ENTITY EXTRACTION, ANIMAL DISEASE-RELATED EVENT RECOGNITION AND CLASSIFICATION FROM WEB

by

SVITLANA VOLKOVA

B.S., Petro Mohyla State University, Ukraine, 2004
M.S., Petro Mohyla State University, Ukraine, 2006

A THESIS

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Computing and Information Sciences
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2010

Approved by:

Major Professor
William H. Hsu

Abstract

Global epidemic surveillance is an essential task for national biosecurity management and bioterrorism prevention. The main goal is to protect the public from major health threads. To perform this task effectively one requires reliable, timely and accurate medical information from a wide range of sources. Towards this goal, we present a framework for epidemiological analytics that can be used to extract and visualize infectious disease outbreaks from the variety of unstructured web sources automatically. More precisely, in this thesis, we consider several research tasks including document relevance classification, entity extraction and animal disease-related event recognition in the veterinary epidemiology domain. First, we crawl web sources and classify collected documents by topical relevance using supervised learning algorithms. Next, we propose a novel approach for automated ontology construction in the veterinary medicine domain. Our approach is based on semantic relationship discovery using syntactic patterns. We then apply our automatically-constructed ontology for the domain-specific entity extraction task. Moreover, we compare our ontology-based entity extraction results with an alternative sequence labeling approach. We introduce a sequence labeling method for the entity tagging that relies on syntactic feature extraction using a sliding window. Finally, we present our novel sentence-based event recognition approach that includes three main steps: entity extraction of animal diseases, species, locations, dates and the confirmation status n-grams; event-related sentence classification into two categories - suspected or confirmed; automated event tuple generation and aggregation. We show that our document relevance classification results as well as entity extraction and disease-related event recognition results are significantly better compared to the results reported by other animal disease surveillance systems.

Table of Contents

List of Figures

List of Tables

Acknowledgments

It is a pleasure to thank those who made this thesis possible including Dr. William Hsu, Dr. Doina Caragea, Dr. Gurdip Singh, my former professors from Mykolayiv Black Sea State University Dr. Oleksandr Trunov and Dr. Yuriy Kondrateko, my family and friends.

I want to thank my adviser Dr. William Hsu for guiding the research presented in this thesis. I appreciate the opportunity to participate in NABC project. I also thank him and National Agriculture Biosecurity Center for providing funding for my research.

I owe my deepest gratitude to Dr. Doina Caragea for her support and guidance. She has always been a constant source of motivation and encouragement for me. I appreciate her time and effort helping me with research papers and giving a valuable feedback. I am grateful to her for writing the recommendation letters for my PhD applications and other scholarships. I dedicate my acