 2 and a variance for error terms in Equation 4. I have more unknowns than the number of periods, which complicates the estimation process. I briefly discuss the model estimation issues below.

3.1.6 Model Estimation for the Markov Switching Model

Model estimation for the Markov switching model is much more complicated than that of the standard time series models such as the ARIMA models. The technical difficulty arises from the presence of unknown hidden states. In a simplified case

involving only one hidden outbreak state variable with two possible states and a total of T periods, a direct evaluation of the likelihood function involves a summation of all possible trajectories of hidden states. The time complexity is $O(2^T)$, which is intractable in practice. More sophisticated algorithms, which compute the posterior distribution of the hidden states using a forward-filtering-backward-smoothing (FFBS) procedure (Kim & Nelson 1999; Scott 2002) take only $O(2^3T)$ steps. The computation of the posterior distribution of the hidden states is required by many estimation methods such as the expectation-maximization (EM) algorithm (Dempster *et al.* 1977; Popescu & Wong 2005; Song *et al.* 2007) Gibbs sampling, and Markov Chain Monte Carlo (MCMC) (Albert & Chib 1993; Carter & Kohn 1994; Chib & Greenberg 1995). Note that to deliver the final optimal parameter estimation, these algorithms need to execute repeatedly until certain convergence criteria are met.

The EM algorithm finds the maximum of the likelihood function by iterating between calculating the expected value of state variables given current parameters and calculating the maximum of log likelihood given the expected state variables. It was applied to estimate the hidden Markov model in a previous outbreak detection study (Strat & Carrat 1999). Compared to other numerical optimization methods, the EM algorithm is more robust and usually converges if a maximum exists. However, it is possible that the algorithm converges to a local maximum instead. In practice, the EM algorithm is run with multiple initial values.

A serious drawback of the EM algorithm is the label switching problem (Scott 2002). The Markov switching model (and the hidden Markov model) is invariant under

arbitrary permutations of the state labels. As a result, I cannot be sure whether $s_t = 0$ is representing an outbreak or non-outbreak state before the estimation procedure is completed. The label switching problem is especially an issue when the Markov switching model is part of a larger automatic disease outbreak detection system.

Gibbs sampling (Albert & Chib 1993; Carter & Kohn 1994; Madigan 2005) is an alternative estimation method that can avoid the label switching problem. The Gibbs sampling iterates to draw random variables from conditional posterior distributions of parameters and state variables to simulate the full posterior distribution of parameters and state variables. Specifically, let $\Theta = \{\theta_1,...,\theta_k\}$ denote the unknown parameters (and state variables). By the Bayes Theorem, the posterior distribution $p(\Theta|Y)$ is proportional to the likelihood of $p(Y|\Theta)$ multiplying the prior of parameters $p(\Theta)$. The label switching problem can be avoided by imposing proper constraints on $p(\Theta)$. Gibbs sampling estimates parameters using a simulation-based method. The following steps can be used to simulate Θ from its posterior distribution. First, select initial values $\Theta^{(0)} = \{\theta_1^{(0)},...,\theta_k^{(0)}\}$. For $i = 1,2,...I$, iterate through the following steps:

Draw $\theta_1^{(i)}$ from $p(\theta_1|Y,\theta_2^{(i-1)},...,\theta_k^{(i-1)})$.

Draw $\theta_2^{(i)}$ from $p(\theta_2|Y,\theta_1^{(i)},\theta_3^{(i-1)},...,\theta_k^{(i-1)})$.

...

Draw $\theta_k(i)$ from $p(\theta_k|Y,\theta_1^{(i)},\theta_2^{(i)},...,\theta_{k-1}^{(i)})$.

Record $\Theta^{(i)} \equiv \{\theta_1^{(i)},\theta_2^{(i)},\ldots,\theta_k^{(i)}\}$

It has been shown that $\{\Theta^{(i)}\}$ converges to $p(\Theta|Y)$ (Besag 1974; Geman & Geman 1984). As a result, the posterior mean of θ_j can be estimated by the average of $\{\theta_j^{(i)}\}$, excluding certain "burn-in" iterations to minimize the effect of the initial value. The confidence intervals of the estimated parameters can also be calculated directly from $\{\theta_j^{(i)}\}$.

3.2 Outbreak Detection Using Markov Switching with Jumps (MSJ) Models

I developed my disease outbreak detection algorithm based on the Markov switching models (Hamilton 1989). Two hidden disease outbreak states (0 or 1; non-outbreak or outbreak) were assumed. To handle the sporadic extreme values, I included a jump component to filter their negative effects on outbreak detection. Seasonality was handled based on the concept of seasonal random walk.

My proposed MSJ model is described below:

$$y_t = g\left(Y^{t-1}\right) + z_t \tag{5}$$

$$z_t = \xi_t J_t + x_t \tag{6}$$

$$x_t = a_{0,0} + a_{0,1}s_t + \left(a_{1,0} + a_{1,1}s_t\right)x_{t-1} + \sum_{i=1}^{6} w_i d_{t,i} + \sum_{i=1}^{K} b_i v_{t,i} + e_t \tag{7}$$

$$s_t \in \{0,1\} \tag{8}$$

$$J_t \in \{0,1\} \tag{9}$$

$$p\left(s_t = j \mid s_{t-1} = i\right) = p_{ij} \tag{10}$$

$$e_t \sim N\left(0, \sigma^2\right) \tag{11}$$

$$\xi_t \sim N\left(0, \sigma_a^2\right) \tag{12}$$

$$g\left(Y_{t-1}\right) = med\left\{\bar{y}_{t-m}, \bar{y}_{t-2m}, \bar{y}_{t-3m}\right\} \tag{13}$$

$$\bar{y}_{t-im} = \frac{y_{t-im-3} + y_{t-im-2} + \cdots + y_{t-im+3}}{7} \tag{14}$$

where $Y^{t-1} = (y_1, y_2, \ldots, y_{t-1})$ and $m = 365$. The hidden state variable $s_t = 1$ indicates period t is an outbreak period, 0 otherwise.

Equation 5 filters out the seasonal fluctuation by subtracting the day-of-year expectation from observed time series values. The day-of-year expectation is estimated using the historical values within a day-of-year window in the past three years (Eq. 13-14). The next equation (Eq. 6) further decomposes the residual (z_t) into normal variation (x_t) and a possible jump component. If a jump exists ($J_t = 1$), then ξ_t is the size of the jump. Equation 7 articulates the dynamic behavior during outbreak and non-outbreak periods. The hidden state variable s_t controls the constant term and an autoregressive coefficient. The variables $d_{t,i}$ are day-of-week dummies. The exogenous variables $v_{t,i}$ are optional controlling factors. Environmental variables such as pollen level and temperature are two possibilities. If necessary, more lagged dependent variables can also be included. For example, I can set $v_{t,1} = x_{t-2}$, $v_{t,2} = x_{t-3}$, ..., $v_{t,6} = x_{t-7}$. As defined in Equation 10, the transition of s_t follows a first-order Markov process.

Compared to conducting outbreak detection using a baseline time series model combined with a statistical surveillance method, my approach provides the following advantages. First, the alert scores ($p(s_t = 1|Y_t)$) of my approach have a clear and intuitive interpretation. Most existing outbreak detecting methods output alert scores that do not have clear meanings. The only way to make sense of the alert scores is to compare the scores with an established threshold. The alert score of my detection algorithm, without reference to any thresholds, can be interpreted as the outbreak probability given available information.

Second, my algorithm provides an estimated outbreak size in addition to outbreak probability. In traditional outbreak detection methods, it is not easy to estimate the outbreak size directly from the alert statistics or estimated parameters. My method allows the model to recognize different temporal dynamics in different periods. The outbreak size can be calculated directly from the estimated parameters. The information could be valuable for the planning of public health intervention.

Third, the jump component gives my algorithm the ability to separate sporadic extreme values from slow-moving disease outbreaks. The additional information provides flexibility that is valuable for different surveillance needs.

3.2.1 Changing Dynamics and Outbreak Size

The hidden variable s_t plays an important role in determining the dynamics of x_t. Consider a simplified setting with no day-of-week effect ($w_i = 0$) nor exogenous variables ($b_i = 0$). If I have $s_t = 0$ for all time except $t = t_1$, then the observed value increased by $\Delta_{t_1} \equiv a_{0,1} + a_{1,1}y_{n-1}$ at t_1, ignoring the effect of the noise (e_t). Note that the autoregressive coefficient $a_{1,1}$ also plays an role in determining the magnitude of the increase at time t_1. After this time point, the effect of Δ_{t_1} decreases exponentially. The scenario is similar to dropping a group of infected persons in a large community at period t_1 and seeing the disease starting to spread. However, since infected persons recover from the disease quickly, the disease dies out quickly as well.

If $s_t = 1$ for $t = t_1, t_1 + 1, \ldots, t_1 + q$, the effect of increased constant term and autoregressive coefficients accumulates during the outbreak periods until it reaches the new stable level. The new long-term mean can be found by writing x_t as a function of $a_{i,j}$

and e_t only. A simple computation gives $E[x_t|s_t = 1] \equiv \overline{m}_2 = (a_{0,0} + a_{0,1}) / (1 - a_{1,0} - a_{1,1})$. Similarly, the long-term mean of non-outbreak periods is $E[x_t|s_t = 0] \equiv \overline{m}_1 = a_{0,0} / (1 - a_{1,0})$. The outbreak size is the difference between $\overline{m}_2$ and $\overline{m}_1$.

3.2.2 Model Estimation

Gibbs sampling is used for model estimation. I need to estimate the following sets of coefficients and hidden states: time series coefficients $A = (a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1})$, day-of-week coefficients $W = (w_1, w_2, \ldots, w_6)$, exogenous variable coefficients $B = (b_1, b_2, \ldots, b_k)$, variance of the error term (σ^2), transition probability $P = (p_{00}, p_{11})$, hidden outbreak state $S^T = (s_1, s_2, \ldots, s_T)$, hidden jump state $J^T = (J_1, J_2, \ldots, J_T)$, hidden jump size $\Xi^T = (\xi_1, \xi_2, \ldots, \xi_T)$, and variance of jumps (σ_a^2).

To facilitate the simulation of random variables from the posterior distributions, conjugate priors are used for all parameters. As discussed in the Appendix, all conditional posteriors follow well known statistical distributions and are summarized in Table 1. The dot ($\bullet$) in Table 3.1 indicates the conditioning on other parameters and hidden states. To increase the efficiency of sampling s_t, the FFBS procedure is used.

Table 3.1 Conditional Posterior Distributions

$(A, W, B)\| \bullet \sim$ Multivariate Normal
$\sigma^2\| \bullet \sim$ Inverse Gamma
$\xi_t\| \bullet \sim$ Normal
$J_t\| \bullet \sim$ Binomial
$s_t\| \bullet \sim$ Binomial
$\sigma_a^2\| \bullet \sim$ Inverse Gamma
$p_{ii}\| \bullet \sim$ Beta

It should be noted that to avoid the label switching problem, I constraint the parameter sampling results so that $\overline{m}_1 < \overline{m}_2$ is satisfied. If the constraint is violated, (A, W, B) are redrawn until the constraint is satisfied.

3.2.3 Prospective Outbreak Detection

Given an up-to-date time series, prospective outbreak detection answers the question "What is the probability of having a disease outbreak today?" Letting t denote the current time period, I want to estimate $p(s_t = 1|Y^t)$, where s_t is the hidden outbreak state and Y^t is the vector contains all time series values up to time t. When a new time series value arrives in the next period, the system needs to re-run the model and provide the estimation of $p(s_{t+1}|Y^{t+1})$.

My preliminary experiments found that direct implementation of the estimation algorithm provides little valuable outbreak information because the algorithm became too sensitive to small changes. The algorithm tried to scrutinize all small changes and tended to over react to those changes. To overcome this difficulty, I developed a regulation technique to desensitize the algorithm so that small, unimportant changes would be ignored.

3.2.4 Desensitization for Prospective Outbreak Detection

The desensitization technique is an extension of the solution for the label switching problem. To make the algorithm ignore small, unimportant changes, I rejected the parameter sampling results that indicated small changes. Specifically, I chose g as the minimal outbreak size that I wanted to detect. I let $a_{0,0}^{(c)}, a_{0,1}^{(c)}, a_{1,0}^{(c)}, a_{1,1}^{(c)}$ be the sampling

result of the *c*-th iteration. I rejected the sampling result if $\overline{m}_1^{(c)} \geq \overline{m}_2^{(c)}$ - *g*. The coefficient *g* is set to 5% of the time series mean during the training period. Also, the autoregressive coefficient needs to have a value between -1 and 1 to ensure that the time series is stationary. The desensitization procedure is summarized in Algorithm 1.

Algorithm 1 Desensitization Procedure

repeat

Draw $(A^{(c)}, B^{(c)}, W^{(c)})$ from $(A, B, W)|\bullet$

$\overline{m}_1^{(c)} \leftarrow a_{0,0}^{(c)} / \left(1 - a_{1,0}^{(c)}\right)$

$\overline{m}_2^{(c)} = \left(a_{0,0}^{(c)} + a_{0,1}^{(c)}\right) / \left(1 - a_{1,0}^{(c)} - a_{1,1}^{(c)}\right)$

until $\overline{m}_1^{(c)} < \overline{m}_2^{(c)} - g$ and $\left|a_{1,0}\right| < 1$ and $\left|a_{1,0} + a_{1,1}\right| < 1$

return $(A^{(c)}, B^{(c)}, W^{(c)})$

3.2.5 Prior Distributions

While some parameters of the prior distributions are quite robust to various circumstances, others need to be customized case by case. I applied a simple AR(1) model with day-of-week effect on the training data with seasonality removed. The estimated variance of the error term is used to set up the parameters for the prior of σ^2 and σ_a^2. The estimated day-of-week effects are used to set up the prior of w_i. The prior distributions used in this study are summarized in Table 3.2.

Table 3.2 Prior Distributions

Parameter	Distribution	Parameter
$\{a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1}\}$	Multivariate Normal(*M*, *V*)	$M = \{0,0,0.15,0.6\}$, $\{v_{ii}\}_{i=1}^{4} = \{400,400,3,3\}$
σ^2	Inverse Gamma	$\alpha = 3, \beta =$ est. variance $\times (\alpha - 1)$
σ_a^2	Inverse Gamma	$\alpha = 3, \beta =$ est. variance $\times 5(\alpha - 1)$

$\{w_1,\cdots,w_6\}$	Multivariate Normal(M, V)	M is est. from the training data $\{v_{ii}\}_{i=1}^{6}=\{v_a,v_a,v_a,v_a,v_a,v_a\}$; $v_a=\max(100,5\max(M))$
P_{11}	Beta	a = 2, b = 0.2
P_{22}	Beta	a = 2, b = 0.1

The off-diagnose elements of V is set to zero

3.2.6 Summary of the Estimation Procedure

Given a time series covering period 1 to t_1, my goal is to estimate the outbreak probability of period t_1, together with other relevant parameters and hidden state variables. Using Gibbs sampling for estimation, I need to choose the total number of iteration B and the "burn-in" iteration b. The sampling results between iteration b + 1 and B are then used to compute the outbreak probability (alert score) and the estimates of other parameters. The pseudo code that summarizes the procedure can be found in Algorithm 2. I implemented my approach on R, an open-source statistical software (http://www.r-project.org/).

Algorithm 2 Prospective Outbreak Detection Using the Markov Switching with Jumps (MSJ) Model

for c = 1 to B **do**
  $(A^{(c)},B^{(c)},W^{(c)}) \leftarrow$ Desensitization()
  Draw $\sigma^{2(c)}$ from $\sigma^2|\bullet$
  Draw $s_{t1}^{(c)},s_{t1-1}^{(c)},\ldots,s_1^{(c)}$ using FFBS
  Draw $J_t^{(c)}$ from $J_t|\bullet$ for $t = 1,2,\ldots,t_1$
  Draw $\xi_t^{(c)}$ from $\xi_t|\bullet$ for $t = 1,2,\ldots,t_1$
  Draw $\sigma_a^{2(c)}$ from $\sigma_a|\bullet$
end for
$\hat{p}(s_{t1}=1\mid Y^t)\leftarrow \sum_{c=b+1}^{B} s_{t1}^{(c)}/(B-b+1)$

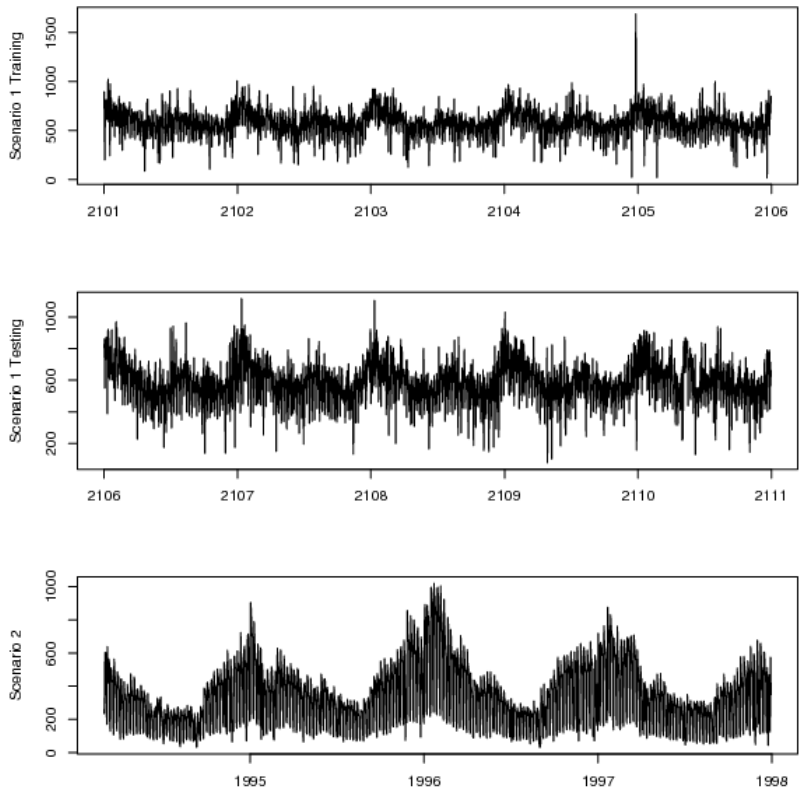
3.3 Evaluation Study

I developed two disease outbreak scenarios to evaluate my approach. Scenario 1 is aggregated over-the-counter (OTC) anti-diarrheal and anti-nauseant sales simulated based on a real-world dataset developed by the International Society for Disease Surveillance (ISDS) for a three-year time period (ISDS 2008). The outbreaks in this scenario were simulated based on "a large waterborne outbreak of Cryptosporidium [which] occurred in the Battleford area of Saskatchewan during the spring of 2001. Due to the prolonged, less severe nature of Cryptosporidium, many infected residents self-medicated, evidenced by an increase of OTC anti-diarrheal and anti-nauseant product sales during the outbreak." (cf. https://wiki.cirg.washington.edu/pub/bin/view /Isds/TechnicalContest).

This dataset contains 5 years of training data and 30 5-year time series datasets with outbreaks for model testing. The starting date of training data is marked "1/1/2010." The training data and one of the 30 testing time series are plotted at the top and middle panels of Fig. 2. Note that the plotted testing dataset contains an outbreak starting from "4/15/2110" that lasts for 54 days.

Scenario 2 used a real-world clinic visit time series and simulated outbreaks following the standard approach widely-used in the syndromic surveillance literature (Reis & Mandl 2003b; Reis *et al.* 2003; Burkom *et al.* 2007). I imposed simulated anthrax outbreaks on the clinic visit time series collected from a metropolitan area. The clinic visit is classified into syndromes using ICD-9 code according to the definitions from CDC (cf. http://www.bt.cdc.gov/ surveillance/ syndromedef/ word/ syndromedefinitions.doc).

The dataset covers the period from 2/28/1994 to 12/30/1997 with a total of 1402 days. Observations before 12/31/1995 were reserved for model training. Outbreak periods were randomly chosen between 1/1/1996 to 12/30/1997. Since the respiratory syndrome is the most common syndrome for early infection of inhalational anthrax, I focused on detection disease outbreak using aggregated clinic visits with the respiratory syndrome in this study. The lower panel of Figure 3.2 plots the respiratory syndrome count time series used in this study.



*The upper panel is the training data of Scenario 1. The middle panel is one of the testing data of Scenario 1. The lower panel is the original real-world time series used in Scenario 2.

Figure 3.2 Time Series Plots of Research Testbeds

I chose $B = 300$ and $b = 100$. Preliminary experiments on simulated data showed that this setting was appropriate for estimating the hidden states and parameters.

I discuss my outbreak simulation method for Scenario 2 in more detail and benchmark surveillance methods for both Scenarios in sequence. The last subsection presents experimental results.

3.3.1 Simulating Disease Outbreaks for Scenario 2

There are two major components in simulating the disease outbreak caused by inhalational anthrax (Buckeridge *et al.* 2005). The first component is the disease progression of infected persons (Wein *et al.* 2003). The second component is the health-care seeking behavior of infected persons. Since I aimed to focus on temporal disease outbreak detection, the spatial dispersion of anthrax spores and infection in different areas (Wein *et al.* 2003) were not considered in the simulation.

At the beginning of the simulation, I assumed that, in total, there are S infected persons. For each infected person, the disease progresses through three states: incubation, prodromal, and fulminant. The length of each state follows a log-normal distribution. Disease symptoms start to appear in the prodromal state. An infected person may have respiratory, gastrointestinal, or fever syndromes in the prodromal state. When the infected person is in the fulminant state, the person may exhibit shock syndrome or neurological syndrome. Since syndromes in the fulminant state are not the focus in this study, this state is not simulated. Parameters used in the simulation are summarized in Table 3.3 and Table 3.4.

Table 3.3 Parameters for Anthrax Outbreak Simulation: Disease Progression

Parameter	Value	Source
Incubation duration, median	11 days	[68]
Incubation duration, dispersion	2.04 days	[68]
Prodromal duration, median	2.50 days	[68]
Prodromal duration, dispersion	1.44 days	[68]

Table 3.4 Parameters for Anthrax Outbreak Simulation: Health-care Seeking at Prodromal State

Parameter	Value	Source
Prob. of seeking care	0.4	[21]
Prob. of respiratory syndrome	0.7	[69]
Prob. of gastrointestinal syndrome	0.2	[69]
Prob. of fever syndrome	0.1	[69]

I set S to 15,000. This setting corresponds to an average peak of 566 patients. The outbreak period begins when anthrax spores are released and ends when more than 90% of infected persons with the respiratory syndrome have realized. Since there are usually a small number of patients with a long incubation period, the outbreak period may be artificially long just because few persons have late onset of the syndrome. The 90% cut-off ensures that the outbreak period covers the most intense period of anthrax outbreaks.

The outbreak signals were imposed on the real-world time series with a starting time chosen randomly between 1/1/1996 to 12/30/1997. I generated 50 synthetic datasets for evaluation, each containing one simulated outbreak.

3.3.2 Benchmark Temporal Detection Methods

I chose the Serfling model with the CUSUM method as my first benchmark detection method. The Serfling model is one of the most popular time series models that

incorporate seasonal fluctuations (Serfling 1963; Brillman *et al.* 2005). The model can be written as follows:

$$y_t = a_0 + \sum_{i=1}^{6} d_{t,i} w_i + b_1 \cos(2\pi t / 365.25) + b_2 \sin(2\pi t / 365.25) + e_t$$
$$e_t \sim N(0, \sigma^2)$$

where $d_{t,i}$ is day-of-week dummies. Note that I only need 6 day-of-week dummies.

As mentioned earlier, the observations in the training period were reserved for model training. When the detection system started, one-step-ahead prediction was made using the most up-to-date parameters. The standardized prediction error was fed into the CUSUM method. The process was repeated until the end of the testing period. This benchmark detection method is referred to as the S+CUSUM method in the subsequent discussions.

The second benchmark method, the trimmed-mean seasonal ARMA model, was implemented following [15],[14]. Observations made in the training period were used to estimate the overall mean, the mean for day-of-week and the trimmed-mean for day-of-year. For observations in the testing periods, the overall mean, the mean for day-of-week and the trimmed-mean for day-of-year were subtracted from the raw count. The de-meaned counts were then fed into an ARMA(p,q) to filter out high-frequency dependency. I chose $p = 1$ and $q = 0$ for Scenario 1 and $p = 7$ and $q = 0$ for Scenario 2 according to Akaike Information Criteria (AIC). One step ahead predictions were calculated and prediction errors from the ARMA model were then weighted according to a linear-increasing pattern to compute the alert score. New observations were included for

parameter estimation when available. This method is referred to as T+MA (Moving Average) and my approach is referred to as MSJ in subsequent discussion.

3.3.3 Evaluation Metrics

I use two common syndromic surveillance evaluation metrics in this study (Brillman *et al.* 2005; Burr *et al.* 2006; Rolka *et al.* 2007). The first metric is detection timeliness (ISDS 2008). It measures the delay from the onset of the disease outbreak to the first detection of the disease outbreak at a given level of false alarm rate. If an outbreak is not detected across the whole outbreak period, the delay time is counted as the maximum outbreak length in all testing runs (65 days for Scenario 1 and 28 days for Scenario 2).

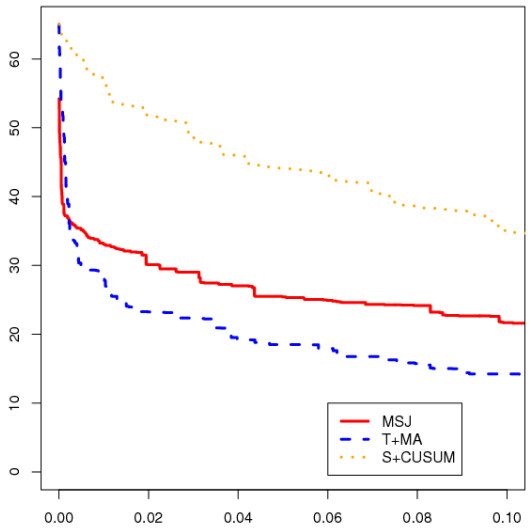
The false alarm rate (FAR) is defined as the probability of having an alarm for non-outbreak days (Reis & Mandl 2003b; Reis *et al.* 2003; Jackson *et al.* 2007; Wieland *et al.* 2007). For example, an FAR of 0.1 means that, on average, there are about 365 × 0.1 = 36 days with false alarms in a year with no outbreaks.

The second metric is per-day detection sensitivity (Reis & Mandl 2003b). This metric measures the probability of detecting an outbreak on an outbreak day. Given an alert threshold h, let o_h be the number of outbreak days that have alert scores exceeding h and Q be the total outbreak days in the testing dataset, the detection sensitivity is o_h/Q.

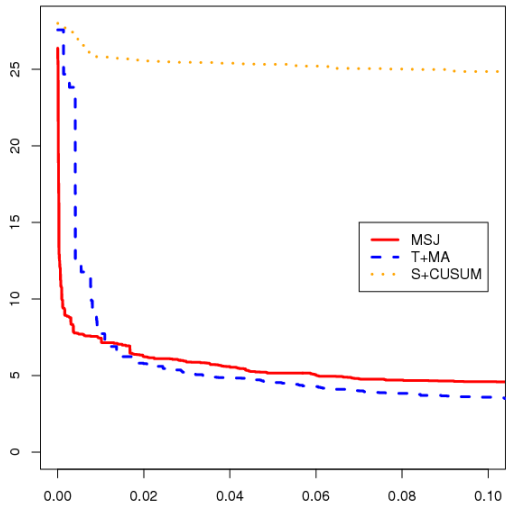
3.3.4 Results

Figure 3.3 (a and b) plot the detection timeliness of Scenario 1 and 2 at different false alarm rates. The solid line corresponds to the delay of my approach. Dashed and

dotted lines indicate the T+MA and S+CUSUM methods. I consider the false alarm rate only when it is less than or equal to 0.1. A detection system with a false alarm rate higher than 0.1 is usually considered impractical because of the high cost associated with confirming the false alarms. As clearly observed in the figures, my approach started with the lowest detection delay compared to other benchmark methods. As the FAR increased, the delay decreased for all methods. The detection delay of my approach remained the lowest for a range of FAR and then the T+MA method became the lowest.



(a) Scenario 1 Detection Timeliness at Different False Alarm Rates



(b) Scenario 2 Detection Timeliness at Different False Alarm Rates

Figure 3.3 Performance Comparison: Timeliness

The S+CUSUM method had the worst detection speed. The detection delay is at the maximum outbreak length (65 days for Scenario 1 and 28 days for Scenario 2) when the FAR is 0. It indicated that the S+CUSUM method could detect no outbreaks when no

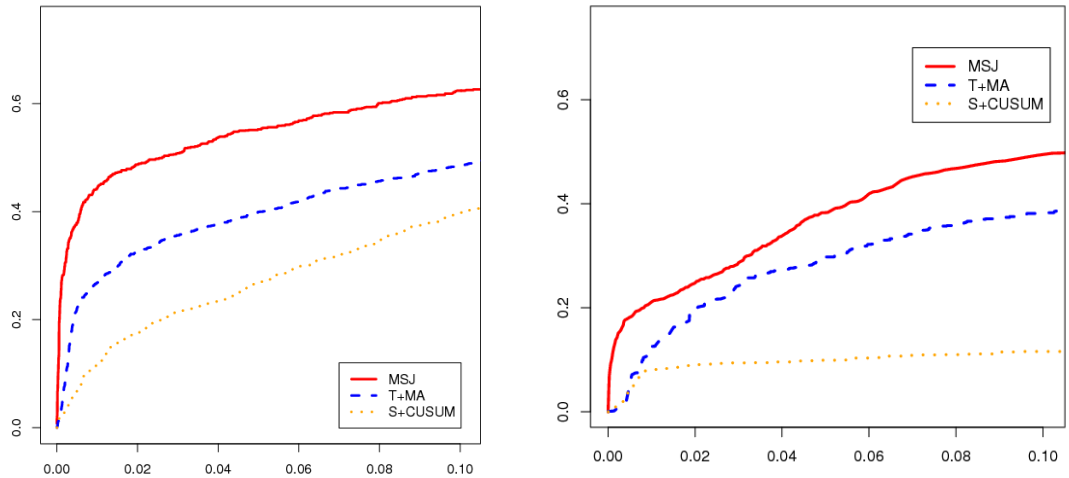
false alarms were allowed. The improvement of detection delay was the smallest among all methods when FAR increases.

To further analyze the detection delay, I test the hypothesis that my approach has a detection delay equal or larger than the benchmark methods. The last two columns of Table 3.5 report the detection delay differences and the p-values (in the parenthesis) for hypothesis tests. My approach has lower detection delay than the S+CUSUM method. The difference is, in most cases, significant at conventional confidence levels for Scenario 2 but not Scenario 1. On the other hand, my method has higher detection delay compared to the T+MA method except for the case of $FAR = 0$. The difference, nevertheless, is not significant. In general, the detection timeliness of my approach proved to be better than the S+CUSUM method and is at the same level as the T+MA method.

Table 3.5 Comparison of Detection Timeliness

FAR	MSJ (d_{msj})	S+CUSUM (d_{cu})	T+MA (d_{ma})	d_{cu} - d_{msj}	d_{ma} - d_{msj}
			Scenario 1		
0.0000	53.4	65.0	65.0	11.57 (0.22)	11.57 (0.22)
0.0125	30.2	52.8	22.7	22.57 (0.10)	-7.57 (0.73)
0.0250	27.0	50.0	20.1	23.07 (0.11)	-6.82 (0.74)
0.0500	22.6	42.6	15.2	20.04 (0.16)	-7.43 (0.74)
0.0750	21.4	37.4	12.8	16.00 (0.20)	-8.57 (0.77)
0.1000	18.6	32.9	10.6	14.36 (0.22)	-7.96 (0.82)
			Scenario 2		
0.000	13.4	28.0	27.6	14.64 (0.06)	14.22 (0.07)
0.012	7.1	25.7	6.9	18.60 (0.00)	-0.24 (0.52)
0.025	6.1	25.5	5.5	19.38 (0.00)	-0.64 (0.70)
0.050	5.2	25.3	4.5	20.16 (0.00)	-0.62 (0.69)
0.075	4.7	25.0	3.9	20.30 (0.00)	-0.86 (0.77)
0.100	4.7	25.0	3.7	20.32 (0.00)	-0.96 (0.78)

Figure 3.4 (a and b) plots the detection sensitivity of all detection methods under the FAR I considered. My approach had the highest detection sensitivity compared to the benchmark methods. The T+MA came in second, followed by the S+CUSUM method. The performance gaps remained consistent at FAR greater than 0.0125. In fact, the sensitivity of my approach was on average 0.15 higher than the T+MA method in Scenario 1 and 0.09 higher in Scenario 2. The performance gap was even larger (0.28 and 0.26) compared with the S+CUSUM method. In some FARs, the gaps represented a relative difference of more than 100%.



(a) Scenario 1 Sensitivity at Different False Alarm Rates

(b) Scenario 2 Sensitivity at Different False Alarm Rates

Figure 3.4 Performance Comparison: Sensitivity

Table 3.6 summarizes the detection sensitivity of the surveillance methods considered. The first three columns report the sensitivity at different FARs. The last two columns report the differences of sensitivity between my approach and the benchmark methods. The parentheses in the last two columns are the p-values of the statistical tests

hypothesizing equal or worse performance of my approach. The p-value is computed using a bootstrapping method based on the paired comparison of all testing days (Lu *et al.* 2008b).

Table 3.6 Comparison of Detection Sensitivity

FAR	MSJ (s_{msj})	S+CUSUM (s_{cu})	T+MA (s_{ma})	s_{ma} - s_{msj}	s_{cu} - s_{msj}
Scenario 1					
0.0000	0.008	0.000	0.000	0.008 (0.00)	0.008 (0.00)
0.0125	0.459	0.140	0.286	0.320 (0.00)	0.174 (0.00)
0.0250	0.498	0.195	0.341	0.302 (0.00)	0.156 (0.00)
0.0500	0.552	0.267	0.399	0.286 (0.00)	0.153 (0.00)
0.0750	0.590	0.331	0.450	0.259 (0.00)	0.141 (0.00)
0.1000	0.624	0.395	0.484	0.228 (0.00)	0.140 (0.00)
Scenario 2					
0.0000	0.070	0.000	0.001	0.070 (0.00)	0.069 (0.00)
0.0125	0.217	0.083	0.142	0.134 (0.00)	0.075 (0.00)
0.0250	0.266	0.093	0.216	0.173 (0.00)	0.050 (0.00)
0.0500	0.383	0.099	0.298	0.284 (0.00)	0.086 (0.00)
0.0750	0.461	0.109	0.355	0.352 (0.00)	0.106 (0.00)
0.1000	0.497	0.116	0.382	0.381 (0.00)	0.115 (0.00)

Because of the large number of testing days, it is not surprising to see significant testing results across all FARs. The p-values indicate that my approach is significantly better than the two benchmark methods across all FARs under consideration. The testing results confirm that my approach indeed performs better in term of sensitivity across all FARs under consideration.

To better understand the intuition behind the performance difference, I present the alert scores around an outbreak period from the three surveillance methods. As plotted in Figure 3.5, the top panel is the input time series to the surveillance methods. The following three panels present the alert scores from my approach, the S+CUSUM method,

and the T+MA method. The solid blue lines in the three lower panels mark the beginning and ending of the outbreak period. The horizontal green dashed lines mark the thresholds corresponding to a FAR of 0.0125. The alert scores higher than the thresholds are marked as outbreak days by the three methods.

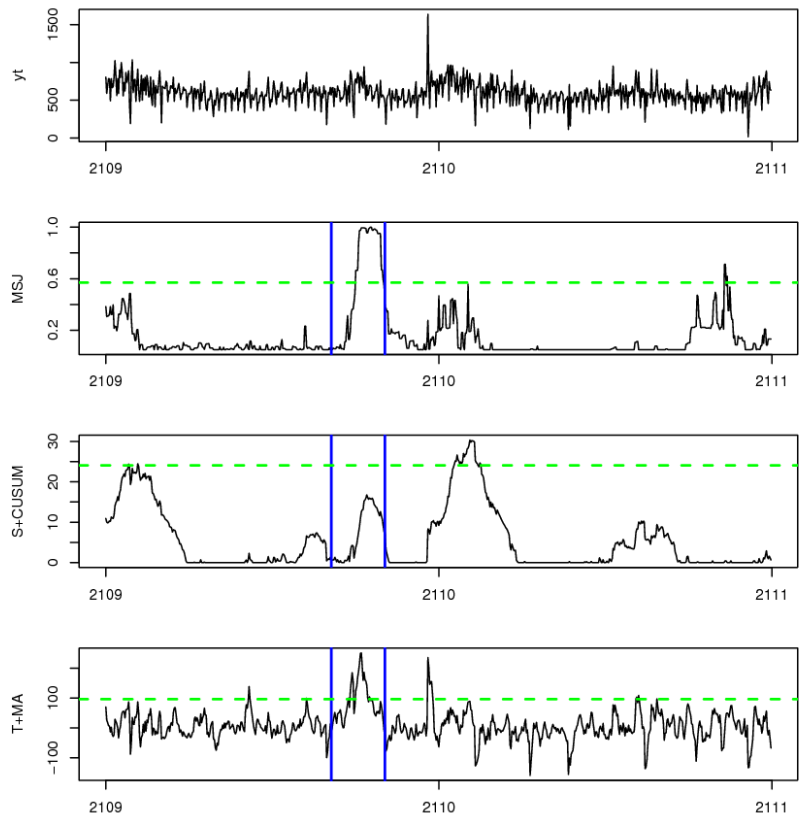


Figure 3.5 A Comparison of the Alert Scores from Three Surveillance Methods

In this example, the outbreak lasted for 60 days. The MSJ method detected the outbreak first at the 28th day and continued through the 58th day. The T+MA method first detected the outbreak at the 20th day. However, the alert score fell below the

threshold the next day and fluctuated around the threshold for the next few days. It was not until the 28th day when the alert scores move beyond the threshold steadily and fell below the threshold again at the 42th day. The S+CUSUM method did not output scores higher than the threshold during the outbreak period. In total, my method detected 31 outbreak days, compared to 20 days by the T+MA method and 0 day by the S+CUSUM method.

Clearly, my approach had the best sensitivity. More than half of the outbreak days were correctly detected. Compared to my approach, the T+MA method detected the onset of the outbreak with similar speed. However, my approach did a much better job in detecting the end of the outbreak. The ability to better detecting the ending of an outbreak contributes to the higher detection sensitivity of my approach.

One salient characteristic of this input time series is a jump at the end of the year "2109." Both the T+MA method and the S+CUSUM method were seriously disrupted by the jump, causing elevated alert scores. My approach, on the other hand, filtered out the jump effectively and output reasonably low scores around the jump day. The unique ability to filter out the jumps keeps the false alarm rate low and leads to a better detection sensitivity.

3.4 Conclusions

Disease outbreak detection using time series data is an important function for syndromic surveillance systems. I treated the disease outbreak as hidden outbreak states and developed a Markov switching with jumps model for syndromic surveillance. To handle the negative effect caused by the jumps in the observed time series, I extended the

Markov switching model to include an extreme value filtering component. The negative effect of jumps can be successfully filtered out, which led to a lower false alarm rate.

I evaluated my disease outbreak detection approach using both simulated and real-world baseline time series, together with outbreaks simulated following established methods. Two benchmark surveillance methods were included. The first benchmark method, S+CUSUM, uses the Serfling model to filter out seasonal fluctuation and then applies the CUSUM method on standardized prediction errors. The second benchmark method, T+MA, uses trimmed-mean seasonal ARMA model and computes the alert scores using linear increasing weights. The evaluation results showed that my method achieved a similar level of detection timeliness and higher detection sensitivity compared to the benchmark outbreak detection methods. My approach had a detection sensitivity 23% to 328% higher than the benchmark methods.

I applied an earlier version of the detection methods reported in this paper in a disease outbreak detection algorithm competition organized by the International Society for Disease Surveillance in 2007. My method ranked the third among participating methods. The performance gap between my method and the best-performing algorithm is within 6%.

The results reported in my study suggest a promising future for the use of hidden state variables to model the changing dynamics of observed surveillance time series. I plan to extend my approach to outbreak detection with multiple data streams through multivariate time series analysis based on Markov switching. I am also exploring

opportunities to apply the approach developed in this paper in areas beyond infectious disease informatics. One such area is sensor data integration and anomaly detection.

CHAPTER 4. TEXT-BASED RISK RECOGNITION FOR BUSINESS DECISION MAKING

Risk can be interpreted as the potential events and trends that may impact a business's growth trajectory and shareholder value (COSO 2004; Slywotzky & Drzik 2005). Some of these potential events and trends may provide growth opportunities while others may damage a business's future. Increased worldwide competition, new technology development, and the Internet have all contributed to an increasingly turbulent business environment. Systematically collecting and analyzing risk-related information has become critical for businesses facing today's volatile environment.

Decision makers rely on timely and accurate reports of risk-related information to avoid surprises, identify threats and opportunities, and gain competitive advantages. While decision makers may have access to more complete information about their own businesses, information about other firms is mainly obtained from various external sources such as on-line news websites, analysts' reports, government websites (e.g., the EDGAR system), and e-commerce, blog, forum, and social media websites. A business may want to monitor risk-related information about other firms for at least three reasons. First, firms in the same industry often share the same challenges and opportunities because of similar production technologies. Second, new developments of competitors are often directly related to future growth opportunities. Finally, a firm depends on its suppliers to provide inputs, and on buyers to distribute its products and services. Risks faced by its suppliers and consumers may potentially impact the firm as well. Monitoring

risks faced by same-industry firms, competitors, and firms in the supply chain is an expedient way to monitor the changing business environment.

While part of the risk-related information may be quantitative in nature, it is often conveyed in qualitative textual descriptions due to its ambiguous nature. Understanding, tracking, and analyzing risk-related information embedded in textual data often require processing large amounts of textual data. Conducting large scale manual analysis is labor intensive and costly. Moreover, the boredom and fatigue associated with processing large amounts of textual data may reduce overall accuracy, thus diminishing one of the main advantages of manual processing.

The development of text mining and information retrieval approaches provides an attractive alternative for large scale textual analysis. Current text mining and information retrieval studies, nonetheless, are silent on supporting the analysis of risk-related information in textual data for business decision making. Studies on opinion mining, an important subfield of text mining, have identified important linguistic cues that signal subjectivity and sentiment in statements (Wiebe 2000). Machine learning approaches and lexicon-based approaches have been proposed to measure sentiment in textual data (Pang *et al.* 2002; Tetlock *et al.* 2008). Current opinion mining and sentiment analysis studies neglect the characteristics such as uncertainty, ambiguity, and estimation that may directly signal the existence of risk-related information in textual data. Moreover, general information retrieval (IR) tools (Storey *et al.* 2008) focus mostly on the topical information in documents and may not be an effective approach for recognizing risk-related information across different topics.

The importance of risk-related information and the inadequacy of existing IT tools motivate this study. Informed by theories for decision making under uncertainty, text-based risk measures were developed to signal risk-related information. These measures tap into the core inputs for decision making under uncertainty and can be used to filter and analyze large numbers of business documents based on their contributions to the decision making process. The proposed measures not only can support decision making practice but also can benefit research associated with risk-related information.

I operationalized my proposed risk analysis approach by developing the AZRisk (risk from A to Z) design framework. At the core of the AZRisk framework is the state-of-art machine learning approaches and comprehensive feature representations designed to perform classification tasks derived from my text-based risk measures. My research contributes in several important dimensions to the design framework for enterprise risk recognition, including: text-based risk measures for enterprise risk recognition, and the associated feature representing and analytical techniques.

The remainder of this paper is organized as follows. I review relevant decision making theories, present my text-based risk measures, and summarize related opinion mining studies in Section 4.1. Section 4.2 presents the design framework for text-based risk recognition and my research hypotheses. Section 4.3 presents the research hypotheses. Section 4.4 summarizes the experimental results. I conclude my discussions and future research directions in Section 4.5.

4.1 Text-Based Risk Recognition

My study focuses on developing novel text-based risk recognition approaches to signal risk-related information. As depicted in Figure 4.1, decision makers are faced with large numbers of documents that may support their decision making process. However, only a portion of the documents may contain risk-related information that contributes directly to this process. Grounded in decision making theories, three text-based risk measures have been proposed to signal risk-related information embedded in business documents. I first summarize relevant theories and present the proposed text-based risk measures. Related opinion mining studies are then reviewed to explicate the technical aspects of the underlying problem.

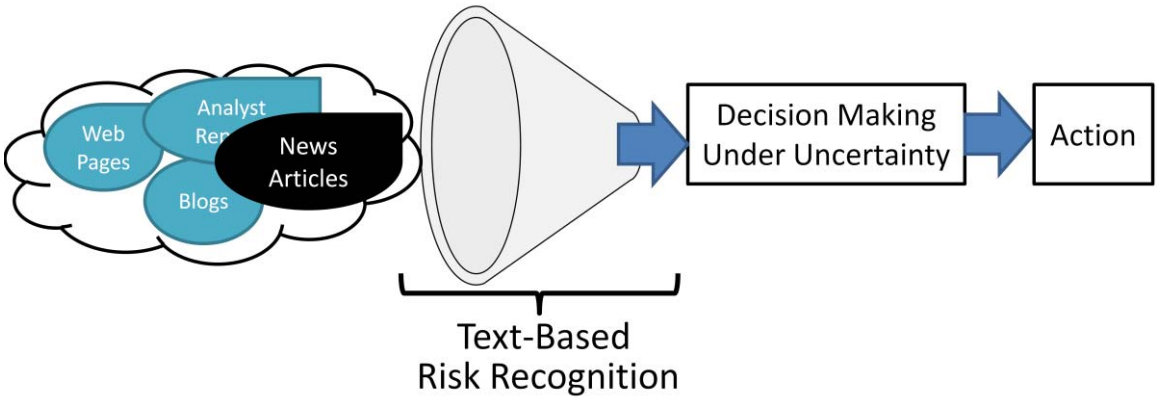


Figure 4.1 Text-Based Risk Recognition

4.1.1 Decision Making Under Uncertainty

Various theories have been developed to explain and guide decision making under uncertainty. My discussion focuses on two aspects that are directly relevant to my

research. The first aspect is the formation and representation of the decision making under uncertainty problem. In other words, I focus on the inputs these theories assumed. The second aspect is strategies used to obtain the inputs essential for choice under uncertainty, especially when quantitative information is not available.

The expected utility theory (Von Neumann & Morgenstern 1944), which is widely used in economic modeling, assumes three types of decision inputs: the possible outcomes, probability distribution over possible outcomes, and the preference of the decision maker. The preference is represented by a utility function that takes an outcome as the input and produces a real number called the utility level. Outcomes associated with larger utility levels are preferred over those with smaller utility levels.

Without loss of generality, assume that a decision maker is facing two options and the consequences of the two options can be described by two distributions over the n possible outcomes. Let T1 = ($p_{1,1}$, $p_{1,2}$, …, $p_{1,n}$) be the probability distribution of the n outcomes if the decision maker chooses the first option and T2 = ($p_{2,1}$, $p_{2,2}$, …, $p_{2,n}$) represents the second. Using the utility function of the decision maker, I can compute n numbers, u_1, u_2, … u_n, that are associated with the utility level of each outcome should it be realized. The expected utility theory asserts that a decision maker will choose the first option over the second if $\sum_{i=1}^{n} u_i \, p_{1,i} \geq \sum_{i=1}^{n} u_i \, p_{2,i}$.

Later theories generalize the expected utility theory and provide alternative explanations for decision making behavior that cannot be explained by the expected utility theory. One notable example is the prospect theory (Kahneman & Tversky 1979). This theory provides a descriptive model for decision making under uncertainty. The

preference of a decision maker is represented by a value function that represents the gain and loss of possible outcomes compared to a reference point. The value function for losses is convex and relatively steeper compared to the gain side, which is usually concave. As a result, phenomena such as loss aversion can be explained under this framework. The prospect theory also replaces probabilities with decision weights, which reflect the observed real-world decision making process.

While the prospect theory was created to address the shortcomings of the expected utility theory, the inputs to these two theories are almost the same. Both theories take probability distributions over potential outcomes and the preference of the decision maker as the inputs. The difference is how the information is incorporated into the decision making process. In fact, most theories about decision making under uncertainty are related to the expected utility theory and assume similar inputs.

The second important aspect of the decision making under uncertainty problem is the strategy used to derive the decision inputs. Two strategies are commonly used. The first strategy assumes that the probability distributions of potential outcomes are objective information that is commonly observable. This strategy also assumes that the preference of the decision maker is given. These assumptions raise the question about quantifying the probability associated with potential states or outcomes. Probabilities of certain future outcomes may be extremely difficult to determine objectively. For example, unrepeatable events, such as worldwide nuclear war, cannot be determined objectively based on relative frequencies (Plous 1993). The second strategy addresses the problem by allowing decision making based on subjective probabilities. The subjective probability

theory (Savage 1954; Anscombe & Aumann 1963) argues that even if states of the world are not associated with recognizable, objective probabilities, restrictions on preferences among decision makers still imply that they behave as if there was a set of utilities and probabilities assigned the outcomes. Decisions are made by taking expectations over the unobservable utilities and probabilities (Mas-Colell *et al.* 1995). The belief about possible states, together with the utility function, can be extracted by observing the choices made by the decision maker.

I summarize the discussion with several observations. First, subjective probability is generated from the utility function. That is, subjective probability does not exist independently of a decision maker's preference and opinion. There is no reason that I cannot think of subjective probability as one kind of opinion. Second, both strategies discussed above are commonly used in economic modeling depending on the problem on hand. As a result, probability can be regarded as either objective or subjective information. Decision makers' preferences, on the other hand, are completely subjective. In order to make decisions, the potential outcomes, which can be considered objective information, need to be evaluated subjectively and converted to utility levels during the decision making process.

4.1.2 A Conceptual Model for Risk Recognition

To recognize risk-related statements that contribute directly to the inputs for decision making, I proposed three text-based risk measures based on decision theories. As shown in Figure 4.2, potential outcomes, probability distribution over potential outcomes, and a decision maker's preferences are three main inputs to decision making

under uncertainty. In the context of business decision making, potential outcomes are often associated with a firm's cash flow over time. Probability distributions describe the uncertainty associated with the cash flow. Preference is used to evaluate available information and decide on the actions.

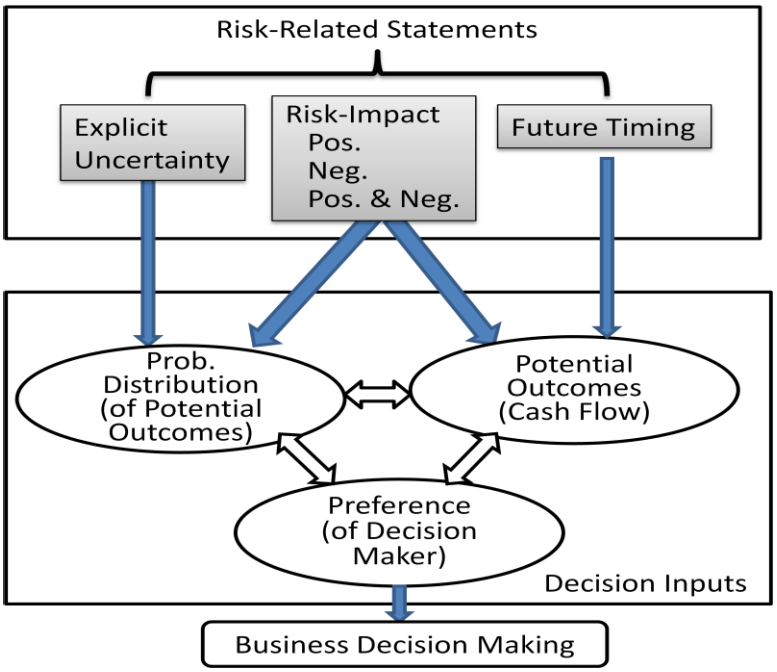


Figure 4.2 A Conceptual Model of Risk-Related Statements and Decision Inputs

Business documents contribute to the decision making process mainly through the refinement of potential outcomes and their probability distributions. I define a statement to be risk-related if it can provide relevant information for business decision making under uncertainty. While preference may also be influenced by business documents, in my discussion it is assumed to be fixed. The level of risk aversion, as a result, is is assumed to be fixed and will not be affected by new information.

The above discussion suggests that risk-related statements may contribute to the decision making process by informing potential outcomes, the risks associated with the outcomes, or both. I propose three text-based risk measures to capture the contributions through these channels: future timing, explicit uncertainty, and risk impact. A summary of these measures can be found in Table 4.1. These measures are presented in sequence below.

Table 4.1 Text-Based Measures for Risk-Related Information

Measure	Definition
Future Timing	Primary content is about future events or states
Explicit Uncertainty	Explicit accounts of doubt or unreliability toward reported information
Risk Impact	Affect decision maker's belief about a firm's future cash flow

Future timing is designed to signal future events or states in textual data. Future events or states are often associated with potential outcomes a decision maker may need to consider. Some of these events or states may be linked to a firm's future cash flow. The future timing measure is broader than the potential outcomes in the conceptual model because certain future events or states may not have implications for business decision making. On the other hand, not every potential outcome will be mentioned in business documents. Moreover, past and ongoing events may also hint at possible future developments. The overlap between future timing and potential outcomes, as a result, is not perfect. This weakness, nonetheless, comes with the advantage of a simple and clear definition that allows effective judgment.

Explicit uncertainty refers to explicit accounts of doubt or unreliability toward reported information. Explicit mentioning of doubt or unreliability often conveys information regarding the probability distribution over potential events or states. These expressions seldom contain quantitative information. Instead, a decision maker may incorporate qualitative information into the risk assessment of potential outcomes by evaluating and weighting its contribution to current belief. Studying the mechanism behind this process is beyond the scope of my study. This measure aims at signaling statements that may contribute to the risk assessment of potential outcomes.

I define that a statement has a risk impact if it contains information that affects a decision maker's belief about a firm's future cash flow. The goal is to summarize potential impact directions that can be rationally inferred from a statement. As opposed to future timing and explicit uncertainty that focus on surface quality, risk impact intends to estimate the directions of aggregated impact based on a reasonable level of world knowledge. The impact direction could be positive, negative, or both. Having both positive and negative impact simultaneously could be caused by the complexity or ambiguity of an event. For example, a business decision may have a positive short term impact but a long term negative impact. Aggregating across time results in an impact of both directions.

Refining the proposed measures is possible. For example, in addition to signaling whether the primary content is about future events and states, the future timing measure can be extended to extract a list of future events and states as well as their expected timing. Explicit uncertainty can be extended to extract entities associated with the

expressed uncertainty. Risk impact can ideally be characterized by the timing, direction, and size of impact. Expected future cash flow can be depicted by a time series starting from now and continuing into the future. Risk impact is the change in the expected cash flow time series caused by a risk-related statement. These refinements, nonetheless, are associated with complicated assessments and may be difficult to answer consistently even by experts. My subsequent analysis focuses on these three proposed measures and leaves the refinements to future research.

4.1.3 Previous Related Opinion Mining Studies

As discussed above, text-based risk recognition aims at recognizing statements that may directly impact a decision maker's belief about a firm's future cash flow. Since the impact is not directly observable, it can only be inferred through interpreting a given statement. Text-based risk recognition, to a certain degree, is similar to the central problem of opinion mining: identifying the private state in each sentence (Wiebe *et al.* 2005). Private state is not open to objective observation or verification (Quirk *et al.* 1985) and can only be inferred through the interpretation of a reader. The major difference between text-based risk recognition and opinion mining is the focus of the underlying entity. Text-based risk recognition focuses on the impact to a business entity. Opinion mining, on the other hand, focuses on the source of the private state. Moreover, opinion mining often focuses on subjective statements while text-based risk recognition focuses on both subjective and objective statements.

For the purpose of providing a technical foundation to the text-based risk recognition problem, the similarities are more important than the differences. I

summarize opinion mining studies by research topics, features, and techniques in sequence.

4.1.3.1 Research Topics in Opinion Mining Studies

Opinion mining covers three major topics: subjectivity identification, sentiment analysis, and certainty identification. Subjectivity identification aims at distinguishing whether information is presented as fact or opinion (Bruce & Wiebe 1999). Separating fact from opinion is valuable for information retrieval systems designed to provide objective information (Wiebe *et al.* 2001). Sentiment analysis aims at identifying positive and negative opinions, emotions, and evaluations (Wiebe *et al.* 2005). The extracted information may be valuable for business intelligence applications and recommender systems (Pang *et al.* 2002). The goal of certainty identification is to identify the level of certainty, together with other relevant information such as the holder of the underlying perspective and the timing associated with expressed certainty (Rubin *et al.* 2005).

To the best of my knowledge, text-based risk recognition has not yet been investigated in previous opinion mining studies. Certainty identification is associated with the proposed explicit uncertainty measure and sentiment analysis may hint on the directions of risk impact. Previous studies, nonetheless, do not focus on the role of textual data for business decision making under uncertainty.

4.1.3.2 Features

For the purpose of constructing automatic processes for opinion mining, various features are used to represent the underlying document, sentence, or expression as a numerical

vector. For example, the conventional bag-of-word (or "unigram") representation converts a document to a long vector of word frequency (see, e.g., (Zhang & Oles 2001)). Each element in the vector corresponds to a unique word in the document. Unigrams and extensions (bigrams and trigrams; the co-occurrence of two or three words) are commonly used in opinion mining (Pang *et al.* 2002; Abbasi *et al.* 2008). Syntactic features such as part-of-speech (POS) tags were also shown to be useful (Yu & Hatzivassiloglou 2003). Stylistic features such as the existence of special characters (e.g., @#$%^), average word length, and vocabulary richness have been found to be useful for sentiment analysis for text from forums (Abbasi *et al.* 2008).

The semantic aspect of given textual data is often captured by lexicons, which are organized according to the meanings of words. A commonly used lexicon is the General Inquirer lexicon (GI; cf. http://www.wjh.harvard.edu/~inquirer/), which contains 182 categories that are designed for content analysis. For example, the positive category contains 1915 words of positive outlook and the negative category contains 2291 words of negative outlook. These semantic categories have been used in previous text mining research and are able to capture important information from textual data (see, e.g., (Tetlock 2007; Pang & Lee 2008)). Other features such as the link structure between web pages are also valuable for text classification (Efron 2004).

4.1.3.3 Techniques

Techniques adopted in opinion mining can be roughly divided into two types. The first type of technique uses a semi-automatic approach to identify linguistic cues that may be associated with subjectivity or polarity. The basic idea is to manually assign scores to

a small set of seeds and extend the scores using certain structures. It has been shown to be useful in the context of identifying linguistic cues for subjectivity (Wiebe 2000) and polarity (Turney 2002).

The other type of technique adopts statistical machine learning models to construct IT artifacts that can automatically identify subjectivity or polarity. Statistical machine learning approaches such as naïve Bayes classifier, support vector machines (SVM), and maximum entropy classifier, have been adopted in previous studies. Naïve Bayes classifier is commonly used in sentiment and subjectivity classification. However, consistent with other text classification studies (Dumais *et al.* 1998), the performance is not as good as that of the other classifiers (Pang *et al.* 2002).

SVM with a linear kernel has been shown to consistently deliver good performance across different text classification tasks (Joachims 1998; Pang *et al.* 2002). Given a set of training examples that can be classified into two classes, SVM searches for a decision hyper-plane that maximizes the margin between the two classes in the transformed feature space (Scholkopf *et al.* 1999). It should be noted that only a subset of training examples will affect the decision hyper-plane because of the margin maximization characteristic. Previous opinion mining studies also found that the performance may increase if the input features are pre-processed using various feature selection techniques (Abbasi *et al.* 2008).

Maximum entropy classifier has also been considered in opinion mining study (Pang *et al.* 2002). The maximum entropy model is mathematically equivalent to logistic regression, which has been widely used to model the discrete choice problem (Ben-Akiva

& Lerman 1985). While logistic regressions are a powerful modeling approach, applying them under the opinion mining context may encounter the curse of dimensionality problem. Recall that before applying statistical machine learning approaches, each sentence (or document) needs to be converted to a long vector according to the features considered. It is common in text classification problems that the length of the vector is longer than the total number of training examples, which leads to serious over-fitting during the learning (model estimation) phrase. Early stopping (i.e., a fixed run of numerical optimization) was used in the previous study to overcome the problem (Pang *et al.* 2002).

4.1.4 Challenges in Text-Based Risk Recognition

While a wide range of topics have been investigated in opinion mining studies, previous studies are silent on the design framework that can generate IT artifacts for text-based risk recognition. Various prototype systems have been built to evaluate the effectiveness of opinion mining approaches. A notable example is the OpinionFinder system (Wilson *et al.* 2005), which provides automatic subjectivity identification at the sentence level and sentiment polarity classification at the expression level (cf. http://www.cs.pitt.edu/mpqa/opinionfinderrelease/). Studies on certainty identification, nonetheless, only documented preliminary annotation results and did not provide guidelines for IT artifact development. It is my intent to provide a design framework for text-based risk recognition and to evaluate and validate the framework.

4.2 AZRisk (Risk from A to Z): A Design Framework for Text-Based Risk Recognition

The importance of text-based risk recognition and the insufficiency of existing opinion mining systems warrant the development of a novel design framework to help inform future system development. Through the development and evaluation of the proposed design framework, I intend to answer the following research questions:

- How do I develop a design framework for text-based risk recognition?
- How does text-based risk recognition differ from opinion mining and information retrieval problems?
- What is the most suitable approach for text-based risk recognition?
- What are useful features for text-based risk recognition?

Guided by the conceptual framework presented in Figure 4.2 and previous opinion mining studies, I propose the AZRisk design framework for text-based risk recognition. As presented in Figure 4.3, the AZRisk design framework aims at recognizing risk-related information embedded in individual sentences by operationalizing the proposed text-based risk measures. The AZRisk design framework involves three major phases: annotation, learning, and production. In the annotation phase, the three proposed text-based risk measures are formulated as binary text classification tasks, which facilitate the creation of reference standard datasets. The learning phase is a search process which identifies suitable feature representation and statistical machine learning approaches. After the best model has been identified, it enters the production phase, which provides predicted classification results for each input sentence. In the following discussion, I present the three major phases and discuss important components involved in each phase.

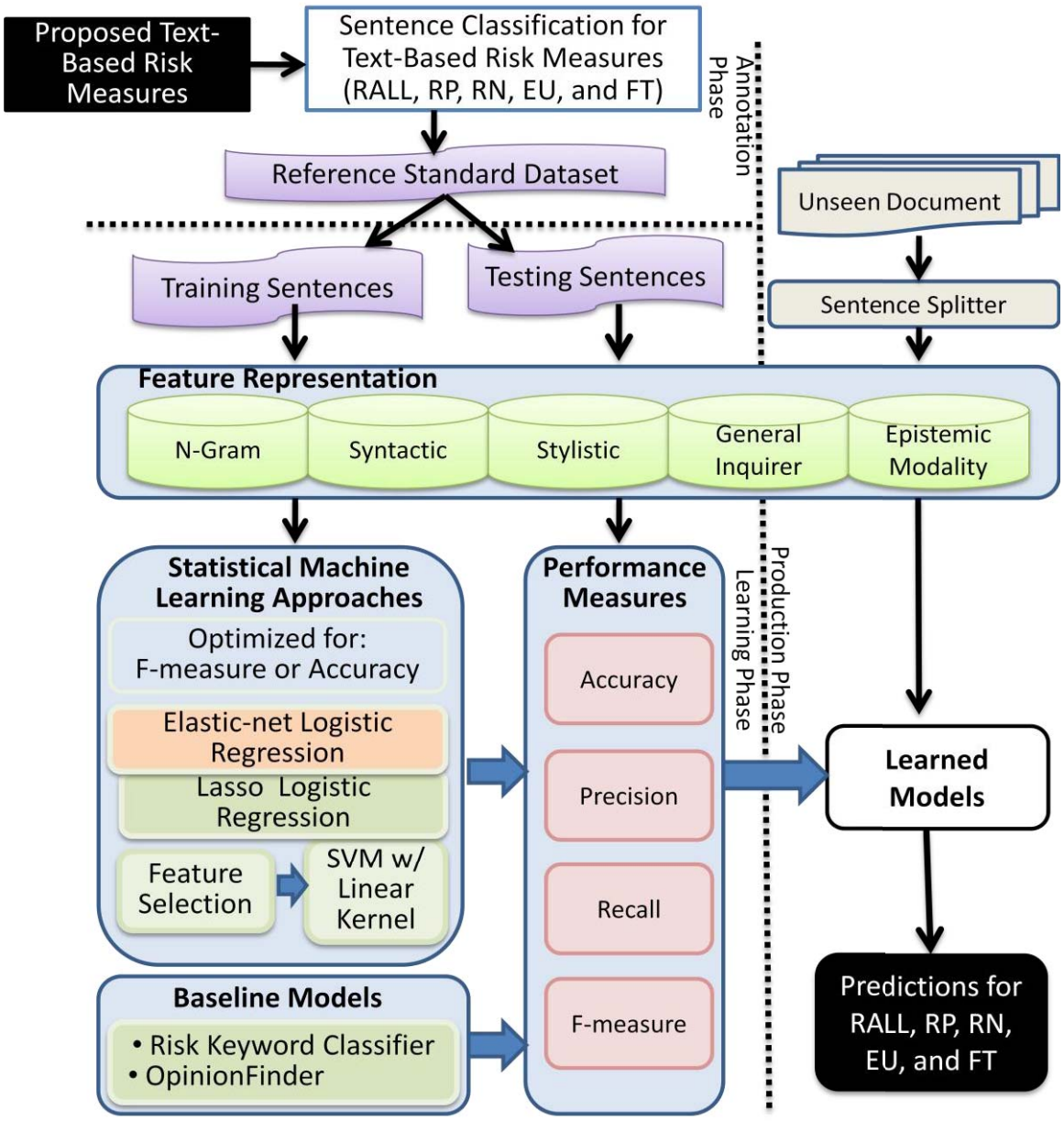


Figure 4.3 AZRisk: A Design Framework for Text-Based Risk Recognition

4.2.1 Annotation Phase

The annotation phase bridges the conceptual model in Figure 4.2 and the learning phase of AZRisk by converting the three proposed text-based risk measures into binary text

classification problems at the sentence level. Following previous opinion mining studies (Bruce & Wiebe 1999; Wiebe *et al.* 2001), I choose to conduct my analysis at sentence level. I consider document level analysis too coarse and it may also introduce unnecessary noise that would adversely affect performance. Word level analysis is associated with higher annotation costs and longer completion time that may not justify the additional benefit.

4.2.1.1 Sentence Classification for Text-Based Risk Measures

Table 4.2 summarizes the mapping between risk measures and three text classification tasks. Future timing is designed to signal the existence of future events or states. The FT classifier achieves this goal by identifying sentences with primary content about future events or states.

Table 4.2. Sentence Classification Tasks for Text-Based Measures.

Risk Measure	Classification Task (Binary Decision)	Definition
Future Timing	FT	Whether the primary content of this sentence is about future events or states.
Explicit Uncertainty	EU	Whether this sentence contains explicit accounts of doubt or unreliability toward reported information.
Risk Impact	RP	Whether this sentence positively affects the decision maker's belief about a firm's future cash flow.
	RN	Whether this sentence negatively affects the decision maker's belief about a firm's future cash flow.
	RALL	Whether this sentence positively or negatively affects the decision maker's belief about a firm's future cash flow.

The explicit uncertainty measure aims at identifying sentences with explicit accounts of doubt or unreliability toward the reported information. Previous studies on certainty identification divided the certainty-uncertainty spectrum into four categories and found that human experts tend to disagree on which level of certainty the expression belongs to (Rubin 2007). Rubin and Liddy (Rubin *et al.* 2005) thus suggest a binary classification scheme for certainty/uncertainty recognition. Following their suggestion, I define the EU classifier to be dichotomous. It assigns a sentence to the positive class if it contains explicit accounts of doubt or unreliability toward reported information.

Risk impact captures information affecting decision makers' beliefs over a firm's future cash flow. Four values are possible: positive, negative, positive and negative, and none. Two classifiers, RP and RN, were used to capture the positive and negative impact directions, respectively. These two classifiers operate independently. Combining the output from these two classifiers gives us the four possible values of the risk impact measures.

In addition to RP and RN, I introduce another classifier, RALL, which captures risk impact regardless of the impact directions. This classifier provides a means to discriminate among sentences according to the existence of risk impact, which can be used to filter out sentences with risk impact in subsequent analysis. It should be noted that other combinations of classifiers can be used to achieve the same outcomes. For instance, I can first apply an RALL classifier and then define another pair of directional classifiers that operate on the positive cases identified by RALL. I leave the exploration of the alternative combinations to future research.

To provide concrete examples, consider the sentences listed in Table 4.3. The first sentence describes an on-going plan about adjusting Hitachi's operations. While the final outcome remains unknown, it is clear that this on-going plan may impact Hitachi's future cash flow. It is thus reasonable to assign it to RP, RN, and RALL, indicting possible impact in both directions. This sentence, which appeared in a WSJ article on Feb. 5, 2004, was mainly about an ongoing plan expected to be finished by 2006. It thus contains references to future timing and explicit expressions of uncertainty toward the information ("it planned to withdraw").

Table 4.3 Examples of Risk-Related and Non Risk-Related Sentences

No.	Sentence	RALL	RP	RN	FT	EU
1	Faced with an industry downturn and criticism of its sprawling business structure, Hitachi, which makes everything from tiny mobile phones to bulky construction machinery, said last year that by March 2006 it planned to withdraw from unprofitable, noncore operations accounting for about 20% of its more than eight trillion yen in annual group sales (Wall Street Journal, Feb. 5, 2004).	√	√	√	√	√
2	While many analysts had predicted the market for ICDs would grow about 20% a year due to an aging population, many now forecast only single-digit percentage growth for the year (Wall Street Journal, Oct. 19, 2006).	√		√	√	√
3	Many personal-computer applications send hundreds, even thousands of messages back and forth before completing a task such as transferring a file (Wall Street Journal, Apr. 27, 2004).					

The second sentence contains information about decreasing future market shares, which would clearly have a negative impact on a decision maker's assessment of the firm's future cash flow. It thus belongs to RN, RALL, and FT. The expression ("forecast"

and "predicted") in the sentence explicitly conveys the uncertainty toward the reported information. The last sentence describes the operation of computer applications and does not contain information directly associated with a firm's future cash flow. It is not about future events or states; nor does it convey uncertainty toward reported information.

4.2.1.2 Reference Standard Dataset

A wide range of business documents can be adopted to create the reference standard dataset. I chose to sample news articles from the *Wall Street Journal* (WSJ) for the following reasons. First, WSJ is a highly-circulated newspaper with a business focus (cf. Audit Bureau of Circulation, http://www.accessabc.com/). WSJ reaches a broad range of audiences and can be thought of as commonly accepted business documents. Second, WSJ covers many different companies in various industries, which fits my goal of constructing and verifying a general-purpose text-based risk recognition framework.

My research testbed was created from a random sample of firm-specific WSJ news articles published between 8/4/1999 and 3/2/2007. The selected articles were sorted in a random order and then split into sentences. The original publication date and a document ID were attached to each sentence. Articles with more than 30 sentences were truncated to avoid the dominance of long documents. An Excel file that contained the sentences, publication dates, and document IDs was created. The original order of the sentence in an article was preserved. My testbed contains 2,539 sentences from 164 firm-specific news articles. Previous text mining studies typically have a sample size of around one thousand (for example: 1,000 messages (Abbasi *et al.* 2008), 1,140 messages (Wiebe *et al.* 2001),

504 sentences (Wiebe *et al.* 1999), and 685 sentence (Rubin 2004)). A sample size of 2,539 sentences is comparable to previous related studies.

I divided the 2,539 sentences into three sets while preserving the article boundaries. The development set contains the first 521 sentences. The testing set contains the following 491 sentences. The production set contains the remaining 1,527 sentences. The development set was used to develop the coding manual, which was applied to the testing set by two annotators independently. The sentences in the production set were then coded following the coding manual.

Table 4.4 reports agreement and Cohen's kappa computed based on the testing set. Agreement, which is the proportion that the tagging from two coders matches, is higher than 89% for all tasks considered. Moreover, all tasks have Cohen's kappa (chance-corrected agreement) above 0.75, which is considered to represent excellent agreement beyond chance (Fleiss 1981). Note that the agreement can be interpreted as the accuracy upper bound for the underlying classification tasks. That is, if I consider the annotation from one annotator as the correct answer, then the accuracy of the other annotation is the agreement, and vice-versa. Based on the observation, the accuracy upper bound is 0.89, 0.90, and 0.91 for RN, RP, and RALL, respectively. FT and EU have higher upper bounds (0.95 and 0.94) compared to the other tasks considered.

Table 4.4. Inter-Rater Agreement

Task	Agreement*	Cohen's Kappa*
FT	0.95 (0.92, 0.96)	0.85 (0.79, 0.90)
EU	0.94 (0.91, 0.96)	0.81 (0.75, 0.88)
RP	0.90 (0.88, 0.93)	0.79 (0.73, 0.85)
RN	0.89 (0.86, 0.91)	0.77 (0.71, 0.82)
RALL	0.91 (0.88, 0.93)	0.81 (0.76, 0.86)

*The 95% confidences intervals (reported in the parentheses) were computed

based on 5,000 bootstrappings.

My reference standard dataset was generated by merging the annotation results from the development, testing, and production sets. Discrepancies were resolved based on consensus. Table 4.5 summarizes the prevalence of the annotation tasks considered. Forty-six percent of the sentences are annotated as RALL, which indicates that about half of the sentences in my sample are risk-related. If I look at the direction of impact, 33% of the sentences contain favorable interpretations of the future while 36% of the sentences contain unfavorable interpretations. The difference between RP and RN is significant at 95% confidence interval. One possible reason is that the WSJ prefers unfavorable news to favorable news.

Table 4.5. Prevalence of Risk-Related Sentences.

Task	Count	Percent*
FT	642	25% (24%, 27%)
EU	634	25% (23%, 27%)
RP	836	33% (31%, 35%)
RN	904	36% (34%, 38%)
RALL	1157	46% (44%, 47%)

*The values in the parentheses report 95% confidence intervals.

Both EU and FT have a prevalence of 25%, which means that most sentences in firm-specific news do not contain explicit uncertainty expression, and that the majority of sentences in firm-specific news are not directly related to future events or states. Note that the prevalence of RALL is much higher than that of FT and EU. It indicates that sentences may not be directly related to future events or states but are still tagged as risk-

related. Similarly, sentences without explicit uncertainty accounts may also be tagged as risk-related sentences.

4.2.2 Learning Phase

The learning phase involves the design of feature representation, the selection of statistical machine learning approaches and baseline models, and performance comparisons (see Figure 4.3). Using ten-fold cross validation (Dietterich 1998), the performance of selected statistical machine learning approaches and baseline models were compared in order to verify the effectiveness of the design framework and gain insights into the risk recognition problem. Hyperparameters of statistical machine learning models were tuned by further splitting a training dataset and conducting grid search. Decision thresholds were determined in a similar manner. I discuss important components involved in the learning phase in sequence.

4.2.2.1 Feature Representation

For the purpose of machine learning, each sentence needs to be converted to a numerical representation. I designed the feature representation based on previous studies and the characteristic of the recognition tasks at hand. Four types of features including n-grams, syntactic features, stylistics features, and semantic features were selected.

N-grams are one of the most commonly used features in text classification. The de-facto bag-of-word (unigram; i.e., "1-gram") representation converts a sentence into a long vector of 0s and 1s that represents the existence of a list of unique words (Forman 2003). The collocation of two words (bi-gram) and three words (tri-gram) are also commonly

used (Abbasi *et al.* 2008). Before converting words to N-gram, a morphological analyzer based on WordNet (cf. http://wordnet.princeton.edu/) was used to find the base form of verbs and nouns; all numbers were replaced with "NUM."

Part-of-speech (POS) tags of each word (e.g. adjective, adverb, plural noun) reflect the syntactic aspect of a sentence (Wiebe *et al.* 1999). Syntactic features provide additional information about the usage of each word in the context of a given sentence.

Stylistic features have been found to be useful in recognizing sentiments among user-generated documents (Abbasi & Chen 2008) as well as discriminating deception in on-line communication (Zhou *et al.* 2004). This empirical evidence suggests that stylistic features may capture intentional or unintentional changes in writing style caused by the changes in mental status. I included popular stylistic features such as average sentence length (as measured by the number of words), average word length (as measured by the number of characters), and redundancy (as measured by the portion of function words in a sentence). Features that appear in only one (or two) sentences was replaced with "HAPAX" (or "DIS") to indicate the existence of low frequency features.

As suggested by previous certainty identification studies (Rubin 2007), linguistic devices such as epistemic modality (Coates 1987; Nuyts 2001), evidentiality (Mushin 2001), and hedging (Hyland 1998) may play an important role in conveying the evaluation of chance, the degree of reliability of expressed information, and the lack of commitment to the truth value of an accompanying proposition. These linguistic devices are often associated with function words and high frequency words that are removed in text classification studies (Manning & Schuze 1999). In order to fully capture the effect

of these linguistic cues, stop words were not removed during the process of generating feature representation. Furthermore, I enhanced the feature representation by adding features directly associated with uncertainty and risks. Specifically, an epistematic modality (EPI) lexicon with nine categories was included (Rizomilioti 2006). EPI categories such as epistematic verb (e.g., suspect, seem) and epistematic adjectives (e.g., suggestive, seemly) provide more detailed information than those captured by POS tags and can potentially help the recognition tasks. I also included a commonly used lexicon, General Inquirer (GI; cf. http://www.wjh. harvard.edu/~inquirer/), in my feature representation to capture the semantic aspects of sentences (Tetlock 2007).

I adopted a frequency cutoff of two when converting a sentence to the underlying representation. If a feature appears in two or less sentences in the whole testbed, then the feature is dropped. The cutoff of two is considered a conservative and practical choice (Forman 2003).

4.2.2.2 Statistical Machine Learning Approaches

Before discussing the statistical machine learning approaches adopted in this research, it should be noted that all classification tasks are binary. In the subsequent discussion, a positive case refers to a sentence that possesses the underlying characteristic. For example, a sentence containing favorable (unfavorable) information about future cash flow is a positive case for a RP (RN) classifier.

Two types of machine learning approaches, support vector machines (SVM) and regularized logistic regression models, are considered in this study. Previous studies found SVM an attractive option for opinion mining (Pang *et al.* 2002; Abbasi *et al.* 2008).

While direct application of logistic regression may encounter the curse of dimensionality problem, recent studies in text classification have shown that regularized logistic regression can address the curse of dimensionality problem and achieve performance higher than or comparable to SVM (Figueiredo & Jain 2001; Zhang & Oles 2001). Among various regularized logistic regressions, elastic-net (Zou & Hastie 2005; Friedman *et al.* 2009) and lasso (Tibshirani 1996) logistic regressions perform feature selection and model learning in one, integrated step. Elastic-net and lasso logistic regressions exclude features that do not contribute to the underlying classification tasks during model training. Only features that are considered useful during the learning process are included. The resulting models are usually parsimonious and effective. I briefly introduce elastic-net logistic regression, lasso logistic regression, and SVM.

The dependent variable Y_i of elastic-net logistic regressions (ENET) is either 1 (positive case) or 0 (negative case). The independent variable X_i is a long vector representing the sentence according to the feature representation. Similar to logistic regression, the probability that a given sentence is a positive case can be written as $\text{Prob}(Y_i = 1|X_i; B) = \frac{1}{1+e^{-X_i B}} \equiv P(X_i; B)$, where *B* is the coefficient vector that needs to be learned (or estimated) from the training examples.

If the length of B is larger than the number of training examples, then the likelihood function is under-determined. The likelihood function will increase without bound during the model learning. ENET addresses the problem by adding additional regularization terms that penalize growing coefficients. Specifically, given N training examples, ENET learns the coefficient vector B by maximizing the following objective function:

$$\max_B \frac{1}{N} \sum_{i=1}^{N} [Y_i \log P(X_i; B) + (1 - Y_i) \log (1 - P(X_i; B))] - \lambda [\alpha \|B\|_{L1} + (1 - \alpha) \|B\|_{L2}^2] \quad (1)$$

where $\|B\|_{L1} = \sum_{j=0}^{k} |B_j|$ (i.e., L1 norm) and $\|B\|_{L2}^2 = \sum_{j=0}^{k} B_j^2$ (i.e., L2 norm). The first half of the objective function, $\frac{1}{N} \sum_{i=1}^{N} [Y_i \log P(X_i; B) + (1 - Y_i) \log (1 - P(X_i; B))]$, is the likelihood function of a corresponding logistic regression divided by the number of observations. The second half of the objective function, $[\alpha \|B\|_{L1} + (1 - \alpha) \|B\|_{L2}^2]$, is the regularization terms for ENET, which linearly combine L1 and L2 norms. The parameter α is a real number between 0 and 1 that controls the mix between these two types of regularization. L1 regularization corresponds to the term $\|B\|_{L1}$, which can be interpreted as imposing a Laplace prior with mean zero on the coefficient vector B. L2 regularization corresponds to the term $\|B\|_{L2}^2$, which can be interpreted as imposing a Gaussian prior with mean 0 on B.

The intuitive interpretation is that the regularization terms of ENET tell the model that if no additional information is provided, the coefficients are zero. The parameter α determines the shape of the prior distribution that carries this information. The parameter λ determines how aggressive this "zero tendency" is carried out during the training process. Higher λ corresponds to a more aggressive selection of features.

After the training process is completed, the learned coefficient $\widehat{B}$ is obtained. The probability that an unseen sentence z is a positive case can be determined by computing $P(X_z; \widehat{B})$. Binary classification can be obtained by applying a threshold θ to $P(X_z; \widehat{B})$. That is, a sentence is predicted to be a positive case if $P(X_z; \widehat{B}) \geq \theta$. In this study, I

consider ENET with $\alpha = 0.25$, 0.5, and 0.75. The three ENET models are denoted as ENET25, ENET50, and ENET75, respectively.

Lasso logistic regression (LASSO) can be considered as a special case of ENET with $\alpha = 1$. LASSO is similar to ENET in the sense that both models only select a subset of features to construct their models. A major difference is how features are selected during the training phase. Given a group of features with high pairwise correlations, LASSO tends to select only one feature from the group and leave the others out (Zou & Hastie 2005). ENET, on the other hand, promotes a grouping effect and tends to select variables in the group simultaneously.

I also consider the support vector machine (SVM) classifier with a linear kernel (Vapnik 1995; Joachims 1999) in this study. Previous studies have shown that SVM has achieved good performance on text classification problems (Dumais *et al.* 1998; Joachims 1998). A salient characteristic of the SVM model is that it maximizes the margin between two classes when choosing the decision hyperplane. A SVM model can be written as:

$$\min_{S,S_0} \frac{1}{2}\|S\|_{L2}^2 + C\sum_{i=1}^{N}\xi_i \qquad (2)$$

$$\text{subject to } \xi_i > 0, Y_i(X_i S + S_0) \geq 1 - \xi_i \quad \forall i$$

where S_0 and S are the intercept and slopes of the decision hyperplane. A slightly different representation for dependent variable is used: $Y_i = 1$ or -1 for a positive or a negative case. The constraints say that given the decision hyperplane, each training example needs reside at the correct side of the hyperplane and the distance to the hyperplane needs to be at least $1/\|S\|_{L2}$. If the condition is not satisfied, a penalty of ξ_i

will be charged. The training process searches for the hyperplane that minimizes the total penalty of training examples and the L2 norm of the normal vector S. Note that minimizing the L2 norm of the normal vector S is equivalent to maximizing the margin between two classes, which is $2/\|S\|_{L2}$. I used a popular algorithm, SVM-Light (http://svmlight.joachims.org/) (Joachims 1999), to solve the optimization problem.

4.2.2.3 Feature Selection

Feature selection is a common technique to reduce the number of features used in statistical machine learning approaches. ENET and LASSO have built-in mechanisms to handle large numbers of features. As a result, I did not apply any feature selection techniques in conjunction with these two approaches.

Previous studies have reported increased performance by combining SVM with feature selection (Abbasi *et al.* 2008). Feature selection techniques can be roughly divided into two types: filter and wrapper (Li *et al.* 2007). The first type of technique evaluates individual features and selects those that are considered the best. This type of feature selection technique is often referred to as the filter. The second type of technique, the wrapper, evaluates a subset of features by invoking a learning algorithm and computing the classification accuracy associated with the subset of features. A wrapper may achieve a higher performance level, but at a higher computational cost. In addition, the results of feature selection may have lower generality (Li *et al.* 2007). I adopted one of the popular filter techniques, information gain (IG), to preselect features for SVM.

4.2.2.4 Baseline Models

The baseline models provide reference points from which the performance of various statistical machine learning approaches can be interpreted. I considered two baseline models. The first baseline model, Risk Keywords Classifer (RKC), adopted a simple keyword matching strategy to perform various classification tasks considered. If one or more keywords appear in a sentence, then this classifier assigns it as a positive case. This approach is applied to RALL, RP, RN, EU, and FT. I collected risk keywords from the indexed terms of a book that covers topics about risk management (Young & Tippins 2001). This approach treats the classification tasks as one kind of information retrieval problem. The performance of RKC relative to other approaches can be interpreted as the effectiveness of applying information retrieval techniques to the risk recognition problem.

The second baseline model, OF classifier, recognizes risk-related sentences based on the output of a state-of-the-art opinion mining tool, OpinionFinder (Riloff & Wiebe 2003). This baseline was included to investigate the effectiveness of current opinion mining techniques for the risk recognition problem at hand. OF classifies a sentence as RALL if the OpinionFinder tags any words or phrases in the sentence as having positive or negative sentiment. A sentence is classified as RP (RN) if the OpinoinFinder tags any words or phrases in the sentence as having positive (negative) sentiment. The three classifiers for RALL, RP, and RN are referred to as OP_PN, OF_P, and OF_N, respectively.

I also constructed two classifiers for FT and EU based on OpinionFinder. I adopted a naïve assumption that all subjective sentences are related to uncertainty and all objective sentences are about future events or states. The OF_O (OF_S) classifier assigns a sentence to the FT (EU) class if it is classified as an objective (subjective) sentence by OpinionFinder.

4.2.2.5 Performance Measures

I measured the quality of the classification results using classical text classification measures, including accuracy, recall, precision, and F-measure (Witten & Frank 2005; Abbasi & Chen 2008). Specifically, let TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative of a classification results. These measures can be computed as follows:

$$\text{Accuracy} = (TP+TN) / (TP+FP+TN+FN)$$

$$\text{Recall} = TP / (TP+FN)$$

$$\text{Precision} = TP / (TP+FP)$$

$$\text{F-measure} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$$

Accuracy is the probability that a case is correctly assigned. Recall is the probability that a positive case is correctly assigned; precision is the probability that an assigned positive case is correct. There is a natural trade-off between recall and precision. That is, if a model is tuned to increase recall, the precision usually decreases. The relative importance between recall and precision varies from scenario to scenario. I argue that in the context of risk-related sentence recognition, recall and precision are equally important. A classification system with high precision but low recall may overlook a large chunk of

documents containing risk-related information. On the other hand, high recall and low precision introduces a higher level of noise and impacts the usefulness of the classification results. One way to balance precision and recall is by computing the F-measure, which is the harmonic mean of precision and recall.

I emphasize accuracy and F-measure when training the statistical machine learning models. The emphasis reflects my preference for high quality classification results.

4.2.3 Production Phase

After the statistical machine learning approaches and baseline models have been evaluated, the most suitable models were selected and recorded for the production phase. For an unseen document, each sentence is converted to a numerical vector according to the feature representation used in the training phase. The learned models then can be used predict classification outcomes. While the learning phase often requires sophisticated numerical optimization procedures, applying learned models are much less computationally intensive and can process large numbers of documents in a short time period. Aggregating and sorting according to the prediction results allows us to differentiate each document according to its implication to decision making under uncertainty. Precious labor hours than can be prioritized according to the classification results.

4.3 Research Hypotheses

Although previous opinion mining studies have touched on some aspects of text-based risk recognition, it exhibits unique characteristics that cannot be fully captured by

existing approaches. By incorporating rich sets of features, statistical machine learning approaches can learn the unique characteristics of text-based risk recognition and generate effective models. I hypothesize that statistical machine learning approaches can successfully incorporate useful features into the learned model and surpass opinion mining tools for text-based risk recognition:

H1: Statistical machine learning approaches outperform baseline models based on OpinionFinder.

While topical information may be critical for traditional information retrieval systems, text-based risk recognition involves judgments beyond topical information. As a result, keyword-matching will not be an effective approach for text-based risk recognition:

H2: Statistical machine learning approaches outperforms RKC, the keyword-matching baseline models.

Among the three statistical machine learning approaches considered, ENET and LASSO often generated parsimonious and interpretable models. Moreover, the grouping effect of ENET can handle the multicollinearity problem that may exist in a text classification dataset. LASSO, on the other hand, is less attractive along this line. Moreover, elastic-net logistic regressions have been shown to achieve higher or equal performance compared to LASSO and SVM (Zou & Hastie 2005). ENET would be an attractive option if its performance were comparable to other approaches. I hypothesize that ENETs performance is equal to or better than other statistical machine learning approaches for my text-based risk recognition tasks.

H3: Elastic-net logistic regressions have equal or better performance as compared to SVMs and LASSO.

4.4 An Experimental Study

I conducted an experimental study to evaluate AZRisk using the reference standard dataset created from WSJ. Subsequent discussions summarize the input features to the statistical machine learning approaches, followed by hypotheses testing results. The most suitable approach then was selected based on the outcomes. I conclude this section with a discussion of important features and their implications.

4.4.1 Input Features

Table 4.6 summarizes the input features to statistical machine learning approaches. Trigram generated the largest number of unique features (47,700), followed by bigram (34,094) and unigram (6,566). Many of the features, nonetheless, appear in only one or two sentences. After applying the frequency threshold (2), the number of features was greatly reduced. Bigram contained the largest number of retained features (2,708), followed by unigram (2,514) and trigram (1,154). All semantic categories in POS and EPI achieved the threshold. All but one semantic category in GI exceeded the given threshold. Examples of different feature types are listed in the last column of Table 4.6. Using my feature representation, each sentence was converted to a numeric vector of length 6,604 before statistical machine learning approaches were applied.

Table 4.6 Summary of Input Features

Feature	Feature Sets	# of Unique	Examples; Definitions

Types		Features; Mean (Std)	
N-Gram	Unigram	2514/6566*	audit, stay, rule
	Bigram	2708/34094*	a_share, attorney_general
	Trigram	1154/47700*	%_decline_in
Syntactic	POS	35/35*	JJ, NNS, VB
Stylistic	Sent. Len.	22.7 (10.7)	# of words per sent.
	Avg. Word Len.	5.1 (0.7)	Average word length
	Redundancy	0.3 (0.1)	# fun. words / # words in sent.
Semantic	General Inquirer (GI)	181/182*	Positive, Negative, Understatement
	Epistemic Mod.	9/9*	Epi_lex_verb, Epi_noun

* # of unique features with sentence frequency>2 / # of unique features

4.4.2 Performance Comparisons: Statistical Machine Learning Approaches and Baseline Models

Table 4.7 summarizes the performance of RALL classification. The second column reports classification performance with classifiers optimized for accuracy during the training phase; the third to the fifth columns report performance with classifiers optimized for F-measure during the training phase.

Table 4.7 Performance of RALL Classification

Model	Accuracy†	F-Measure‡	Recall‡	Precision‡
LASSO	67.1%	66.5%	83.8%	55.1%
ENET75	69.3%	68.0%	87.7%	55.6%
ENET50	68.9%	68.7%	90.5%	55.4%
ENET25	69.4%	68.9%	**91.2%**	55.4%
SVM	**69.5%**	**70.2%**	83.9%	**60.3%**
SVM w/IG	69.1%	68.9%	84.3%	58.3%
RKC	53.0%	30.4%	47.0%	22.5%
OF_PN	54.8%	27.9%	19.1%	51.4%

†Model training was optimized for accuracy
‡Model training was optimized for F-measure

Both baseline models, RKC and OF_PN, performed badly for RALL classification. The accuracy of RKC was 53.0%, similar to that of OF_PN (54.8%). The F-measure of RKC was 30.4%, 2.5% higher than that of OF_PN (27.9%). It is interesting to note that by using keywords, I were able to locate 47.0% of RALL sentences (i.e., recall). This method, nonetheless, generated noisy results – the precision was only 22.5%.

All statistical machine learning approaches performed better than the two baseline models. Among them, the SVM was the best performer, with an accuracy of 69.5% and an F-measure of 70.2%. ENET25 came in second place, with an accuracy of 69.4% and an F-measure of 68.9%. LASSO was the worst performer for RALL. The average accuracy of all machine learning approaches was 68.9%, 15.9% higher than the accuracy of the RKC and 14.1% higher than the accuracy of OF_PN. The average F-measure of statistical machine learning approaches was 68.5%. The F-measure of baseline models was at least 38.13% lower than the average F-measure of statistical machine learning approaches.

I computed the 95% confidence interval and the p-value of the null hypothesis that a given pair of performance was equal. Pairwise comparisons showed that the two baseline models had worse performance compared to the statistical machine learning approaches. All accuracy and F-measure differences between these groups were statistically significant at the 95% confidence level after the Bonferroni correction. The results provide strong evidence supporting Hypotheses H1 and H2 for RALL classification, which asserts the superior performance of the statistical machine learning approaches.

Hypotheses H1 and H2 are also supported for RP, RN, EU, and FT classification. Table 4.8 summarizes the performance of classifying RP, RN, EU, and FT. The results reveal similar patterns. The performance gaps between the baseline models and the statistical machine learning approaches remain. The difference is statistically significant after Bonferroni corrections. The use of simple keyword matching techniques (RKC) for EU classification causes an accuracy loss of 18.0% compared to the average accuracy of machine learning approaches. This gap is smaller (11.8%) if I consider the sentiment-based OF_S model. Performance gaps of similar or larger magnitude can be found across the other three classification tasks. These performance gaps indicate that keyword matching and existing sentiment analysis tools are not effective alternatives for risk recognition.

Table 4.8 Performance of RP, RN, EU, and FT Classification

	RP Classification				RN Classification			
Models	Acc.†	F-Meas.‡	Recall‡	Prec.‡	Acc.†	F-Meas.‡	Recall‡	Prec.‡
Lasso LR	72.8%	56.8%	72.5%	46.7%	68.5%	59.1%	84.2%	45.6%
ENET75	72.7%	57.2%	69.9%	48.5%	69.9%	59.1%	85.4%	45.2%
ENET50	72.6%	56.5%	**72.8%**	46.1%	70.3%	60.0%	**89.8%**	45.0%
ENET25	72.6%	56.9%	71.7%	47.2%	70.4%	60.2%	89.5%	45.3%
SVM	**73.3%**	**59.8%**	**72.8%**	**50.8%**	**71.0%**	**63.1%**	80.4%	**51.9%**
SVM w/IG	73.0%	57.2%	66.6%	50.1%	69.7%	60.7%	77.5%	49.9%
RKC	58.6%	24.5%	30.7%	20.3%	59.2%	29.0%	38.2%	23.3%
OF_P/OF_N	66.9%	14.7%	8.6%	49.3%	62.4%	20.9%	13.9%	42.0%
	EU Classification				FT Classification			
Models	Acc.†	F-Meas.‡	Recall‡	Prec.‡	Acc.†	F-Meas.‡	Recall‡	Prec.‡
Lasso LR	81.1%	61.9%	83.3%	49.3%	86.7%	71.1%	67.9%	74.5%
ENET75	81.2%	61.7%	**84.2%**	48.7%	86.4%	71.3%	67.8%	75.1%
ENET75	81.1%	62.0%	83.1%	49.5%	86.4%	71.0%	66.8%	75.8%
ENET25	81.4%	61.8%	83.4%	49.0%	**87.1%**	71.7%	66.5%	**77.8%**
SVM	81.3%	59.7%	70.2%	51.9%	86.9%	72.1%	72.6%	71.7%
SVM w/IG	**82.6%**	**62.6%**	71.3%	**55.9%**	87.0%	**72.3%**	68.1%	77.1%
RKC	63.5%	22.1%	23.7%	20.7%	62.7%	20.9%	22.6%	19.5%
OF_S/OF_O	69.7%	20.0%	15.1%	29.4%	32.2%	39.7%	**88.2%**	25.6%

†Model training was optimized for accuracy
‡Model training was optimized for F-measure

4.4.3 Performance Comparison: Among Statistical Machine Learning Approaches

The statistical tests showed that, for RALL classification, the pairwise accuracy and F-measure differences among the three ENET models and two SVM models (ENET25, ENET50, ENET75, SVM, and SVM w/IG) were not significant at a 95% confidence level after the Bonferroni correction for multiple comparisons. LASSO, on the other hand, had a significantly worse F-measure compared to all other statistical machine learning approaches. Two out of five accuracy differences between LASSO and other statistical machine learning approaches are significant. The results indicate that ENETs and SVMs have comparable performance while the performance of LASSO is worse than other statistical machine learning approaches. Hypothesis H3, as a result, is supported based on RALL classification results.

For RP, RN, EU, and FT, the accuracy difference among ENET25, ENET50, ENET75, SVM, and SVM w/IG was insignificant after the Bonferroni correction. The F-measure difference among ENET25, ENET50, ENET75, SVM, and SVM w/IG was also statistically insignificant with the exception that SVM has higher F-measure compared to the other 4 approaches for RN classification. While LASSO performed significantly worse than other statistical machine learning approaches for RALL classification, the performance difference between LASSO and other statistical machine learning approaches in RP, EU, and FT classification was insignificant. The empirical evidence is in line with Hypothesis H3 for RP, EU, and FT classification.

As discussed in the previous section, the agreement between human annotators can be considered as the upper bound of the classification task. The average agreement for the five recognition task was 91.8% while the average accuracy of statistical machine learning approaches was 76.0%. This suggests that statistical machine learning approaches achieved 82.8% of human performance. The pairwise difference between human performance and individual statistical machine learning approaches are statistically significant, which means that statistical machine learning approaches indeed perform worse than humans. However, compared to the two baseline models, I note that the performance level of the statistical machine learning approaches was not trivial and they indeed captured the gist of risk-related information in textual data.

4.4.4 Identifying the Most Suitable Model for Risk Recognition

Based on the results of the five recognition tasks, I can make the following observations. First, statistical machine learning approaches outperform baseline approaches based on opinion mining and keyword matching. Second, the best performer varies among the 5 classification tasks. Finally, the performance differences among SVMs and ENETs are not statistically significant while LASSO performed worse than other approaches for RALL classification.

One of my research questions is to identify the most suitable statistical machine learning approaches for risk recognition. The evaluation results indicate that SVM and ENETs have similar levels of performance. SVM and ENETs, nonetheless, are very different in terms of model interpretability. I argue that ENETs generate learned models

that are simpler and easier to interpret and that ENETs are more suitable for the tasks under consideration.

Models learned by ENETs are simpler and easier to interpret compared to SVM for the following reasons. First, there is a clear probability interpretation of the learned coefficients for ENETs. As discussed before, given the learned coefficient vector $\widehat{\mathrm{B}}$, the probability that a given sentence X_z is a positive case is $\mathrm{Prob}\left(\mathrm{Y}=1|\mathrm{X_z};\widehat{\mathrm{B}}\right)=\frac{1}{1+\mathrm{e}^{-\mathrm{X_z}\widehat{\mathrm{B}}}}$. Straight forward calculations show that $\log\frac{\mathrm{Prob}\left(\mathrm{Y}=1|\mathrm{X_z};\widehat{\mathrm{B}}\right)}{\mathrm{Prob}\left(\mathrm{Y}=0|\mathrm{X_z};\widehat{\mathrm{B}}\right)}=\mathrm{X_z}\widehat{\mathrm{B}}=\prod_{i=1}^{K}X_{z,i}\hat{B}_i$. The left hand side of the equation is the logarithm of odds ratio; the right hand side is the sum of the coefficient times the feature values (independent variables). It is clear from the equation that the coefficient of a feature is the marginal contribution to the log odds ratio. Coefficients larger than zero increase the odds ratio, while coefficients smaller than zero decrease the odds ratio. Zero coefficients have no contribution to the odds ratio.

The SVM model, on the other hand, does not start with a probability setting. The decision hyperplane is constructed according to the support vectors (i.e., a small set of training instances) identified during the optimization process. It is possible to compute the weights associated with each input feature from the support vectors. The absolute values of the weights hint at the relative importance of the features. However, there is no probability interpretation associated with the weights.

The shortcoming of lacking probability interpretation can be addressed by retro-fitting the signed distances with a sigmoid function to map the SVM outputs into probabilities (Platt 1999). This approach transfers the original (-1, 1) representation for

negative and positive cases back to the (0, 1) representation and fits a logistic regression with a constant term and the outputted SVM signed distance as the independent variable. The fitted logistic regression then provides probability interpretation for the SVM outputs. It is possible to derive the odds ratio associated with the input features from here.

I argue that any functions that map from the real line to the interval [0, 1] can potentially be used to assign the probability. Retro-fitting a logistic regression is a reasonable choice but it may not be the optimal one for assigning posterior probability to a given sentence. The probability computed based on ENETs, nonetheless, is based on the learned coefficients that maximized the posterior distribution given the training instances. ENET models are thus considered more rigorous compared to SVM.

The other reason that I prefer ENETs to SVM is because of the endogenous feature selection characteristics of ENET models. As discussed in the previous section, ENET automatically selects important features during model training and usually generates a model that includes only a subset of the original input features. For example, ENET25 constructed RALL classification using 216 out of about 6604 features. RP and RN classifiers involved only 50 and 412 features, respectively. While SVM had a higher F-measure for RN recognition, the learned model involved most of the 6604 features and was relatively difficulty to interpret. Reducing the number of features before training the SVM models did not help boost performance in most cases.

Based on the above discussion, I choose ENET25 for risk recognition. ENET25 performed relatively better among the three competing setting (ENET25, ENET50, and

ENET75) and possesses some of the desirable characteristics of the ENET models. My subsequent discussions focus on ENET25 only.

4.4.5 Feature Analysis

Analyzing how different features contribute to the classification tasks for risk recognition informs us the nature of the classification problems as well as the validity of the model. Table 4.9 lists the top 30 features for recognizing RALL sentences. As discussed above, the estimated coefficient can be interpreted as the marginal contribution to the logarithm of odds ratio, which eventually determines the classification results.

Table 4.9 Important Features for RALL Recognition

Feature	Coef.	Feature	Coef.	Feature	Coef.
expect	0.63	rating	0.38	last_week	0.29
will	0.59	the_ipod	0.37	it	0.28
who_have	0.55	wa_n't	0.35	all	0.27
volkswagen	0.55	investigation	0.34	down_from	0.27
much_as_$	0.52	riot	0.33	regulatory	0.27
ipod	0.50	plan	0.32	agree_acquire	0.27
infinity	0.48	say_it	0.32	outlook_for	0.26
breakup	0.45	in_this	0.32	VBZ	0.24
VB	0.43	coming	0.31	separately	0.24
next	0.43	MD	0.30	from_NUM_%	0.23

Four types of features are listed in Table 4.9. The first type of feature includes linguistic cues directly associated with modality. The most important feature of this type is “will,” which comes with a coefficient of 0.59. The POS tags, MD (modal) and VB (verb base form), are also associated with modality. These tags, to some extent, reflect the grammatical structure involved in delivering risk-related expressions. Future tense and hedging are examples that may associate MD and VB to risk-related sentences.

The second type of feature conveys risk-related meanings. The most prominent feature of this type is "expect," which has a coefficient of 0.63. Having this keyword in a sentence increased the odds ratio by exp(0.63) = 1.8 and contributes strongly to the classification of a RALL case. In addition to the top two features mentioned above," "next," "plan," "coming," and "outlook_for" also belong to this group. These features remind us about the disclaimer regarding forward-looking statement commonly seen in financial reports (e.g., 10-K):

This release includes forward-looking statements within the meaning of Section 27A of the Securities Act of 1933 and Section 21E of the Securities Exchange Act of 1934. Forward-looking statements are based on management's beliefs and assumptions. These forward-looking statements are identified by terms and phrases such as "anticipate," "believe," "intend," "estimate," "target," "expect," "continue," "should," "could," "may," "plan," "project," "predict," "will," "potential," "forecast," and similar expressions.

In fact, the definitions of forward looking statements are consistent with my definition of text-based risk measures since most of these statements may directly impact decision makers' belief about a firm's future cash flow. The similarity between the identified features and the keywords mentioned in the forward-looking statement disclaimer suggests that my classification models are valid and that there is a consistent understanding of risk-related information in business documents.

The third type of feature reflects language usage associated with risk-related sentences. For example, "much_as_$ (i.e., [as] "much as" followed by "$")" "sai_it," "in_this," "down_from," "from_NUM_% (i.e., "from" followed by a number and "%")"

do not have explicit meanings regarding risks. However, these words or phrases are strongly associated with risk-related information possibly because of the reporters' writing style.

The last type of feature includes proper nouns, such as "Volkswagen," "iPod," and "Infinity." These proper nouns reflect the specific content that has entered my random sample. It is arguable that proper nouns will be useful for unseen sentences. Larger testing samples over time and from different domains may help us design a better strategy to handle proper nouns.

An interesting observation of the top 30 features is that many of them are considered stop words. As discussed in the Feature Representation Section, stop words are associated with linguistic devices that are used to convey chance, reliability, and writers' commitment to reported information. My finding confirms the importance of these linguistic cues for text-based risk recognition. It also suggest that filtering out stop words may seriously damage the ability to recognize risk-related sentences given the fact that features related to stop words are often associated with large coefficients. This observation indicates that text-based risk recognition is very different from traditional information retrieval tasks that focus on identifying topical information.

4.5 Conclusion

Risk-related statements provide relevant information for business decision making under uncertainty. My research aims at developing a design framework for risk-related statement recognition. Informed by decision theories, I proposed three text-based risk

measures in a conceptual model that can be used to signal the existence of risk-related information in business documents. These measures cover the core inputs to decision making under uncertainty and form the basis for the AZRisk design framework.

The AZRisk design framework tackles the risk recognition problem by operationalizing the three proposed measures using statistical machine learning approaches. The proposed risk measures were transformed to binary sentence classification problems to facilitate the creation of a reference standard dataset and subsequent learning process. Using firm-specific news articles from the WSJ, the AZRisk design framework was evaluated using standard text classification performance measures. My results demonstrate the effectiveness of the AZRisk design framework. My approach achieved 82.8% of human performance and is significantly better than approaches using keyword search or opinion mining tools. Although the underlying theories clearly rely on these inputs, effective approaches to signal the existence of risk-related information did not exist before.

Moreover, my work represents an effort to understand the problem of recognizing risk-related information in business documents. I find that recognizing risk-related information at the sentence level is very different from topic-based information retrieval or identifying opinions in documents. Many of the important features identified in the experimental studies fall into the category of stop words, which may be removed by search engines and information retrieval systems. My results suggest that the unique characteristics of text-based risk recognition and the potential benefits associated with this type of information deserve the efforts of developing a novel design framework.

I am working on risk-related document surveillance systems that take advantage of AZRisk to provide visual summaries of stake holders, their relationships, and the potential risks involved. These efforts can provide more fine-grained summary information to support business decision making. Another potential research direction is to investigate the effects of risk-related information in news, newswire, and social media websites. I am particularly interested in the relative importance, interaction, and timeliness of risk-related information in these different channels.

CHAPTER 5. GIVING CONTEXT TO ACCOUNTING NUMBERS: THE ROLE OF NEWS SENTIMENT AND COVERAGE

Accounting numbers are an important means for management to communicate firm performance to outside investors. Through regularized financial statements, investors receive creditable and useful firm-specific information, which helps them better evaluate the true value of firms. High quality accounting numbers not only reduce the information asymmetry between managers and outside investors, but also facilitate sound investment decisions and the efficiency of security markets.

The usefulness of accounting numbers has been an important issue for accounting researchers and general investors. By measuring the magnitude of the relationship between stock returns and earnings using the earnings response coefficient (ERC), previous studies have concluded that the ERC is significantly positive (Kothari 2001) and numerical earnings information indeed conveys value-relevant information to the markets. The empirical results suggest that a favorable earnings surprise induces positive abnormal stock returns, while an unfavorable earnings surprise induces negative abnormal stock returns.

Accounting numbers, nonetheless, are not the only source of information conveying the fundamental value of firms. Other sources, such as financial news, trade association publications, and reports issued by analysts and brokerage houses may also contain useful information. These alternative information sources often provide timely updates between earnings announcements and may play an important role in shaping investors' beliefs. The magnitude of the ERC, as a result, may be influenced by these information sources.

While highly circulated financial news has been shown to impact short-term market returns (Tetlock 2007), few studies have investigated how financial news impacts the return-earnings relation. Tetlock et al. (Tetlock *et al.* 2008) quantified sentiment in news articles by counting words associated with negative outlooks. Their empirical results showed that the fraction of negative words in firm-specific news stories forecasts low firm earnings. Previous studies on ERCs have identified four important determinates: earnings persistency, firm risk, firm growth, and interest rate (Collins & Kothari 1989; Easton & Zmijewski 1989; Kothari 2001). However, these studies have not examined the interaction between the information content of earnings and news articles.

Given the limitations of previous studies, my research aims at investigating how financial news sentiment and coverage impact the return-earnings relation. To the best of my knowledge, this is the first study that documents how financial news sentiment and coverage affect investors' reactions to accounting earnings. I used the *Wall Street Journal* (WSJ) as the representative source of financial news and collected news articles discussing S&P 500 companies from August 1999 to February 2007. My collection contains 283,457 news articles and spans more than seven years. This testbed provides a solid ground for statistical inference. Firm-level news sentiment and coverage computed from my news collection facilitates my investigations on the interaction between financial news coverage and return-earnings relation.

The remainder of the chapter is organized as follows. Section 5.1 provides a review of related literature followed by the discussion of the research objectives and hypotheses in Section 5.2. In Section 5.3, I describe my data sources and empirical models. The

findings of the study are presented in Section 5.4. Section 5.5 discusses the implications of my results. I conclude with a summary and future research directions in Section 5.6.

5.1 Literature Review

The seminal works of Ball and Brown (Ball & Brown 1968) and Beaver (Beaver 1968) spawned the study of the information content of accounting numbers. Researchers study a wide range of topics via the return-earnings relation and the event study framework (Fama *et al.* 1969; Kothari 2001). In this section I first summarize the earnings response coefficient research and then focus on two major aspects that are directly relevant to this study: lagged performance information in accounting earnings and asymmetry in the return-earnings relation. I then summarize recent development on ambiguous information processing, which provides the theoretical foundation about how news sentiment impacts the return-earnings relation.

5.1.1 Earnings Response Coefficient

The information content of accounting numbers can be measured by the extent to which security prices change in response to the announcement of financial statements. One of the most commonly used measures is the Earnings Response Coefficient, ERC. Specifically, an ERC is estimated using the following model:

$$CAR_{it} = a + b\ UE_{it} + e_{it}$$

where e_{it} is white noise and CAR_{it} (Cumulated Abnormal Return; CAR) is the measure of risk-adjusted return for security i cumulated over an event window around the earnings announcement at time t. The ERC literature often adopts Fama and French three-factor

model (Fama & French 1993), which incorporates excess market returns (market returns minus risk-free interest rates), small minus large firm returns (SML), and high minus low book-to-market firm returns (HML) as main risk factors. Expected returns can be computed after the loadings of risk factors are estimated using historical data. Abnormal returns are the difference between actual stock returns and their expected returns computed using the three-factor model. It is a common practice to accumulate the abnormal returns over a period of time to capture the market reaction to earnings events.

UE_{it} is the unexpected earnings divided by security price at time t. Unexpected earnings captures the earnings variation that has not yet been incorporated into investors' belief. There are two approaches commonly used to compute unexpected earnings. The first approach estimates expected earnings based on a time series model. The second approach uses analysts' forecasts as a proxy for market expectation. Empirical evidence suggests that analysts' forecasts are a better proxy for market expectation of earnings (Livnat & Mendenhall 2006).

The estimated value of coefficient b is the ERC. While there is consensus that the ERC is significantly positive (Kothari 2001), the magnitude of the ERC may vary across firms. Collins and Kothari (Collins & Kothari 1989) suggested that the information environment, which can be broadly defined to include all sources of information relevant to firm value assessment, is an important determinant of ERCs. Their study, nonetheless, did not directly measure the information environment but instead looked at other determinants such as earnings persistency, firm risk, firm growth, and interest rates. Other studies that looked at the economic determinants of ERCs have identified various

characteristics that affect ERCs. Notable determinants include: firm size (Collins *et al.* 1987; Easton & Zmijewski 1989), capital structure (Dhaliwal *et al.* 1991), earnings persistence (Kormendi & Lipe 1987), earnings quality (Dechow & Dichev 2002; Francis *et al.* 2004), and similarity of investor expectations (Abarbanell *et al.* 1995).

5.1.2 Lagged Performance Information in Accounting Earnings

Accounting earnings measurement emphasizes transaction-based revenue recognition. As a result, a large portion of the information embedded in earnings is historical in nature. Security prices, on the other hand, reflect both current earnings as well as future earnings information that is available to the market (Kothari & Sloan 1992). As a result, value relevant information in accounting earnings may have been incorporated into stock prices before earnings announcements have been issued. Only a small portion of the earnings information content can be captured by measuring the stock price reaction around earnings events. The "price lead earnings" viewpoint (Beaver *et al.* 1980) provides a compelling explanation of why estimated earnings response coefficients are small in comparison to theoretical predictions (Kothari 2001).

The "price lead earnings" viewpoint has been verified by empirical studies that expand the return-earnings measurement window (Easton *et al.* 1992), include leading period return (Kothari & Sloan 1992; Jacobson & Aaker 1993), and include future earnings and future returns (Collins *et al.* 1994). However, most of these studies failed to explore other information sources that may have contained future earnings information. Financial news is one such information source (Ma *et al.* 2009; Muntermann 2009). Before being officially announced, information related to a firm's earnings is sometimes

disclosed through various news reports. Investors could incorporate the information into their determination of reasonable security prices. The return-earnings relation can be better modeled if relevant information from financial news could be captured.

5.1.3 Asymmetry in the Return-Earnings Relation

Previous studies have documented that security markets react to positive and negative earnings surprises differently. Positive earnings surprises were found to be associated with larger price responses (Hayn 1995; Lopez & Rees 2002). One possible reason is that widely adopted accounting conservatism recognizes probable losses as they are discovered but defers revenue until it is verified (Harrison & Horngren 2003). This practice increases the speed that accounting numbers reflect economic losses compared to economic gains. Timely recognition of economic losses leads to lower autocorrelations of negative shocks in earnings time series. The consequence is that accounting losses during the current period are less likely to signal future losses while accounting gains are more likely to signal future gains. This time series property is often referred to as the lower persistency of accounting losses relative to accounting gains (Basu 1997). Lower persistency is known to decrease ERCs (Kormendi & Lipe 1987).

Studies on firm disclosure activities point out that firm management is motivated to disclose bad news earlier in fear of litigation risk (Graham *et al.* 2005). This explanation suggests that information about a forthcoming negative earnings surprise may have been released earlier, which leads to lower ERC when the bad earnings are officially announced. The asymmetry in the return-earnings relation should be controlled in the empirical study so that valid results could be obtained.

5.1.4 The Effect of Ambiguity in Textual Data

Textual data such as news articles and financial reports contain qualitative information. Compared to accounting numbers which are quantitative in nature, the interpretation to textual data is usually less precise. A range of conclusions can often be reasonably supported based on a set of textual data. This characteristic is referred to as the ambiguity of textual data in this study.

Epstein and Schneider (Epstein & Schneider 2008) have developed a novel theoretical model for ambiguous information processing by treating ambiguous information using multiple likelihoods instead of a single likelihood in the classical Bayesian framework. In their model, decision makers are assumed to maximize the expected utility under a worst-case belief that is chosen from a set of conditional probabilities.

One important prediction of the model is the asymmetric response to the sentiment in news articles (Epstein & Schneider 2008). Bad news affects investors' decisions more than good news. The reason is that the worst case of good (bad) news sentiment is that the signal is unreliable (very reliable). Unreliable good news, as a result, is largely ignored while very reliable bad news may often trigger actions.

Under the context of earnings announcements, the news sentiment before earnings announcement is one type of ambiguous signal that affect investors' pricing decisions. The theoretical model predicts that investors react on negative news sentiment and positive news sentiment differently. The directions of news sentiment therefore, may interact with the return-earnings relation.

5.2 Hypotheses Development

In this study, I aim to investigate the role of financial news in determining the ERC. Financial news contains timely updates on firm value. The "price lead earnings" viewpoint suggests that investors would have incorporated the information from financial news into stock prices before the earnings announcements. The information content of earnings announcements, as a result, is reduced by news coverage. In other words, I expect a negative relationship between news coverage and ERCs:

H1: Firms with higher news coverage before their earnings announcements are associated with lower ERCs.

Previous studies on the asymmetry return-earnings relationship have documented that positive earnings surprises were found to be associated with larger price responses (Hayn 1995; Lopez & Rees 2002). In addition to this effect, the ambiguous information processing model of Epstein and Schneider (Epstein & Schneider 2008) predicts that investors respond to negative news sentiment but ignore positive news sentiment. Stock prices, as a result, decrease in response to negative news sentiment but remain the same when the pre-announcement period is associated positive news sentiment. Given that bad news has been incorporated into stock prices, positive unexpected earnings triggered larger adjustments because of the "surprise" caused by the inconsistency between news sentiment and unexpected earnings. The inconsistency between news sentiment and unexpected earnings thus pushes the ERC higher than that implied by the asymmetry return-earnings relationship:

H2: Other things been equal, negative news sentiment followed by positive unexpected earnings increases the ERC.

Positive news sentiment, on the other hand, does not affect stock prices. Consequently, the consistency between the sign of news sentiment and the sign of unexpected earnings is not associated with the variation of ERC:

H3: Other things been equal, positive news sentiment followed by negative unexpected earnings is not associated with additional increase in ERC.

Figure 5.1 schematically shows the hypothesized relationships in the research model. Measured by the ERC, the thick horizontal arrow indicates the relationship between unexpected earnings and cumulated abnormal returns. Hypothesis H1 asserts that news coverage moderates this relationship; higher news coverage reduces the ERC. Hypothesis H2 asserts that there is a three-way interaction between news sentiment, sign of unexpected earnings, and the return-earnings relation if news sentiment is negative.

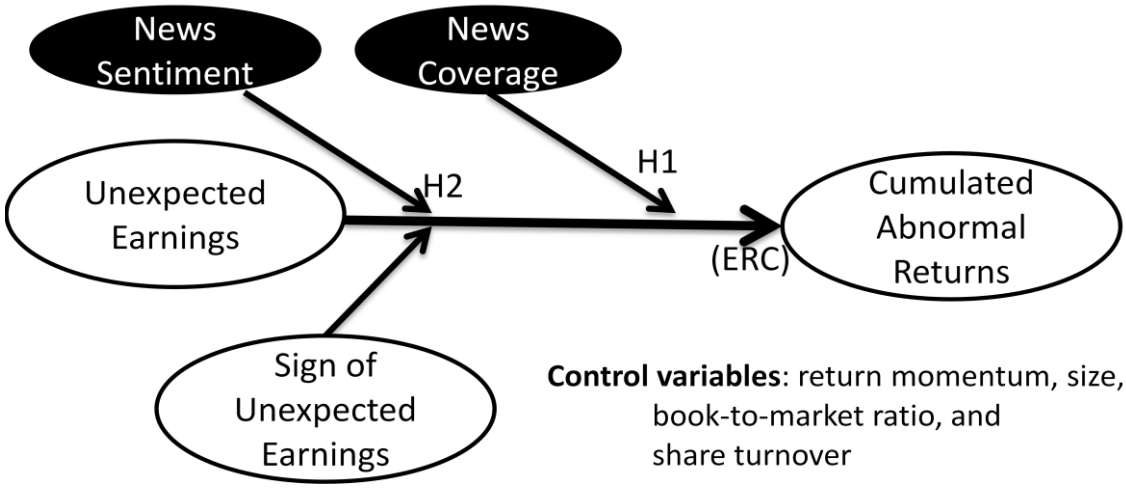


Figure 5.1 Research Model

5.3Research Methodology

5.3.1 Research Testbed

I used articles from the WSJ to develop my research testbed. News articles from August 1999 through February 2007 were collected from the ProQuest database. I retrieved 283,457 news articles in total for the 91-month period and developed a system for automatic analysis of firm news sentiment and coverage. I focused on S&P 500 companies because financial news tends to cover large firms. The monthly S&P 500 companies list was obtained from the Center for Research on Security Prices (CRSP) database.

Daily stock prices were obtained from CRSP and analysts' forecasts were retrieved from the Institutional Brokers' Estimate System (IBES). While accounting numbers are publicly available from the EDGAR system (Gerdes 2003), the data are not in a consistent format across firms or time period. I instead downloaded accounting numbers from the Compustat North American Annual and Quarterly database, which contained the same numerical data but in a structured format.

5.3.2 Firm-Based News Sentiment and Coverage Analysis

I conducted firm-based news sentiment and coverage analysis by creating a system that automatically extracts firm names and computes news sentiment. I defined that a firm received news coverage in a news article if the firm's name was mentioned at least once in the article. One news article may be associated with zero, one, or more firms.

News sentiment was computed by counting positive and negative words in a article. Each firm in the same article is associated with the same news sentiment measure.

Figure 5.2 presents the system design for firm-based news sentiment and coverage analysis. My system consists of two major modules: firm name extraction and sentiment analysis. The firm name extraction module performs named entity recognition and standardizes the recognized company names by consulting the "stocknames" table in the CRSP monthly stock price dataset (SM). Standardized firm IDs (PERMCO) are then attached to matched entities.

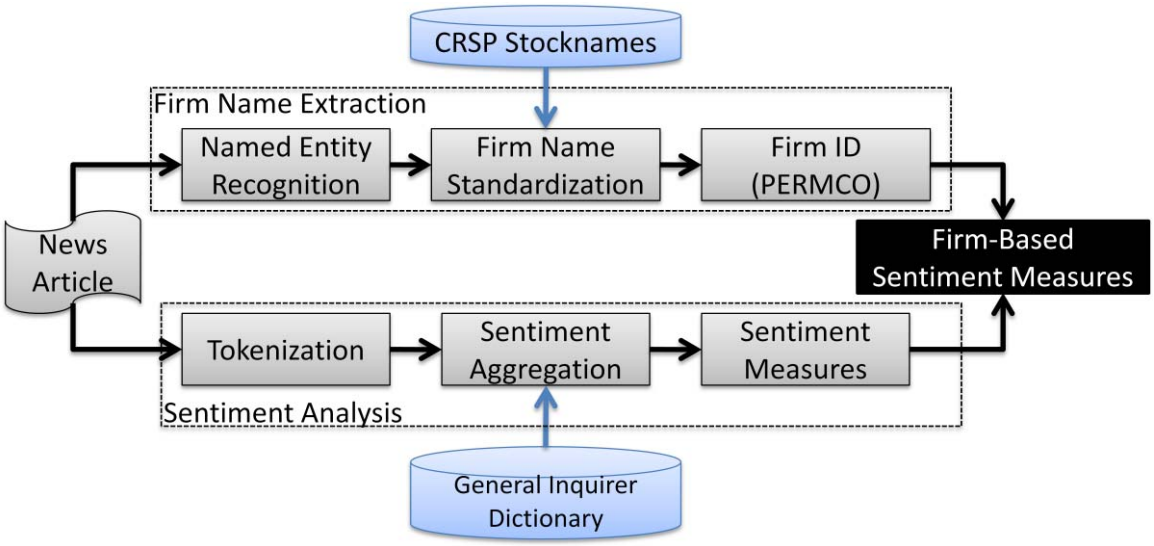


Figure 5.2 Firm-Based News Sentiment and Coverage Analysis

The design of my firm name matching process follows a tight-to-loose approach. Each recognized named entity goes through a three-stage matching process. The first stage matches the full named entity string against the firm names in the stocknames table. Since the company names in the stocknames table do not contain punctuation, all

punctuation marks in the original named entity string are replaced with white spaces. Extra white spaces (two or more consecutive white spaces) are removed. The process stops if the named entity string matches with an entry in the stocknames table.

In the second stage, the named entity string is gradually truncated when matching against firm names in the stocknames table. Each time the last word in the named entity string is removed if the previous string does not match with any entries. A match is obtained if the truncated named entity string is identical to the beginning part of a firm name in the stocknames table. The process stops if the truncated named entity string matches with an entry or the truncated string contains fewer than 2 words.

The third stage handles possible complications that involve acronyms. If part of a company name is an acronym, the company names in stocknames table often contains additional white spaces (e.g., "U S AIRWAY GROUP INC"). I address this issue by detecting acronyms in the recognized named entity and inserting additional white space between characters of an acronym before the matching process.

The sentiment analysis module conducts sentiment analysis by performing word-level sentiment aggregation. The input news articles are first divided into word tokens. Following the sentiment aggregation procedure proposed by Tetlock et al. (Tetlock *et al.* 2008), I compute the number of word tokens that belong to the positive or negative semantic categories in the General Inquirer (GI) dictionary. A dictionary based morphological analyzer is incorporated to find the base form of input word tokens before matching with the GI dictionary. The number of positive words and the number of negative words in an article are used to compute sentiment scores.

Combining the output from these two modules provides the input for my firm-based sentiment measures. I assume that the positive and negative words in an article apply equally to all firms that appear in the article. Two sentiment scores were considered in this study. Similar to Antweiler and Frank (Antweiler & Frank 2004), the first score is defined as

$$\mathrm{SENT1}_{i,[t_1,t_2]} = \frac{\sum_{t=t_1}^{t_2} \mathrm{PW}_{t,i} - \sum_{t=t_1}^{t_2} \mathrm{NW}_{t,i}}{\sum_{t=t_1}^{t_2} \mathrm{PW}_{t,i} + \sum_{t=t_1}^{t_2} \mathrm{NW}_{t,i}}$$

where $\mathrm{PW}_{t,i}$ is the total number of positive words in news articles that have mentioned firm i at date t; $\mathrm{NW}_{t,i}$ is the total number of positive words in news articles that have mentioned firm i at date t. The score is computed over a time period t_1 to t_2. This score, which reflects the relative portion of positive and negative words, is bounded by -1 and 1. A score of 1 (-1) is associated with all positive (negative) words for the firm and the time period under consideration.

The other sentiment score is a simple ratio between positive and negative words:

$$\mathrm{SENT2}_{i,[t_1,t_2]} = \frac{\sum_{t=t_1}^{t_2} \mathrm{PW}_{t,i}}{\sum_{t=t_1}^{t_2} \mathrm{NW}_{t,i}}.$$

This score is undefined when the total number of negative words is zero. Zero negative word, nonetheless, is rare for a time period with news coverage.

I computed news sentiment and frequency for each earnings announcement event using a [-20, -1] trading day window relative to the earnings announcement date. This twenty-trading-day window, which is roughly equal to one month, allows most firms to have news coverage before earnings announcements and facilitates the subsequent

statistical inference. Using trading day instead of calendar day can mitigate coverage variations caused by trading holidays.

For firm i that announced its earning at day t, both $\mathrm{SENT1}_{\mathrm{i},[\mathrm{t}-20,\mathrm{t}-1]}$ and $\mathrm{SENT2}_{\mathrm{i},[\mathrm{t}-20,\mathrm{t}-1]}$ were computed and record for subsequent analysis. If a firm has no news coverage during the predefined period, both sentiment scores were considered missing.

5.3.3 Empirical Model Specification

The basic abnormal return/unexpected earnings specification was used as a baseline model to evaluate all subsequent modifications (Collins *et al.* 1987). The regression model takes the following form:

$$\begin{aligned}\mathrm{CAR}_{\mathrm{it},[0,2]} = {} & \mathrm{a} + \mathrm{b}_1\mathrm{UE}_{\mathrm{it}} + \mathrm{b}_2\mathrm{Due}_{\mathrm{it}}\mathrm{UE}_{\mathrm{it}} + \mathrm{g}_1\mathrm{FFalpha}_{\mathrm{it}} + \\ & \mathrm{g}_2\log(\mathrm{Size}_{\mathrm{it}}) + \mathrm{g}_3\log(\mathrm{BM}_{\mathrm{it}}) + \mathrm{g}_4\log(\mathrm{STurnover}_{\mathrm{it}}) + \\ & \mathrm{h}_1\mathrm{FFalpha}_{\mathrm{it}}\mathrm{UE}_{\mathrm{it}} + \mathrm{h}_2\log(\mathrm{Size}_{\mathrm{it}})\,\mathrm{UE}_{\mathrm{it}} + \mathrm{h}_3\log(\mathrm{BM}_{\mathrm{it}})\,\mathrm{UE}_{\mathrm{it}} + \\ & \mathrm{h}_4\log(\mathrm{Sturnover}_{\mathrm{it}})\,\mathrm{UE}_{\mathrm{it}} + \mathrm{e}_{\mathrm{it}} \qquad (5.1)\end{aligned}$$

where Due_{it} is 1 if UE_{it} is positive and 0 if UE_{it} is negative. $\mathrm{CAR}_{\mathrm{it},[0,2]}$ is the cumulated abnormal return of firm i over the trading day window [0, 2] relative to an earnings announcement day t. The three-trading-day window captures the immediate response of earnings announcements (Lopez & Rees 2002). While a longer event window may increase the magnitude of ERC, the effects of other events during a longer window may interfere with the change of stock prices and bias the results (Kothari 2001).

I adopted the Fama-French three-factor model to control for common risks factors (Fama & French 1993). The values of the three risk factors (excess market return, SML and HML) were downloaded from Dr. Kenny French's website (http://mba.tuck.dartmouth.edu/ pages/faculty/ ken.french/index.html). For each earnings event, the regression coefficients estimated over the trading day window [-252, -21] relative to an earnings announcement day were used to compute abnormal returns during the [0, 2] window. The sum of the abnormal returns during the [0, 2] event window is my dependent variable. Excluding 20 trading days before earnings announcements avoided the confounding of the estimated coefficient with earnings announcements. The last term e_{it} is white noise.

The covariate UE_{it} is the unexpected earnings computed based on the difference between the realized earnings and analysts' forecasts, divided by the stock price of firm i at day t. Both the realized earnings per share (EPS) and analysts' forecast were obtained from the IBES database. I retrieved the analyst's most recent monthly mean forecast before earnings announcements. $FFalpha_{it}$ is the estimated intercept in the Fama-French three-factor model used to compute abnormal returns. This variable is used to control the return momentum effect (Jegadeesh & Titman 1993). Other independent variables, including $\log(Size_{it})$, $\log(BM_{it})$, and $\log(STurnover_{it})$, are control variables for firms' market value at time t, book-to-market ratio at time t, and share turnover rate during the reporting quarter associated with time t. Firm size ($\log(Size_{it})$) has been shown to influence the ERC (Lopez & Rees 2002). Book-to-market ratio ($\log(BM_{it})$) is related to future growth of a firm. A lower book-to-market ratio signals higher growth and vice-

versa (Skinner & Sloan 2002). Previous studies reported that growth firms tend to have higher ERCs (Lopez & Rees 2002). All interaction terms between control variables and UE_{it} are included to capture the potential moderating effect of control variables.

Note that by rearranging Equation (5.1), it implies that the ERC of firm i at time t is $b_1 + b_2 Due_{it} + h_1 FFalpha_{it} + h_2 \log(Size_{it}) + h_3 \log(BM_{it}) + h_4 \log(STurnover_{it})$. That is, the ERC is a function of the sign of unexpected earnings as well as other control variables. Coefficient b_1 is the ERC of a firm with zero $FFalpha_{it}$, $\log(Size_{it})$, $\log(BM_{it})$, and $\log(STurnover_{it})$ when unexpected earnings are negative. Coefficient b_2 captures the difference in the ERC between positive and negative earnings surprises. Coefficient h_j (j=1,2,3,4) captures the effect of control variables on the ERC. For example, h_1 can be interpreted as the change of the ERC when $FFalpha_{it}$ increases by one.

To study the effect of news sentiment, I classify news sentiment of each firm announcement into three categories according to the two sentiment score, SENT1 and SENT2. Specifically, I define an earnings announcement has positive news sentiment if its SENT1 is above the value corresponding to the 70th percentile. Similarly an earnings announcement has negative news sentiment if SENT1 is below the value corresponding to the 30th percentile. News sentiment scores between 30th and 70th percentiles are considered neutral. Similar definitions apply to SENT2. My discussion focuses on the positive and negative sentiment computed according to SENT1. The results using SENT2 is mainly for robustness check.

Two dummy variables were used to indicate the news sentiment. The dummy variable $Dsentp_{it}$ is set to 1 if the news sentiment about firm i during the [t-20, t-1] window is positive; $Dsentn_{it}$ is another dummy variable indicating negative news sentiment about firm i during the same period.

Hypothesis H1-H3 can be tested by the following model:

$$\begin{aligned} CAR_{it,[0,2]} = {} & a + b_1 UE_{it} + b_2 Due_{it} UE_{it} + b_3 Dsentp_{it} UE_{it} + \\ & b_4 Dsentn_{it} UE_{it} + b_5 Due_{it} Dsentp_{it} UE_{it} + \\ & b_6 Due_{it} Dsentn_{it} UE_{it} + c_1 NewsFreqnency_{it} + \\ & c_2 NewsFrequency_{it} UE_{it} + g_1 FFalpha_{it} + g_2 \log(Size_{it}) + \\ & g_3 \log(BM_{it}) + g_4 \log(STurnover_{it}) + h_1 FFalpha_{it} UE_{it} + \\ & h_2 \log(Size_{it}) UE_{it} + h_3 \log(BM_{it}) UE_{it} + \\ & h_4 \log(STurnover_{it}) UE_{it} + e_{it} \end{aligned} \quad (5.2)$$

where $NewsFrequency_{it}$ is the number of times firm i appears in financial news during the trading day window [-20, -1] relative to the earnings announcement date t. I computed $NewsFrequency_{it}$ via the output of my firm-based news coverage analysis system. $NewsFreqnency_{it}$ is also included following the framework for testing moderator effects (Baron & Kenny 1986).

Hypothesis H1 predicts that firms with higher news coverage before earnings announcements are associated with a lower ERC. By adding news coverage into the empirical model, the ERC of firm i at time t becomes $b_1 + b_2 Due_{it} + b_3 Dsentp_{it} + b_4 Dsentn_{it} + b_5 Due_{it} Dsentp_{it} + b_6 Due_{it} Dsentn_{it} + c_2 NewsFrequency_{it} + h_1 FFalpha_{it} + h_2 \log(Size_{it}) + h_3 \log(BM_{it}) + h_4 \log(STurnover_{it})$. The magnitude

of c_2 can be interpreted as the change in the ERC if firm i is covered in one additional news article during the 20 trading-day pre-announcement window. Hypothesis H1 predicts that c_2 is negative.

The coefficients b_3, b_4, b_5, and b_6 are used to capture the ERC difference under different directional combinations of news sentiment and unexpected earnings. The coefficient b_3 is the effect of positive news sentiment on ERC compared to neutral news sentiment. The coefficient b_4 is the effect of negative news sentiment on ERC compared to neutral news sentiment. The effect of positive news sentiment followed by positive earnings surprises is captured by b_5. The surprise effect associated with negative news sentiment followed positive earnings surprise is captured by b_6.

Hypothesis H2 predicts that $b_6 > 0$; Hypothesis H3 predicts that $b_5 = 0$. The coefficients b_3 and b_4 reflect the overall effect of positive and negative news sentiment on ERC. My hypotheses do not explicitly predict the signs of b_3 and b_4. Moreover, I do not have any reasons to believe that these two coefficients should be positive or negative. Previous studies on asymmetry return-earnings relationship suggest that $b_2 > 0$.

5.4 Empirical Results

I estimated Equations (5.1) and (5.2) via ordinary least square regression. The upper and lower 1% of UE_{it} and $\mathrm{CAR}_{\mathrm{it},[0,2]}$ were winsorized (i.e., extreme values were replaced with the 1% or 99% percentile values) to guard against outliers. The upper 1% of $\mathrm{NewsFrequency}_{\mathrm{it}}$ was also winsorized. Winsorizing selected variables prevents the potential negative effects of extreme values when conducting regression analysis. My

main results would remain qualitatively the same if unwinsorized data were used. The final sample consists of 11,201 firm-quarter observations. Descriptive statistics of variables for Equations (5.1) and (5.2) are summarized in Table 5.1. The average cumulated abnormal return ($CAR_{it,[0,2]}$) during the [0, 2] trading day window relative to earnings announcements is 0.00370 but is not significantly different from zero. The average of unexpected earnings is quite small (0.00046) with a relative large dispersion (std. dev. = 0.00275) even after winsorization. The mean of $NewsFrequency_{it}$ indicates that on average each S&P 500 company is mentioned in 5.971 news articles during the 20 trading days before earnings announcements. Market value ($Size_{it}$), book-to-market ratio (BM_{it}) and share turnover ($STurnover_{it}$) were transformed using the logarithm function to correct their skewed distributions.

Table 5.1 Descriptive Statistics of All Firm-Quarters.

Variable	Obs.	Mean	Median	Std. Dev.	Min.	Max.
CAR_{it}	11,201	0.00370	0.00315	0.05890	-0.17951	0.17239
UE_{it}	11,201	0.00046	0.00033	0.00275	-0.01360	0.01057
$NewsFrequency_{it}$	11,201	5.917	1.000	14.026	0.000	93.000
$FFalpha_{it}$	11,201	0.00028	0.00023	0.00134	-0.00767	0.01200
$\log(Size_{it})$	11,201	23.053	22.963	1.186	16.633	27.128
$\log(BM_{it})$	11,201	-1.118	-1.049	0.760	-8.047	2.066
$\log(STurnover_{it})$	11,201	-0.986	-1.073	0.638	-2.986	2.369

My full model involves news sentiment variables, and can only be estimated using the subsample with news coverage. There are 33% of firm-quarters did not have news coverage, leaving 7,488 firm-quarters in subsample with news coverage. Table 5.2 presents the summary statistics in the subsample with news coverage. The average of

$CAR_{it,[0,2]}$ and UE_{it} is lower compared to the whole sample. The difference, nonetheless, is small compared to their standard deviations. The mean and standard deviation of control variables are similar to those in the full sample. It is not surprising to see that the mean of $NewsFrequency_{it}$ increases from 5.917 to 8.837.

Table 5.2 Descriptive Statistics of Firm-Quarters with News Coverage

Variable	Obs.	Mean	Median	Std. Dev.	Min.	Max.
CAR_{it}	7,488	0.00271	0.00238	0.05892	-0.17951	0.17239
UE_{it}	7,488	0.00044	0.00032	0.00290	-0.01360	0.01057
$NewsFrequency_{it}$	7,488	8.837	3.000	16.385	1.000	93.000
$FFalpha_{it}$	7,488	0.00027	0.00023	0.00135	-0.00767	0.01200
$\log(Size_{it})$	7,488	23.311	23.211	1.219	16.633	27.128
$\log(BM_{it})$	7,488	-1.126	-1.052	0.775	-8.047	1.474
$\log(STurnover_{it})$	7,488	-1.003	-1.095	0.635	-2.986	1.597
SENT1	7,488	0.313	0.303	0.214	-1	1
SENT2	7,343	2.178	1.857	1.680	0	49

5.4.1 Baseline Models

Table 5.3 reports the estimation results of the baseline model (Equation 5.1). The second and third columns of Table 5.3 list a simplified version of Equation 5.1 that excludes the interaction terms between unexpected earnings and control variables. The result shows that the ERC is 4.51 and is significantly positive. The result is consistent with previous studies on ERCs (Collins *et al.* 1987; Kothari & Sloan 1992). Estimating the same model using firm-quarters with news coverage results in a smaller ERC (4.18). Before moving to the results of the more complicated model, I note that this pattern is consistent with my intuition that firms with news coverage have smaller ERCs. Two

control variables, $FFalpha_{it}$ and $\log(Size_{it})$, are significant, indicating a systematic variation of cumulated abnormal returns with respect to momentum and firm size.

Table 5.3 Regression Results of the Baseline Model

	Baseline: Equation 5.1 (Omit Interaction)		Baseline: Equation 5.1 (Full Model)	
	Full Sample	With News	Full Sample	With News
Intercept	0.052***	0.043***	0.042***	0.035**
UE_{it}	4.51***	4.18***	8.76**	7.50
$Due_{it}UE_{it}$			1.046***	1.054***
$FFalpha_{it}$	-3.48***	-4.04***	-3.32***	-3.89***
$\log(Size_{it})$	-0.0022***	-0.0018***	-0.0019***	-0.0016***
$\log(BM_{it})$	-0.00061	0.000073	-0.00056	-0.00012
$\log(STurnover_{it})$	-0.00076	-0.0014	-0.0012	-0.0023**
$FFalpha_{it}UE_{it}$			23.55	142.60
$\log(Size_{it})UE_{it}$			-0.30	-0.23
$\log(BM_{it})UE_{it}$			-2.26***	-1.87***
$\log(STurnover_{it})\ UE_{it}$			-0.84***	-0.35
Adj. R-Square	0.049	0.049	0.057	0.056

The last two columns of Table 5.3 report the estimation results using all covariates in Equation 5.1. Consistent with previous studies on asymmetry in the return earnings relation (Hayn 1995; Lopez & Rees 2002), the coefficient of $Due_{it}UE_{it}$ is significantly positive, which indicates that positive unexpected earnings are associated with larger ERC compared to negative unexpected earnings. The estimated coefficient of $\log(BM_{it})UE_{it}$ is significantly negative (-2.26). As discussed before, a lower book-to-market ratio signals higher growth and vice-versa (Skinner & Sloan 2002). The significant coefficient is consistent with previous studies that document higher ERCs for growth firms (Lopez & Rees 2002). Firm size has also been identified to be a determinant of ERCs. The estimated coefficient of $\log(Size_{it})UE_{it}$, nonetheless, is not significant,

while the sign of the coefficient is consistent with previous studies (Lopez & Rees 2002). The estimated coefficient of $\log(\text{STurnover}_{it})\text{UE}_{it}$ is significant using the full sample but shows insignificant results when the subsample was used. $\text{FFalpha}_{it}\text{UE}_{it}$ is not significant in my baseline model.

5.4.2 The Effect of News Coverage

Table 5.4 reports the estimation results of Equation 5.2. As listed in the forth to the fifth columns, my main concern is the interaction between news coverage and unexpected earnings ($\text{NewsFrequency}_{it}\text{UE}_{it}$). The estimated coefficient of $\text{NewsFrequency}_{it}\text{UE}_{it}$ is significantly negative. The estimation result suggests that, other things being equal, the ERC decreases -0.064 if a firm appears in one additional news article. A lower ERC indicates smaller market reactions given the same unexpected earnings and other firm characteristics. My empirical result supports hypothesis H1, which predicts a negative coefficient for $\text{NewsFrequency}_{it}\text{UE}_{it}$. It is interesting to note that the estimated coefficient for news coverage ($\text{NewsFrequency}_{it}$) is not significant. It means that $\text{NewsFrequency}_{it}$ does not influence CAR directly; the influence is through moderating the return-earnings relation.

Table 5.4 The Effect of News Sentiment and Coverage on ERC

	News Coverage: Omit Interaction		News Coverage		News Sentiment
	Full Sample	With News	Full Sample	With News	With News
Intercept	0.043***	0.038**	0.049***	0.043***	0.036**
UE_{it}	8.82**	7.69	-3.40	-6.87	-4.46
$\text{Due}_{it}\text{UE}_{it}$	1.040***	1.040***	1.11***	1.12***	0.72**
$\text{Dsentp}_{it}\text{UE}_{it}$					-1.56*
$\text{Dsentn}_{it}\text{UE}_{it}$					-1.62**

$Due_{it} Dsentp_{it} UE_{it}$					2.22**
$Due_{it} Dsentn_{it} UE_{it}$					2.14**
$NewsFrequency_{it}$	0.0000081	0.000019	0.000062	0.000070	0.000065
$NewsFrequency_{it} UE_{it}$			-0.064***	-0.065***	-0.062***
$FFalpha_{it}$	-3.31***	-3.88***	-3.27***	-3.84***	-3.82***
$\log(Size_{it})$	-0.0019***	-0.0017**	-0.0022***	-0.0020***	-0.0017**
$\log(BM_{it})$	-0.00057	-0.00016	-0.00078	-0.00034	-0.00071
$\log(STurnover_{it})$	-0.0012	-0.0023**	-0.0015	-0.0026**	-0.0026**
$FFalpha_{it} UE_{it}$	23.48	142.40	-32.16	85.16	57.35
$\log(Size_{it})UE_{it}$	-0.30	-0.24	0.29	0.46*	0.38
$\log(BM_{it})UE_{it}$	-2.26***	-1.87***	-1.97***	-1.56***	-1.50***
$\log(STurnover_{it})\ UE_{it}$	-0.84***	-0.35	-0.53*	0.054	-0.013
Adj. R-Square	0.056	0.055	0.059	0.058	0.058

Given the empirical support for the interaction between news coverage and unexpected earnings, it is important to know whether the significant result is driven by firm size. As reported in the first two columns of Table 5.4, the estimated coefficient of $\log(Size_{it})UE_{it}$ is negative when the interaction between news coverage and unexpected earnings ($NewsFrequency_{it} UE_{it}$) is excluded. The estimated coefficient of $\log(Size_{it})UE_{it}$ becomes positive when $NewsFrequency_{it} UE_{it}$ was included in the empirical model. Moreover, $NewsFrequency_{it} UE_{it}$ is significant at a 99% confidence level across both samples while $\log(Size_{it})UE_{it}$ is only significant at a 90% confidence level in the "with news" subsample. The results suggest that the interaction between news coverage and unexpected earnings is not a size effect despite relatively high correlation between size of a firm and its news coverage (correlation coefficient=0.468; p-value<0.01). Previous studies have documented lower ERCs for larger firms (Collins *et*

al. 1987; Easton & Zmijewski 1989), but the lower ERCs may have been affected or caused by news coverage that was omitted in the empirical models.

5.4.3 Interaction Between News Sentiment and Unexpected Earnings

The estimation results of Equation 5.2 using the subsample with news coverage are summarized in Table 5.4 (the last column). The estimated coefficient of $\mathrm{Due}_{it}\,\mathrm{Dsentn}_{it}\,\mathrm{UE}_{it}$ is significantly positive (2.14), indicating a surprise effect for earnings events with negative news sentiment followed by positive unexpected earnings. The estimation result is consistent with Hypothesis H2. The estimated coefficient of $\mathrm{Due}_{it}\,\mathrm{Dsentp}_{it}\,\mathrm{UE}_{it}$ is also significantly positive (2.22), which result is inconsistent with Hypothesis H3. The significantly positive coefficient indicates that, given positive news sentiment before earnings announcement, a positive unexpected earnings is associated with a higher ERC after controlling for the asymmetric return-earnings relation, the effect of positive news sentiment, and other moderators.

The coefficient of $\mathrm{Dsentp}_{it}\,\mathrm{UE}_{it}$ captures the effect of positive sentiment on ERC, compared to the earnings announcements with neutral news sentiment. The estimated coefficient of $\mathrm{Dsentp}_{it}\,\mathrm{UE}_{it}$ is significantly negative (-1.56). The negative coefficient indicates that, compared to neutral news sentiment, positive news sentiment is associated with a lower ERC. A similar effect exists for negative news sentiment; negative news sentiment is also associated with a lower ERC.

The coefficient of $\mathrm{NewsFrequency}_{it}\,\mathrm{UE}_{it}$, which captures the effect of news coverage on ERC, is still significantly negative and is of the same magnitude compared to the estimation results excluding news sentiment variables. The results suggest that

even after controlling for the interaction between the directional combinations of news sentiment and earnings surprise, news coverage still plays a role in decreasing the information content of earnings surprise. The estimation result is consistent with Hypothesis H1.

The asymmetry response of positive and negative earnings surprise is captured by the coefficient of $Due_{it}UE_{it}$, which has an estimation value of 0.72 and is significantly positive. The estimated value is smaller compared to those estimated without news sentiment variables.

For the purpose of robustness check, I have also replaced SENT1 with SENT2 during the process of classifying news sentiment. The empirical results are qualitatively the same as those presented in this section. My empirical results are robust with alternative sentiment measures.

5.5 Discussion

Based on the estimation results, I compute the ERC under different directional combinations of news sentiments and earnings surprise by setting the control variables and news coverage variables to the mean of the subsample with news coverage. As listed in Table 5.5, the largest ERC is associated with the condition when positive news sentiment is followed by positive unexpected earnings; the smallest ERC is associated with the condition when negative news sentiment is followed by negative unexpected earnings.

Table 5.5 Summary of ERC Under Different Directional Combinations of News Sentiment and Unexpected Earnings

ERC	$UE_{it} \geq 0$ ($\text{Due}_{it} = 1$)	$UE_{it} < 0$ ($\text{Due}_{it} = 0$)
Positive News Sentiment ($Dsentp_{it} = 1$)	6.97 (0.68) *	4.03 (0.74) *
Neutral News Sentiment ($Dsentp_{it} = 0,\ Dsentn_{it} = 0$)	6.31 (0.48) *	5.59 (0.48) *
Negative News Sentiment ($Dsentn_{it} = 1$)	6.83 (0.63) *	3.97 (0.61) *

* Standard deviations are in the parentheses.

The followings are some observations that can be made from the results. First, given the same level of news sentiment, positive unexpected earnings always have higher ERCs compared to negative unexpected earnings. The difference is especially large when the news sentiment is not neutral. Second, given the positive unexpected earnings, moving away from neutral news sentiment increases the ERC. However, having non-neutral news sentiment decreases the ERC given negative unexpected earnings.

While my empirical results is consistent with the prediction of ambiguous information processing model of Epstein and Schneider (Epstein & Schneider 2008) when the news sentiment is negative, the results associated with positive news sentiment reject Hypothesis H3 that is formed based on the ambiguous information processing model. The estimation results suggest that investors overreact to positive unexpected earnings given favorable news coverage before the announcement. One possible scenario is that investors postpone their re-evaluation until the release of numerical earnings information. If the numerical earnings information corroborate with the positive news sentiment, then the positive unexpected earnings trigger larger price movements. On the other hand, if the numerical earnings information disagrees with the positive news

sentiment, then the "contradicting" information canceled out. The price movement is smaller.

The effect of news coverage also provides useful insights. If investors want to profit from the price movement around earnings announcements, they should avoid firms with a high level of news coverage. Financial instruments such as call, put, and straddle options (Hull & Basu 2010) can be used to profit from significant stock price movements. While the investment strategies may vary depending on available information and investors' belief, stock prices need to move significantly in the expected direction in order to offset the costs associated with these financial instruments and profit from the investment strategies. Given the fact that news coverage reduces the ERC, investors should avoid firms with high news coverage if they are betting on large stock price movements.

5.6 Conclusions and Future Research Directions

This study investigates the influence of news coverage on the ERC, which measures the information content of earnings. I collected news articles in the *Wall Street Journal* from August 1999 through February 2007 to construct measures for news coverage on S&P 500 companies. Combined with data from classical financial databases such as IBES, Compustat and CRSP, I were able to study the effect of news coverage on earnings surprise.

My empirical results indicate that news coverage has a significantly negative effect on the ERC; higher news coverage decreases the information content of earnings and reduces market responses to unexpected earnings. While news coverage is correlated with firm size, the empirical evidences suggest that my findings are not a size effect. In

addition, news coverage is not subsumed by book-to-market ratio, share turnover rate, and return momentums.

Moreover, there are complex interaction between financial news and numerical earnings. Negative news sentiment in the pre-announcement period is associated with larger ERCs if the unexpected earnings are positive. Positive news sentiment, on the other hand, is associated with larger ERCs when the numerical earnings corroborate with the favorable news sentiment. The asymmetry effect of positive and negative news sentiment is unexpected given current theoretical framework of information processing. More systematic studies are required to understand the reasons behind the phenomena.

My study highlights the importance of financial news in conveying value-related information to the markets. I plan to include more information sources such as newswires, blogs and forum discussions to further investigate the interaction and relative importance of different sources.

CHAPTER 6. CONCLUSIONS, CONTRIBUTIONS, AND FUTURE DIRECTIONS

Surveillance using textual data provides an effective and efficient way to monitor changing business environments. Detecting new developments and trends in a timely manner can help a business respond to potential threats and opportunities properly and gain long-term competitive advantages. In order to improve the representation, modeling, and analysis of text-based surveillance, my dissertation presents a framework to conduct text quantification, anomaly detection, and empirical evaluation.

Among the proposed text quantification approaches, the text-based risk recognition study (Chapter 4) provides an effective framework to extract risk-related sentences in business documents and may be integrated into existing decision supporting systems to better monitor today's volatile environment. Other studies, such as chief complaint classification (Chapter 2), and time series outbreak detection (Chapter 3), are also valuable for surveillance in the real-world setting.

This chapter concludes my dissertation by summarizing the major contributions, discussing the relevance to research in management information systems, and proposing future research directions.

6.1 Contributions

This dissertation involves 4 studies of surveillance using textual data. I summarize the contributions of these studies in this section. In Chapter 2, I developed and evaluated an ontology-enhanced approach to classify free-text chief complaints into syndromic categories. CCs recorded in both English and Chinese can be processed under this

framework. This approach can cope with multiple sets of syndrome definitions. The core of this approach is the UMLS-based weighted semantic similarity score (WSSS) grouping method that is capable of automatically assigning previously un-encountered symptoms to appropriate symptom groups. An evaluation study shows that this approach can achieve a higher sensitivity, F measure, and F2 measure, when compared to the CC classification subsystem of EARS that has the same symptom grouping table and syndrome rules. This approach also outperforms RODS' CoCo native Bayesian classifier for syndrome categories that covers most CCs under consideration. As a side result, I also applied a bootstrapping-based statistical testing procedure to compare the performance of different methods. This procedure can be applied to compare sensitivity, specificity, positive predictive value, F measure, and F2 measure as long as the systems under consideration share a common reference standard dataset in which the independent assumption among records is reasonable.

I also studied the feasibility of extending an existing English-based CC classification system for Chinese syndromic surveillance. I used a statistical pattern extraction method based on the mutual information to extract important phrases from Chinese CCs and construct mappings to English. The UMLS-based BioPortal CC classifier, which was designed to process CCs in English, was used to process translated CCs. I compared the syndrome classification performance of the proposed translation method with those using the machine translation system provided by the Google Language Tool and a bilingual dictionary. Compared to Google Translation, my approach delivered significantly higher PPV, sensitivity, specificity, F measure, and F2 measure

for most syndromic categories. I found similar results in the comparison between my approach and the translations provided by the bilingual dictionary.

The observed superior performance of my proposed Chinese-English mapping approach indicates that the 470 key phrases extracted from about one million Chinese CCs could cover common triage usage. I believe that with a more comprehensive study of Chinese CC records, a set of standardized vocabulary could be constructed and my approach can be adopted in real-world applications. It should be noted that languages are constantly evolving and periodic reviews of extracted key phrases would be necessary to ensure inclusion of new phrases.

In Chapter 3 I studied disease outbreak detection using time series data, which is an important function for syndromic surveillance systems. I treated the disease outbreak as hidden outbreak states and developed a Markov switching with jumps model for syndromic surveillance. To handle the negative effect caused by the jumps in the observed time series, I extended the Markov switching model to include an extreme value filtering component. The negative effect of jumps can be successfully filtered out, which led to a lower false alarm rate.

I evaluated my disease outbreak detection approach using both simulated and real-world baseline time series, together with outbreaks simulated following established methods. Two benchmark surveillance methods were included. The first benchmark method, S+CUSUM, uses the Serfling model to filter out seasonal fluctuation and then applies the CUSUM method on standardized prediction errors. The second benchmark method, T+MA, uses the trimmed-mean seasonal ARMA model and computes the alert

scores using linear increasing weights. The evaluation results showed that my method achieved a similar level of detection timeliness and higher detection sensitivity compared to the benchmark outbreak detection methods. My approach had a detection sensitivity 23% to 328% higher than the benchmark methods.

The study in Chapter 4 aims at developing a design framework for risk-related statement recognition. Informed by decision theories, I proposed three text-based risk measures in a conceptual model that can be used to signal the existence of risk-related information in business documents. These measures cover the core inputs to decision making under uncertainty and form the basis for the AZRisk design framework.

The AZRisk design framework tackles the risk recognition problem by operationalizing the three proposed measures using statistical machine learning approaches. The proposed risk measures were transformed to binary sentence classification problems to facilitate the creation of a reference standard dataset and the subsequent learning process. Using firm-specific news articles from the WSJ, the AZRisk design framework was evaluated using standard text classification performance measures. My results demonstrate the effectiveness of the AZRisk design framework. My approach achieved 82.8% of human performance and is significantly better than approaches using keyword search or opinion mining tools. Although the underlying theories clearly rely on these inputs, effective approaches to signal the existence of risk-related information did not previously exist.

Moreover, my work represents an effort to understand the problem of recognizing risk-related information in business documents. I find that recognizing risk-related

information at the sentence level is very different from topic-based information retrieval or identifying opinions in documents. Many of the important features identified in the experimental studies fall into the category of stop words, which may be removed by search engines and information retrieval systems. My results suggest that the unique characteristics of text-based risk recognition and the potential benefits associated with this type of information deserve the effort of developing a novel design framework.

Chapter 5 studies the influence of news coverage on the ERC, which measures the information content of earnings. I collected news articles in the *Wall Street Journal* from August 1999 through February 2007 to construct measures for news coverage on S&P 500 companies. Combined with data from classic financial databases such as IBES, Compustat and CRSP, I studied the effect of news coverage on earnings surprise.

My empirical results indicate that news coverage has a significantly negative effect on the ERC; higher news coverage decreases the information content of earnings and reduces market responses to unexpected earnings. While news coverage is correlated with firm size, the empirical evidence suggests that my findings are not a size effect. In addition, news coverage is not subsumed by book-to-market ratio, share turnover rate, and return momentums.

Moreover, the interaction between financial news sentiment and numerical earnings is complex. Negative news sentiment in the pre-announcement period is associated with larger ERCs if the unexpected earnings are positive. Positive news sentiment, on the other hand, is associated with larger ERCs when the numerical earnings corroborate with the favorable news sentiment. The asymmetry effect between positive and negative news

sentiment is unexpected given the current theoretical framework of information processing. More systematic studies are required to understand the reasons behind these phenomena.

6.2 Relevance to Management Information Systems Research

This dissertation largely falls into the category of design science research (Hevner *et al.* 2004). As opposed to behavior science research, design science research generates and accumulates knowledge through action (Owen 1998). Almost all studies presented in this dissertation went through a typical design cycle that started with the awareness of the problem, followed by an abduction process based on existing theories and knowledge, IT artifact instantiations, and evaluation (Takeda *et al.* 1990). My research output covers the five major types of design research output: constructs, models, methods, and instantiations (March & Smith 1995). I discuss my research output in terms of this classification and the "interesting" new knowledge associated with the output.

Table 6.1 presents the mapping between the research output of this dissertation and the types of research output proposed by March and Smith (March & Smith 1995). The text-based risk recognition problem discussed in Chapter 4 is a new construct that looks at risk-related information in textual data developed based on existing decision making under uncertainty theories. Our experimental results show that the text-based risk recognition is unique in the sense that existing opinion mining and information retrieval approaches cannot handle the problem appropriately. It is clear from the results that 1) recognizing risk-related information in textual data has been largely negated despite its importance in terms of decision making under uncertainty, 2) the important features

associated with text-based risk recognition are consistent with the defining keywords for forward looking statements (one interpretation is that there is a consistent definition of risk-related information in the business domain), and 3) one particular type of statistical machine learning approach, i.e., elastic-net logistic regression, is more suitable for this type of task compared to other approaches.

Table 6.1 Research Outputs of my Dissertation

Type of Output	Description	Research Output
Constructs	The conceptual vocabulary of a domain	Text-based risk recognition problem (Chapter 4)
Models	A set of propositions or statements expressing relationships between constructs	The interaction between news sentiment, news coverage, and return-earnings relation (Chapter 5)
Methods	A set of steps used to perform a task	Markov Switching with Jumps models (Chapter 3)
Instantiations	The operationalization of constructs, models, and methods	BioPortal Chief Complaint Classification System (Chapter 2), and AZRisk System (Chapter 4)

The empirical results in Chapter 5 reveal the complex interaction between news sentiment, news coverage, and the return-earnings relation. It is clear that the textual data in news articles indeed influences how the numerical earnings information is interpreted in the stock market. The construction of the IT artifact for firm name standardization and news sentiment computation facilitates a better understanding of the underlying relationships. The empirical results show that news sentiment may provide an important context in which numerical earnings are interpreted. Investors seem to act on negative news sentiment but are more conservative faced with positive news sentiment. My

empirical results are in part consistent with the ambiguity information processing model in the literature (Epstein & Schneider 2008). The relationship for positive news sentiment, nonetheless, cannot be explained by existing theories and may be worth further investigation.

The Markov switching with jumps model discussed in Chapter 3 is a method that is designed to handle the time series outbreak detection problem. While the problem is well defined in the literature, my experimental results show an improvement in detection performance compared to other well-known time series surveillance approaches. It is clear from the results that having a statistical model that can handle both slow-moving and fast-moving outbreaks can outperform those that can only handle one of the two types of outbreaks.

Most of the studies in this dissertation provide certain instantiations for proposed constructs, models, and methods. Among them, the BioPortal Chief Complaint Classification System and the AZRisk System provide stand-alone, ready-to-use prototypes that are can be reused in future research. Moreover, part of the existing systems can also be integrated into existing business intelligence system and syndromic surveillance systems. Additional functionality can be readily achieved using these prototype systems.

6.3 Future Directions

My dissertation addresses various challenges faced in surveillance using textual data. The proposed constructs, methods, and models have shown promising performance

in various studied areas. These works can be extended in the following directions to further improve the effectiveness, efficiency, and social impact.

6.3.1 Chief Complaint Classification Systems

Besides obvious future work concerning additional data collection and testing to further evaluate my approach, several interesting research directions remain. First, the National Ambulatory Medical Care Survey (NAMCS) provides datasets that contain CCs with standardized coding (McCaig & Nawar 2006). These datasets may provide new resources for future CC classification research. Second, CCs are often available in languages other than English in international contexts. How to develop a working CC classification system in a multi-lingual environment poses interesting technical challenges, such as a US/Mexico cross border syndromic surveillance system.

Finally, other uses of a medical ontology in the CC classification process may be worth exploring. For instance, in the current process of producing the symptom grouping table, the experts are completely on their own in coming up with terms. One interesting extension is to use medical ontologies to help experts construct this table in an iterative manner by suggesting terms and groupings interactively.

The Chinese chief complaint classification approach provides additional future research directions. The syndrome definitions used in this study only cover those most commonly used by public health practitioners in Taiwan. I am currently working on identifying other useful syndromes and developing proper training and testing data. I also plan to extend my MIM-based approach and develop an approach that can be flexible enough for international public health situational awareness. In addition to technical

research, I am currently working with selected hospitals in Taipei to operationalize and validate my multilingual BioPortal system for syndromic surveillance. I expect that running the Chinese CC classification system in real-world settings (using original phrases) will validate my ideas and offer new technical insights to motive further research.

6.3.2 Time Series Outbreak Detection Using Markov Switching with Jumps Model

The results reported in my study suggest a promising future for the use of hidden state variables to model the changing dynamics of observed surveillance time series. I plan to extend my approach to outbreak detection with multiple data streams through multivariate time series analysis based on Markov switching. I am also exploring opportunities to apply the approach developed in this paper in areas beyond infectious disease informatics. One such area is sensor data integration and anomaly detection.

6.3.3 Text-Based Risk Recognition

I am working on risk-related document surveillance systems that take advantage of AZRisk to provide visual summaries of stake holders, their relationships, and the potential risks involved. These efforts can provide more fine-grained summary information to support business decision making. Another potential research direction is to investigate the effects of risk-related information in news, newswire, and social media websites. I am particularly interested in the relative importance, interaction, and timeliness of risk-related information in these different channels.

6.3.4 News Sentiment and Accounting Earnings

My study highlights the importance of financial news in conveying value-related information to the markets. I plan to include more information sources such as newswires, blogs and forum discussions to further investigate the interaction between and relative importance of different sources. I am also interested in studying the interaction between news sentiment and return-earnings relations. Sophisticated firm-based sentiment measures may reveal the underlying relationship among various textual information sources and how investors interpret the sentiment under the context of firm valuation.

APPENDIX A SELECTED DERIVATIONS OF THE POSTERIOR DISTRIBUTIONS FOR MARKOV SWITHCING WITH JUMPS MODEL

I provide in this appendix the outline for how to derive the conditional posterior distributions. The conditional posterior distributions play a key role in conducting statistical inference. The estimation process iterates to draw random variables from the conditional posteriors in order to construct the joint posterior distribution of parameters and hidden state variables. The discussion is based on the following model:

$$y_t = x_t + \xi_t J_t \tag{15}$$

$$x_t = a_{0,0} + a_{0,1} s_t + (a_{1,0} + a_{1,1} s_t) x_{t-1} + \sum_{i=1}^{6} d_{t,i} w_i + e_t \tag{16}$$

$$s_t \in \{0,1\} \tag{17}$$

$$J_t \in \{0,1\} \tag{18}$$

$$p(s_t = j \mid s_{t-1} = i) = p_{ij} \tag{19}$$

$$p(J_t = 1) = q_j \tag{20}$$

$$e_t \sim N(0, \sigma^2) \tag{21}$$

$$\xi_t \sim N(0, \sigma_a^2) \tag{22}$$

Note that this model is slightly different from the one used in my study. The main difference is that I assume that the time series has been "preprocessed" to remove seasonality. So the y_t here is equivalent to z_t in Equation 5. Also, without loss of generality, b_i is assumed to be zero. I applied Bayesian inference techniques in this study [73]. Specifically, Gibbs sampling was used. My basic model has the following state

variables: $S^T = (s_1,s_2,...,s_T)$, $J^T = (J_1,J_2,...,J_T)$, and $\Xi^T = (\xi_1,\xi_2,...,\xi_T)$. Although $X^T = (x_1,x_2,...,x_T)$ is not observed either, the values are fully determined if both J^T and Ξ^T are known. The coefficients to be estimated are denoted by $\Theta = (a_{0,0},a_{0,1},a_{1,0},a_{1,1},p_{11},p_{22},q_1,\ \sigma^2,\sigma_a^2,w_1,...,w_6)$.

Using the Gibbs sampling technique, I approximate the posterior distribution of parameters and hidden state variables, $p(\Xi^T,J^T,S^T,\Theta|Y^T)$, by iteratively drawing random variables from the following conditional distributions:

$$
\begin{aligned}
&p\left(\Xi^T \mid J^T,S^T,\Theta,Y^T\right)\\
&p\left(J^T \mid \Xi^T,S^T,\Theta,Y^T\right)\\
&p\left(S^T \mid \Xi^T,J^T,\Theta,Y^T\right)\\
&p\left(\Theta \mid \Xi^T,J^T,S^T,Y^T\right)
\end{aligned}
$$

The following is the derivation of the conditional posterior distributions.

A.0.1 Drawing from $p(\Xi^T \mid J^T,S^T,\Theta,Y^T)$

To draw Ξ^T from its conditional posterior, I iterate through each period and draw ξ_t given $\xi_{-t} = \{\xi_1,...,\xi_{t-1},\xi_{t+1},...,\xi_T\}$ and other random variables. Consider the jump size ξ_t at period t when the corresponding indicator variable $J_t = 1$.

$$
\begin{aligned}
& p\left(\xi_t \mid \xi_{-t},J^T,Y^T,S^T,\Theta\right)\\
\propto\ & p\left(Y^T,\xi_{-t},J^T \mid \xi_t,S^T,\Theta\right)p\left(\xi_t \mid S^T,\Theta\right)\\
\propto\ & p\left(Y^T \mid \xi_{-t},J^T,\xi_t,S^T,\Theta\right)p\left(\xi_{-t},J^T \mid \xi_t,S^T,\Theta\right)p\left(\xi_t \mid S^T,\Theta\right)\\
\propto\ & p\left(y_{t+1} \mid y_t,\Xi_T,J^T,S^T,\Theta\right)p\left(y_t \mid y_{t-1},\Xi_T,J^T,S^T,\Theta\right)p(\xi)\\
\equiv\ & p\left(y_{t+1} \mid y_t\right)p\left(y_t \mid y_{t-1}\right)p(\xi)
\end{aligned}
$$

Note that $p(\xi_t|S^T,\Theta) = p(\xi_t)$ and $p(\xi_{-t},J^T|\xi_t,S^T,\Theta) = p(\xi_{-t})p(J^T)$ by definition. To make the equations easier to read, I suppressed the conditioning on Ξ_T,J^T,S^T,Θ at (23) and in the following discussion.

Since

$$p(y_{t+1} \mid y_t) \propto \frac{-1}{\sigma^2}\left[y_{t+1} - J_{t+1}\xi_{t+1} - \left(a_{0,0} + a_{0,1}s_{t+1}\right) - \left(a_{1,0} + a_{1,1}s_{t+1}\right)\left(y_t - J_t\xi_t\right)\right]^2 \quad (24)$$

$$p(y_t \mid y_{t-1}) \propto \frac{-1}{\sigma^2}\left[y_t - J_t\xi_t - \left(a_{0,0} + a_{0,1}s_t\right) - \left(a_{1,0} + a_{1,1}s_t\right)\left(y_{t-1} - J_{t-1}\xi_{t-1}\right)\right]^2 \quad (25)$$

$$p(\xi) \propto \frac{-1}{\sigma_a^2}\xi_t^2 \quad (26)$$

Substituting (24) – (26) back to (23) and complete square with respect to ξ_t, I get the conditional posterior distribution of ξ_t:

$$\xi_t \mid J_t = 1 \sim N(m_t, v_t) \quad (27)$$

$$\xi_t \mid J_t = 0 \sim N(0, \sigma_a^2) \quad (28)$$

where

$$m_t = \frac{\sigma_a^2\left(z_t - \phi_{t+1}z_{t+1}\right)}{\sigma_a^2\left(1 + \phi_{t+1}^2\right) + \sigma^2}$$

$$v_t = \frac{\sigma^2\sigma_a^2}{\sigma^2 + \left(1 + \phi_{t+1}^2\right)\sigma_a^2}$$

$$z_t = y_t - \alpha_t - \phi_t\left(y_{t-1} - \xi_{t-1}J_{t-1}\right)$$

$$z_{t+1} = y_{t+1} - \xi_{t+1}J_{t+1} - \alpha_{t+1} - \phi_{t+1}y_t$$

$$\alpha_t = a_{0,0} + a_{0,1}s_t$$

$$\phi_t = a_{1,0} + a_{1,1}s_t$$

A.0.2 Drawing from $p(J^T \mid \Xi^T, S^T, \Theta, Y^T)$

The posterior of J_t is

$$\begin{aligned} & p\left(J_t \mid \Xi_T, Y^T, J_{-t}, S^T, \Theta\right) \\ \propto\ & p\left(Y^T \mid J_t, J_{-t}, \Xi_t, S^T, \Theta\right) p\left(J_t \mid \Xi_t, S^T, \Theta\right) \\ \propto\ & p\left(y_{t+1} \mid y_t, \Xi_T, J^T, S^T, \Theta\right) p\left(y_t \mid y_{t-1}, \Xi_T, J^T, S^T, \Theta\right) p\left(J_t\right) \\ \equiv\ & p\left(y_{t+1} \mid y_t\right) p\left(y_t \mid y_{t-1}\right) p\left(J_t\right) \end{aligned} \tag{29}$$

The posterior probability of $J_t = 1$ can be calculated by considering the odd ratio

$$\frac{p\left(J_t = 1 \mid \bullet\right)}{p\left(J_t = 0 \mid \bullet\right)} \tag{30}$$

A.0.3 Drawing from $p(S^T \mid \Xi^T, J^T, \Theta, Y^T)$

Since $x_t = y_t - \xi J_t$, the conditional posterior $p(S^T \mid \Xi^T, J^T, \Theta, Y^T)$ can be written as $p(S^T \mid X^T, \Theta)$. Multi-move Gibbs sampling is used to draw S^T from its posterior. To achieve this, the first step is to calculate the filtered state probabilities, i.e., $p(s_t = l \mid X^t)$, $l \in \{0,1\}$. The calculation can be divided into three steps:

(1) One-step ahead prediction of s_t:

$$p\left(s_t = l \mid X^{t-1}\right) = \sum_{k=0}^{1} p_{kl}\, p\left(s_{t-1} = l \mid X^{t-1}\right) \tag{31}$$

(2) Filtering for s_t

$$p\left(s_t = l \mid X^t\right) = \frac{p\left(x_t \mid s_t = l, X^{t-1}\right) p\left(s_t = l \mid X^{t-1}\right)}{p\left(x_t \mid X^{t-1}\right)} \tag{32}$$

where

$$p\left(x_t \mid X^{t-1}\right)=\sum_{k=0}^{1} p\left(x_t \mid S_t=k, X^{t-1}\right) p\left(s_t=k \mid x^{t-1}\right) \tag{33}$$

The smoothed probability $p(S^T \mid X^T, \Theta)$ can be calculated as follows:

$$p\left(S_t=l \mid X^T\right)=\sum_{k=0}^{1} \frac{p_{lk} p\left(s_t=l \mid X^t\right) p\left(s_{t+1}=k \mid X^T\right)}{\sum_{j=0}^{1} p_{jk} p\left(S_t=j \mid X^t\right)} \tag{34}$$

The multi-move Gibbs sampling makes use of the following expansion for S^T:

$$\begin{aligned} & p\left(S^T \mid X^T\right) \\ = \; & p\left(s_T \mid X^T\right) p\left(s_{T-1} \mid s_T, X^{T-1}\right) p\left(s_{T-2} \mid s_{T-1}, X^{T-2}\right) \cdots p\left(s_1 \mid s_2, x_1\right) \\ = \; & p\left(s_T \mid X^T\right) \prod_{t=1}^{T-1} p\left(s_t \mid s_{t+1}, X^t\right) \end{aligned}$$

where

$$p\left(s_t \mid X^t, s_{t+1}\right) \propto p\left(s_{t+1} \mid s_t\right) p\left(s_t \mid X^t\right) \tag{35}$$

A.0.4 Drawing from $p(\Theta^T \mid \Xi^T, J^T, S^T, Y^T)$

Given state variable S^T and jump variables J^T, Ξ^T, the posterior distribution of $a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1}, w_i, \sigma^2$ follows from the standard Bayesian regression model.

Specifically, let $m_t = \{1, s_t, x_{t-1}, s_t x_{t-1}, d_{t,1}, \ldots, d_{t6}\}$ be a row vector, then $M^T = \{m_1' m_2' \ldots m_T'\}'$ is a matrix with T rows. Then the posterior of $\beta = \{a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1}, w_1, w_2, \ldots w_6\}$ follows a normal distribution

$$\beta \sim N(u_\beta, v_\beta) \quad (36)$$

$$u_\beta = v_\beta\left(\frac{M'X}{\sigma^2} + v_0^{-1}\beta_0\right) \quad (37)$$

$$v_\beta^{-1} = \frac{M'M}{\sigma^2} + v_0^{-1} \quad (38)$$

The posterior distribution of σ^2 follows Inverse Gamma distribution

$$\sigma^2 \sim IG(v_g, \lambda_g) \quad (39)$$

$$v_g = v_{g0} + (T)/2 \quad (40)$$

$$\lambda_g = \lambda_{g0}\left(U^{T'}U^T\right)\Big/2 \quad (41)$$

$$U^T = X^T - M^T\beta \quad (42)$$

The posteriors of p_{00} and p_{11} are

$$p_{00} \mid S^T \sim beta(u_{00} + n_{00}, u_{01} + n_{01}) \quad (43)$$

$$p_{11} \mid S^T \sim beta(u_{11} + n_{11}, u_{10} + n_{10}) \quad (44)$$

where n_{ij} refers to the count of transitions from state i to j, which can be calculated directly from S^T. u_{ij} refers to the parameters of the prior distributions for p_{00} and p_{11}.

The posterior of q_1 is

$$q_1 \mid S^T \sim beta(v_1 + n_{v1}, v_0 + n_{v0}) \quad (45)$$

where n_{v1} is the count of $J_t = 1$ and n_{v0} is the count of $J_t = 0$. v_1 and v_0 are the parameters of the prior distribution of q_1.

REFERENCES

Abarbanell, J.S., Lanen, W.N., Verrecchia, R.E., 1995. Analysts' forecasts as proxies for investor beliefs in empirical research. Journal of Accounting and Economics 20, 31-60.

Abbasi, A., Chen, H., 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems 26, 1-29.

Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems 26, 1-34.

Achour, S.L., Dojat, M., Rieux, C., Bierling, P., Lepage, E., 2001. A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development. Journal of the American Medical Informatics Association 8, 351-360.

Agresti, A., 2002. Categorical data analysis. John Wiley & Sons, Hoboken, New Jersey.

Akaike, H., 1970. Statistical predictor identification. Annals of the Institute of Statistical Mathematics 22, 203-217.

Akaike, H., 1973. Information theory and an extension of the likelihood principle. In: Proceedings of the Second International Symposium of Information Theory Perov BN & Csaki F (Eds.)

Albert, J.H., Chib, S., 1993. Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts. Journal of Business & Economic Statistics 11, 1-15.

Aljlayl, M., Frieder, O., Grossman, D., 2002. On bidirectional English-Arabic search. Journal of the American Society for Information Science and Technology 53, 1139 - 1151.

Anscombe, F., Aumann, R., 1963. A definition of subjective probability. Annals of Mathematical Statistics 34, 199-205.

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of Internet stock message boards. Journal of Finance 59, 1259-1294.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., Sadler, L., 1994. Machine translation: an introductory guide. Blackwells-NCC, London

Aronsky, D., Kendall, D., Merkley, K., James, B.C., Haug, P.J., 2001. A comprehensive set of coded chief complaints for the emergency department. Academic Emergency Medicine 8, 980-989.

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research 6, 159--178.

Baron, R.M., Kenny, D.A., 1986. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 51, 1173-1182.

Basu, S., 1997. The conservatism principle and the asymmetric timeliness of earnings. Journal of Accounting & Economics 24, 3-37.

Baum, L.E., Egon, J.A., 1967. An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology. Bull. Amer. Meteorology Soc. 73, 360-363.

Baum, L.E., Petrie, T., 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. Annals of Math. Statistics 37, 1554-1563.

Beaver, W., Lambert, R., Morse, D., 1980. The information content of security prices. Journal of Accounting and Economics 2, 3-28.

Beaver, W.H., 1968. The information content of annual earnings announcements. Journal of Accounting Research 6, 67-92.

Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete choice analysis : theory and application to travel demand. MIT Press, Cambridge, Mass.

Besag, J., 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society. Series B (Methodological) 36, 192-236.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Box, G., Jenkins, G., 1970. Time series analysis: Forecasting and control, San Francisco: Holden-Day.

Brillman, J.C., Burr, T., Forslund, D., Joyce, E., Picard, R., Umland, E., 2005. Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance. BMC Med Inform Decis Mak 5

Brown, R.D., 1996. Example-based machine translation in the Pangloss system. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pp. 169-174, Copenhagen, Denmark.

Bruce, R.F., Wiebe, J.M., 1999. Recognizing subjectivity: a case study of manual tagging. Natural Language Engineering 1, 1-16.

Buckeridge, D.L., Switzer, P., Owens, D., Siegrist, D., Pavlin, J., Musen, M., 2005. An evaluation model for syndromic surveillance: assessing the performance of a temporal algorithm. MMWR Morb. Mortal. Wkly. Rep. 54 Suppl, 109-115.

Burkom, H., 2007. Disease Surveillance: A Public Health Informatics Approach. In: Lombardo JS & Buckeridge DL (eds.) Alerting Algorithms for Biosurveillance. John Wiley & Sons, pp. 143-192.

Burkom, H.S., Murphy, S.P., Shmueli, G., 2007. Automated time series forecasting for biosurveillance. Stat Med 26, 4202-4218.

Burr, T., Graves, T., Klamann, R., Michalak, S., Picard, R., Hengartner, N., 2006. Accounting for seasonal patterns in syndromic surveillance data for outbreak detection. BMC Medical Informatics and Decision Making 6

Carter, C.K., Kohn, R., 1994. On Gibbs Sampling for State Space Models. Biometrika 81, 541-553.

CDC, 2006. Increased antiviral medication sales before the 2005-06 influenza season--New York City. MMWR Morb. Mortal. Wkly. Rep. 55, 277-279.

Chandrasekaran, S., English, J.R., Disney, R.L., 1995. Modeling and Analysis of EWMA Control Schemes with Varance-Adjusted Control Limits. IIE Transactions 27, 282-290.

Chang, W., Zeng, D., Chen, H., 2005. Prospective spatio-temporal data analysis for security informatics. In: IEEE Conference on Intelligent Transportation Systems, Vienna, Austria.

Chapman, W.W., 2006. Natural language processing for biosurveillance. In: Wagner MM, Moore AW & Aryel RM (eds.) Handbook of Biosurveillance. Elsevier, New York, pp. 255-271.

Chapman, W.W., Christensen, L.M., Wagner, M.M., Haug, P.J., Ivanov, O., Dowling, J.N., Olszewski, R.T., 2005a. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artificial Intelligence in Medicine 33, 31-40.

Chapman, W.W., Dowling, J.N., Wagner, M.M., 2004. Fever detection from free-text clinical records for biosurveillance. Journal of Biomedical Informatics 37, 120-127.

Chapman, W.W., Dowling, J.N., Wagner, M.M., 2005b. Generating a reliable reference standard set for syndromic case classification. Journal of the American Medical Informatics Association 12, 618-629.

Chapman, W.W., Haug, P.J., 1999. Comparing expert systems for identifying chest X-ray reports that support pneumonia. In: Proceedings of the AMIA Annual Symposium, pp. 216-220

Chen, H.-H., 2002. Cross-Language Information Retrieval: Theories and Technologies. Journal of Library and Information Science 28, 19-32.

Cheng, K.S., Young, G.H., Wong, K.F., 1999. A study on word-based and integral-bit Chinese text compression algorithm. Journal of the American Society for Information Science 50, 218-228.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings Algorithm. The American Statistician 49, 327-335.

Chien, L.-F., 1999. PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval. Information Processing and Management 35, 501-521.

Chow, G.C., 1960. Tests of Equality Between Sets of Coefficients in Two Linear Regressions. Econometrica 28, 591-605.

Chu, C.-S.J., Stinchcombe, M., White, H., 1996. Monitoring structural change. Econometrica 64, 1045-1065.

Cleary, J.G., Witten, I.H., 1984. Data compression using adaptive coding and partial string matching. IEEE Transactions on Communications 32, 396-402.

Clements, M.P., Hendry, D.F., 2006. Forecasting with breaks. In: Elliot G, Granger CWJ & Timmermann A (eds.) Handbook of Economic Forecasting. Elsevier, pp. 605-657.

Coates, J., 1987. Epistemic modality and spoken discourse. Transactions of the Philological Society 85, 110-131.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46.

Collins, D.W., Kothari, S.P., 1989. An analysis of intertemporal and cross-sectional determinants of earnings response coefficients. Journal of Accounting and Economics 11, 143-181.

Collins, D.W., Kothari, S.P., Rayburn, J.D., 1987. Firm size and the information content of prices with respect to earnings. Journal of Accounting and Economics 9, 111-138.

Collins, D.W., Kothari, S.P., Shanken, J., Sloan, R.G., 1994. Lack of timeliness and noise as explanations for the low contemporaneuos return-earnings association. Journal of Accounting and Economics 18, 289-324.

COSO, 2004. Enterprise risk management - intergrated framework. Committee of Sponsoring Organizations of the Treadway Commission (COSO)

Crubezy, M., O'Connor, M., Buckeridge, D.L., Pincus, Z., Musen, M.A., 2005. Ontology-centered syndromic surveillance for bioterrorism. IEEE Intelligent Systems, Special Issue on Artificial Intelligence for National and Homeland Security 20, 26-35.

Dahlquist, M., Gray, S.F., 2000. Regime-switching and interest rates in the European monetary system. Journal of International Economics 50, 399-419.

Daumke, P., Marku, K., Poprat, M., Schulz, S., Klar, R., 2007. Biomedical information retrieval across languages. Informatics for Health and Social Care 32, 131-147.

Day, F.C., Schriger, D.L., La, M., 2004. Automated linking of free-text complaints to Reason-for-Visit categories and International Classification of Diseases diagnoses in emergency department patient record databases. Annals of Emergency Medicine 43, 401-409.

Dechow, P.M., Dichev, I.D., 2002. The quality of accruals and earnings: the role of accrual estimation errors. The Accounting Review 77, 35-59.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1-38.

Dhaliwal, D.S., Kyung J, L.E.E., Fargher, N.L., 1991. The association between unexpected earnings and abnormal security returns in the presence of financial leverage. Contemporary Accounting Research 8, 20-41.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10, 1895-1923.

Dumais, S., Platt, J., Heckerman, D., Sahami, M., 1998. Inductive learning algorithms and representations for text categorization. In: Proceedings of the seventh international conference on Information and knowledge management, pp. 148-155. ACM, Bethesda, Maryland, United States.

Easton, P.D., Harris, T.S., Ohlson, J.A., 1992. Aggregate accounting earnings can explain most of security returns : The case of long return intervals. Journal of Accounting and Economics 15, 119-142.

Easton, P.D., Zmijewski, M.E., 1989. Cross-sectional variation in the stock market response to accounting earnings announcements. Journal of Accounting and Economics 11, 117-141.

Efron, B., 1979. Bootstrap methods: Another look at the jackknife. Annals of Statistics 7, 1-26.

Efron, B., 1983. Estimating the error rate of a prediction rule: some improvements on cross-validation. Journal of the American Statistical Association 78, 316-331.

Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. Statistical Science 1, 54-77.

Efron, M., 2004. Cultural Orientation: Classifying Subjective Documents by Cociation Analysis. In: Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design

Epstein, L.G., Schneider, M., 2008. Ambiguity, Information Quality, and Asset Pricing. The Journal of Finance 63, 197-228.

Espino, J.U., Dowling, J., Levander, J., Sutovsky, P., Wagner, M.M., Copper, G.F., 2006. SyCo: a probabilistic machine learning method for classifying chief complaints into symptom and syndrome categories. In: Syndromic Surveillance Conference, Baltimore, Maryland.

Espino, J.U., Wagner, M.M., 2001. The accuracy of ICD-9 coded chief complaints for detection of acute respiratory illness. In: Proceedings of the AMIA Annual Symposium, pp. 164-168

Espino, J.U., Wagner, M.M., Tsui, F.C., Su, H.D., Olszewski, R.T., Lie, Z., Chapman, W., Zeng, X., Ma, L., Lu, Z.W., Dara, J., 2004. The RODS Open Source Project: removing a barrier to syndromic surveillance. Stud Health Technol Inform 107, 1192-1196.

Fama, E.F., Fisher, L., Jensen, M.C., Roll, R., 1969. The adjustment of stock prices to new information. International Economic Review 10, 1-21.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3-56.

Figueiredo, M.A.T., Jain, A.K., 2001. Bayesian learning of sparse classifiers. In: Computer Vision and Pattern Recognition, IEEE Computer Society Conference on

Fisher, F.S., Whaley, F.S., Krushat, W.M., Malenka, D.J., Fleming, C., Baron, J.A., Hsia, D.C., 1999. The accuracy of Medicare's hospital claims data: Progress has been made, but problems remain. American Journal of Public Health 82, 243-248.

Fleiss, J.L., 1981. Statistical methods for rates and proportions. John Wiley & Sons, NY, NY.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289-1305.

Francis, J., LaFond, R., Olsson, P.M., Schipper, K., 2004. Costs of equity and earnings attributes. Accounting Review 79, 967-1010.

Friedman, J., Hastie, T., Tibshirani, R., 2009. Regularization paths for generalized linear models via coordinate descent. Department of Statistics, Stanford University.

Frisen, M., 2003. Statistical surveillance. Optimality and Methods. International Statistical Review 71, 403-434.

Frisen, M., De Mare, J., 1991. Optimal Surveillance. Biometrika 78, 271-280.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721-741.

Gerdes, J., 2003. EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC's EDGAR database. Decision Support Systems 35, 7-29.

Gesteland, P.H., Gardner, R.M., Tsui, F.-C., Espino, J.U., Rolfs, R.T., James, B.C., Chapman, W.W., Wagner, A.W.M.M.M., 2003. Automated syndromic surveillance for the 2002 winter olympics. Journal of the American Medical Informatics Association 10, 547-554.

Goh, C.L., Asahara, M., Matsumoto, Y., 2005. Chinese word segmentation by classification of characters. Computational Linguistics and Chinese Lnaguage Processing 10, 381-396.

Gonnet, G.H., Baeza-Yates, R., Snider, T., 1992. New indices for text: PAT Trees and PAT arrays. In: Information retrieval: data structures and algorithms. Prentice-Hall, pp. 66-82.

Grafstein, E., Unger, B., Bullard, M., Innes, G., 2003. Canadian Emergency Department Information System (CEDIS) presenting complaint list (Version 1.0). Canadian Journal of Emergency Medicine 5, 27-34.

Graham, J., Buckeridge, D., Choy, M., Musen, M., 2002. Conceptual heterogeneity complicates automated syndronmic surveillance for bioterrorism. In: Proceedings of the AMIA Annual Symposium, p. 1030

Graham, J.R., Harvey, C.R., Rajgopal, S., 2005. The economic implications of corporate financial reporting. Journal of Accounting & Economics 40, 3-73.

Greene, W.H., 2000. Econometric Analysis. Prentice Hall, New York, NA.

Hamilton, J.D., 1989. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 57, 357-84.

Hamilton, J.D., 1994. Time Series Analysis. Princeton.

Harrison, W.T., Horngren, C.T., 2003. Financial accounting. Prentice Hall.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning. Springer, NY, NY.

Hayn, C., 1995. The information content of losses. Journal of Accounting & Economics 20, 125-153.

Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design Science in Information Systems Research. MIS Quarterly 28, 75-105.

Hinckley, D.W., 1988. Bootstrap methods. Journal of the Royal Statistical Society, Series B 50, 321-337.

Holt, C.C., 2004. Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting 20, 5-10.

Hripcsak, G., Wilcox, A., 2002. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. Journal of the American Medical Informatics Association 9, 1-15.

Hu, P., Zeng, D., Chen, H., Larson, C., Chang, W., Tseng, C., 2005. Evaluating an infectious disease information sharing and analysis systems. In: IEEE International Conference on Intelligence and Security Informatics Kantor P, Muresan G, Roberts F, Zeng D, Wang F-Y, Chen H & Merkle R (Eds.), Atlanta, Georgia.

Hu, P.J.H., Zeng, D., Chen, H., Larson, C., Chang, W., Tseng, C., Ma, J., 2007. System for Infectious Disease Information Sharing and Analysis: Design and Evaluation. Information Technology in Biomedicine, IEEE Transactions on 11, 483-492.

Hull, J.C., Basu, S., 2010. Options, futures, and other derivatives. Pearson.

Hutwagner, L., Thompson, W., Seeman, G.M., Treadwell, T., 2003. The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS). Journal of Urban Health 80, i89-i96.

Hyland, K., 1998. Hedging in Scientific Research Articles. John Benjamins Publishing Company, Amsterdam, Philadelphia.

ISDS, 2008. My Algorithm Can Out-Detect Your Algorithm: Biosurveillance Using Time Series Data. International Society for Disease Surveillance, Tech. Rep.

Ivanov, O., Wagner, M.M., Chapman, W.W., Olszewski, R.T., 2002. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. In: AMIA Symposium, pp. 345-349

Jackson, M.L., Baer, A., Painter, I., Duchin, J., 2007. A simulation study comparing aberration detection algorithms for syndromic surveillance. BMC Med Inform Decis Mak 7

Jacobson, R., Aaker, D., 1993. Myopic management behavior with efficient, but imperfect, financial markets - a comparison of information asymmetries in the United States and Japan. Journal of Accounting & Economics 16, 383-405.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. The Journal of Finance 48, 65-91.

Joachims, T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In: Machine Learning: ECML-98. pp. 137-142.

Joachims, T., 1999. Making large-scale SVM learning practical. In: Schölkopf B, Burges C & Smola A (eds.) Advances in Kernel Methods - Support Vector Learning. MIT-Press.

Kahneman, D., Tversky, A., 1979. Prospect Theory: An Analysis of Decision under Risk. Econometrica 47, 263-291.

Kim, C.J., Nelson, C.R., 1999. State-space models with regime switching. MIT Press, Cambridge.

Kormendi, R., Lipe, R., 1987. Earnings innovations, earnings persistence, and stock returns. The Journal of Business 60, 323.

Kothari, S.P., 2001. Capital markets research in accounting. Journal of Accounting & Economics 31, 105-231.

Kothari, S.P., Sloan, R.G., 1992. Information in prices about future earnings: Implications for earnings response coefficients. Journal of Accounting and Economics 15, 143-171.

Lee, N., Hui, D., Wu, A., Chan, P., Cameron, P., Joynt, G., Ahuja, A., Yung, M., Leung, C., To, K., Lui, S., Szeto, C., Chung, S., Sung, J., 2003. A major outbreak of severe acute respiratory syndrome in Hong Kong. New England Journal of Medicine 348, 1986-1994.

Leroy, G., Chen, H., 2001. Meeting medical terminology needs - The ontology-enhanced medical concept mapper. IEEE Transactions on Information Technology in Biomedicine 5, 261-270.

Lewis, M., Pavlin, J., Mansfield, J., O'Brien, S., Boomsma, L., Elbert, Y., Kelley, P., 2002. Disease outbreak detection system using syndromic data in the greater Washington DC area. American Journal of Preventive Medicine 23, 180-186.

Li, J., Su, H., Chen, H., Futscher, B.W., 2007. Optimal search-based gene subset selection for gene array cancer classification. IEEE Transactions on Information Technology in Biomedicine 11, 398-405.

Li, K.W., Yang, C.C., 2006. Conceptual Analysis of Parallel Corpus Collected From the Web. Journal of the American Society for Information Science and Technology 57, 632-644.

Livnat, J., Mendenhall, R.R., 2006. Comparing the Post Earnings Announcement Drift for Surprises Calculated from Analyst and Time Series Forecasts. Journal of Accounting Research 44, 177-205.

Lombardo, J., Burkom, H., Elbert, E., Magruder, S., Lewis, S.H., Loschen, W., Sari, J., Sniegoski, C., Wojcik, R., Pavlin, J., 2003. A system overview of the Electronic Surveillance System for Early Notification of Community-Based Epidemics (ESSENCE II). Journal of Urban Health 80, i32-i42.

Lopez, T.J., Rees, L., 2002. The effect of beating and missing analysts' forecasts in the information content of unexpected earnings. Journal of Accounting, Auditing & Finance 17, 155-184.

Lu, H.-M., Zeng, D., Chen, H., 2006. Ontology-based automatic chief complaints classification for syndromic surveillance. In: IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan.

Lu, H.-M., Zeng, D., Chen, H., 2008a. Ontology-enhanced Automatic Chief Complaint Classification for Syndromic Surveillance. Journal of Biomedical Informatics 41, 340-356.

Lu, H.-M., Zeng, D., Trujillo, L., Komatsu, K., Chen, H., 2008b. Ontology-enhanced automatic chief complaint classification for syndromic surveillance. J. of Biomedical Informatics 41, 340-356.

Ma, Z., Sheng, O.R.L., Pant, G., 2009. Discovering company revenue relations from news: A network approach. Decision Support Systems 47, 408-414.

Madigan, D., 2005. Bayesian Data Mining for Health Surveillance. In: Spatial and Syndromic Surveillance for Public Health. John Wiley & Sons, pp. 203-221.

Mandl, K.D., Overhage, M., Wagner, M., Lober, W., Sebastiani, P., Mostashari, F., Pavlin, J., Gesteland, P.H., Treadwell, T., Koski, E., Hutwagner, L., Buckeridge, D.L., Aller, R.D., Grannis, S., 2004. Implementing syndromic surveillance: a practical guide informed by the early experience. Journal of the American Medical Informatics Association 11, 141-150.

Manning, C., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

Manning, C.D., Schuze, H., 1999. Foundations of statistical natural language processing. MIT Press, Cambridge, Mass.

March, S.T., Smith, G.F., 1995. Design and natural science research on information technology. Decision Support Systems 15, 251-266.

Mas-Colell, A., Whinston , M., Green, J.R., 1995. Microeconomic theory. Oxford.

McCaig, L.F., Nawar, E.W., 2006. National hospital ambulatory medical care survey: 2004 Emergency Department Summary. Advance Data from Vital and Health Statistics 372, 1-32.

McNeman, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153-157.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153-157.

Montgomery, D.C., 2005. Introduction to statistical quality control. Wiley, Hoboken, NJ.

Moustakides, G.V., 1986. Optimal Stopping Times for Detecting Changes in Distributions. The Annals of Statistics 14, 1379-1387.

Muntermann, J., 2009. Towards ubiquitous information supply for individual investors: A decision support system design. Decision Support Systems 47, 82-92.

Mushin, I., 2001. Evidentiality and Esistemological Stance: Narrative Retelling. John Benjamins Publishing Company.

Navarro, G., 2001. A guided tour to approximate string amtching. ACM Computing Surveys 33, 31-88.

Ng, H.T., Low, J.K., 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp. 277-284, Barcelona, Spain.

Niiranen, S.T., Yli-Hietanen, J.M., Nathanson, L.A., 2008. Toward Reflective Management of Emergency Department Chief Complaint Information. Information Technology in Biomedicine, IEEE Transactions on 12, 763-767.

NIST, 2006. NIST 2006 Machine Translation Evaluation Official Results. URL http://www.itl.nist.gov/iad/894.01/tests/mt/doc/mt06eval_official_results.html.

Nuyts, J., 2001. Epistemic Modality, Language, and Conceptualization: A Cognitive-Pragmatic Perspective. John Benjamin Publishing Company.

Oard, D.W., 1996. Adaptive vector space text filtering for monolingual and cross-language applications. University of Maryland, College Park.

Olszewski, R.T., 2003. Bayesian classification of triage diagnoses for the early detection of epidemics. In: FLAIRS Conference pp. 412-416, Menlo Park, California.

Ong, T.-H., Chen, H., 1999. Updateable PAT-Tree approach to Chinese key phrase extraction using mutual information: a linguistic foundation for knowledge management. In: Proceedings of the Second Asian Digital Library Conference, Taipei, Taiwan.

Owen, C.L., 1998. Design research: building the knowledge base. Design Studies 19, 9-20.

Page, E.S., 1954. Continuous Inspection Schemes. Biometrics 41, 100-115.

Pakhomov, S.V.S., Buntrock, J.D., Chute, C.G., 2006. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. Journal of the American Medical Informatics Association 13, 516-525.

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2, 1-135.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing

Pirkola, A., Hedlund, T., Keskustalo, H., Jarvelin, K., 2001. Dictionary-based cross-language information retrieval: problems, methods, and research findings. Information Retrieval 4, 209-230.

Platt, J.C., 1999. Probabilistic outputs for support vector machines and comparison to regularize likelihood methods. In: Smola AJ, Bartlett P, Schoelkopf B & Schuurmans D (eds.) Advances in Large Margin Classifiers. pp. 61-74.

Plous, S., 1993. The psychology of judgment and decision making. Temple University Press, Philadelphia.

Popescu, C.A., Wong, Y.S., 2005. Nested Monte Carlo EM Algorithm for Switching State-Space Models. IEEE Transactions on Knowledge and Data Engineering 17, 1653-1663.

Qin, J., Zhou, Y., Chau, M., Chen, H., 2006. Multilingual Web Retrieval: An Experiment in English-Chinese Business Intelligence. Journal of the American Society for Information Science and Technology 57, 671-683.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., 1985. A comprehensive grammar of the English language. Longman, New York.

Reis, B.Y., Mandl, K.D., 2003a. Integrating Syndromic Surveillance Data across Multiple Locations: Effects on OUtbreak Detection Performance. In: AMIA 2003 Symposium Proceedings, pp. 549-553

Reis, B.Y., Mandl, K.D., 2003b. Time series modeling for syndromic surveillance. BMC Med Inform Decis Mak 3

Reis, B.Y., Pagano, M., Mandl, K.D., 2003. Using temporal context to improve biosurveillance. Proc. Natl. Acad. Sci. U.S.A. 100, 1961-1965.

Riloff, E., Wiebe, J., 2003. Learning extraction patterns for subjective expressions. In: Conference on Empirical Methods in Natural Language Processing (EMNLP-03), pp. 105-112

Rizomilioti, V., 2006. Exploring epistemic modality in academic discourse using corpora In: van Lier L (ed.) Information Technology in Languages for Specific Purposes. Springer.

Roberts, S.W., 1966. A comparison of some control chart procedures. Technometrics 8, 411-430.

Rolka, H., Burkom, H., Cooper, G.F., Kulldorff, M., Madigan, D., Wong, W.-K., 2007. Issues in applied statistics for public health bioterrorism surveillance using multiple data stream: research needs. Statistics in Medicine 26, 1834-1856.

Rosse, C., Mejino, J.L., Modayur, B.R., Jakobovits, R., Hinshaw, K.P., Brinkley, J.F., 1998. Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base. Journal of the American Medical Informatics Association 5, 17-40.

Rubin, V.L., 2004. Certainty categorization model. In: AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications, Stanford, CA.

Rubin, V.L., 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In: Proceedings of NAACL HLT 2007, pp. 141-144

Rubin, V.L., Liddy, E.D., Kando, N., 2005. Certainty identification in texts: Categorization model and manual tagging results. In: Shanahan JG, Qu Y & Wiebe J (eds.) Computing Attitude and Affect in Text: Theory and Applications. Springer-Verlag.

Sakai, T., 2000. MT-based Japanese-English cross-language IR experiments using the TREC test collections. In: Fifth International Workshop on Information Retrieval with Asian Language, Hong Kong, China.

Sarawagi, S., 2008. Information extraction. Foundations and Trends in Database 1, 261-377.

Savage, L., 1954. The foundations of statistics. Wiley, New York.

Schneider, D., Appleton, L., McLemore, T., 1979. A reason for visit classification for ambulatory care. National Center for Health Statistics. Vital Health Statistics 2, 1-63.

Scholkopf, B., Mika, S., Burges, C., Knirsch, P., Muller, K., Atsch, G., Smola, A., 1999. Input space versus feature space in kernel-based methods.

Schwarz, G., 1978. Estimating the Dimension of a Model. Annals of Statistics 6, 461-464.

Scott, S.L., 2002. Bayesian methods for hidden Markov models: recursive computing in the 21st century. Journal of the American Statistical Association 97, 337-351.

Serfling, R.E., 1963. Methods for Current Statistical Analysis of Excess Pneumonia-Influenza Deaths. Public Health Reports 78, 494–-506.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica Sinica 7, 221-264.

Shapiro, A.R., 2004. Taming variability in free text: Application to health surveillance. In: Syndromic Surveillance: Report from a National Conference, 2003, 95-100.

Shewhart, W.A., 1939. Statistical method from the viewpoint of quality control. Washington, The Graduate School, The Department of Agriculture.

Shiryaev, A.N., 1963. On optimum methods in quickest detection problems. Theory of Probability and Its Applications 8, 22-46.

Skinner, D., Sloan, R., 2002. Earnings surprises, growth expectations, and stock returns or don't let an earnings torpedo sink your portfolio. Review of Accounting Studies 7, 289-312.

Slywotzky, A.J., Drzik, J., 2005. Countering the biggest risk of all. Harvard Business Review, 1-11.

Sniegoski, C.A., 2004. Automated syndromic classification of chief complaint records. Johns Hopkins APL Technical Digest 25, 68-74.

Sonesson, C., 2003. Evaluations of some Exponentially Weighted Moving Average methods. Journal of Applied Statistics 30, 1115-1133.

Sonesson, C., Book, D., 2003. Review and discussion of prospective statistical surveillance in public health. Journal of the Royal Statistical Society, Series A 166, 5-21.

Song, X., Wu, M., Jermaine, C., Ranka, S., 2007. Conditional Anomaly Detection. IEEE Transactions on Knowledge and Data Engineering 19, 631-645.

Sproat, R., Shin, C.L., 1990. A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese & Oriental Lnaguages 4, 336-351.

Steiner, S.H., 1999. EWMA control charts with time-varying control limits and fast initial response. Journal of Quality Technology 31, 75-86.

Storey, V.C., Burton-Jones, A., Sugumaran, V., Purao, S., 2008. CONQUER: A methodology for context-aware query processing on the World Wide Web. Information Systems Research 19, 3-25.

Strat, Y.L., Carrat, F., 1999. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in Medicine 18, 3463-3478.

Takeda, H., Veerkamp, P., Tomiyama, T., Yoshikawam, H., 1990. Modeling design processes. AI Magazine 11, 37-48.

Takeuchi, J., Yamanishi, K., 2006. A Unifying Framework for Detecting Outliers and Change Points from Time Series. IEEE Transactions on Knowledge and Data Engineering 18, 482-492.

Talvensaari, T., Juhola, M., Laurikkala, J., Jarvelin, K., 2007. Corpus-based cross-language information retrieval in retrieval of highly relevant documents. Journal of the American Society for Information Science and Technology 58, 322-334.

Teahan, W.J., Wen, Y., McNab, R., Witten, I.H., 2001. A compression-based algorithm for Chinese word segmentation. Computational Linguistics 26, 375-393.

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. Journal of Finance 62, 1139-1168.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. Journal of Finance 63, 1437-1467.

Thompson, D.A., Eitel, D., Fernandes, C.M.B., Pines, J.M., Amsterdam, J., Davidson, S.J., 2006. Coded chief complaints - automated analysis of free-text complaints. Academic Emergency Medicine 13, 774-782.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58, 267-288.

Tolle, K.M., Chen, H., 2000. Comparing noun phrasing techniques for use with Medical Digial Library Tools. Journal of the American Medical Informatics Association 51, 352-370.

Travers, D., 2003. Identification of concepts from emergency department text using natural language processing techniques and the unified medical language system. University of North Carolina at Chapel Hill.

Travers, D.A., Haas, S.W., 2003. Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. Journal of Biomedical Informatics 36, 260-270.

Travers, D.A., Haas, S.W., 2004. Evaluation of emergency medical text processor, a system for cleaning chief complaint textual data. Academic Emergency Medicine 11, 1170-1176.

Tsui, F.-C., Espino, J.U., Dato, V.M., Gesteland, P.H., Hutman, J., Wagner, M.M., 2003. Technical description of RODS: a real-time public health surveillance system. Journal of the American Medical Informatics Association 10, 399-408.

Turney, P., 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Lingusitics, pp. 417-424, Philadelphia, Pennsylvania.

van Rijsbergen, C.J., 1979. Information retrieval, Butterworths, London.

Vapnik, V., 1995. The nature of statistical learning theory. Springer-Verlag.

Von Neumann, J., Morgenstern, O., 1944. Theory of games and economic behavior. Princeton University Press, Princeton.

Wang, F.L., Yang, C.C., 2007. Mining web data for Chinese segmentation. Journal of the American Society for Information Science and Technology 58, 1820-1837.

Wang, J., 2006. Automatic thesaurus development: term extraction from title metadata. Journal of the American Society for Information Science and Technology 57, 907-920.

Wein, L.M., Craft, D.L., Kaplan, E.H., 2003. Emergency response to an anthrax attack. Proc. Natl. Acad. Sci. U.S.A. 100, 4346-4351.

White, H., 2006. Approximate nonlinear forecasting methods. In: Elliott G, Granger C & Timmermann A (eds.) Handbook of Economic Forecasting. Elsevier, pp. 459-512.

Wiebe, J., 2000. Learning subjective adjectives from corpora. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. AAAI Press / The MIT Press

Wiebe, J., Bruce, R., Bell, M., Martin, M., Wilson, T., 2001. A corpus study of evaluative and speculative language. In: Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.

Wiebe, J., Wilson, T., Cardie, C., 2005. Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation 39, 165-210.

Wiebe, J.M., Bruce, R.F., O'Hara, T.P., 1999. Development and use of a gold-standard data set for subjectivity classifications. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, College Park, Maryland.

Wieland, S.C., Brownstein, J.S., Berger, B., Mandl, K.D., 2007. Automated real time constant-specificity surveillance for disease outbreaks. BMC Medical Informatics and Decision Making 7

Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis. Computational Linguistics 35, 399-433.

Winters, P.R., 1960. Forecasting Sales by Exponentially Weighted Moving Averages. MANAGEMENT SCIENCE 6, 324-342.

Witten, I.H., Frank, E., 2005. Data mining: Practical machine learning tools and techniques. Elsevier.

Wu, T.-S., 2005. Establishing Emergency Department-Based Infectious Disease Syndromic Surveillance System in Taiwan–Aberration Detection Methods, Epidemiological Characteristics, System Evaluation and Recommendations. National Taiwan University.

Wu, T.-S., Shih, F.-Y., Yen, M.-Y., Wu, J.-S., Lu, S.-W., Chang, K., Hsiung, C., Chou, J.-H., Chu, Y.-T., Chang, H., Chiu, C.-H., Tsui, F.-C., Wagner, M., Su, I.-J., King, C.-C., 2008. Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan. BMC Public Health 8, 18.

Wu, Z., Tseng, G., 1993. Chinese text segmentation for text retrieval: achievements and problems. Journal of the American Society for Information Science 44, 532-542.

Yan, P., Chen, H., Zeng, D., 2006. Syndromic surveillance systems: public health and biodefence. Annual Review of Information Science and Technology forthcoming

Yan, P., Chen, H., Zeng, D., 2008. Syndromic surveillance systems: public health and biodefence. Annual Review of Information Science and Technology 42

Yang, C.C., Luk, J.W.K., Yung, S.K., Yen, J., 2000. Combination and boundary detection approaches on Chinese indexing. Journal of the American Society for Information Science 51, 340-351.

Young, P.C., Tippins, S.C., 2001. Managing business risk: An organization-wide approach to risk management. AMACOM, New York.

Yu, H., Hatzivassiloglou, V., 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 conference on Empirical methods in natural language processing - Volume 10. Association for Computational Linguistics

Zhang, J., Tsui, F.-C., andWilliam R. Hogan, M.M.W., Detection of outbreaks from time series data using wavelet transform. In: Proc AMIA Symp

Zhang, T., Oles, F.J., 2001. Text categorization based on regularized linear classification methods. Information Retrieval 4, 5-31.

Zhou, L., Burgoon, J.K., Twitchell, D.P., Qin, T., Nunamaker, J.F., 2004. A comparison of classification methods for predicting deception in computer-medicated communication. Journal of Management Information Systems 20, 139-165.

Zobel, J., Dart, P., 1996. Phonetic string matching: lessons from information retrieval. In: Proceedings of the 19th International Conferences on Research and Development in Information Retrieval, Zurich, Switzerland.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society B 67, 301-320.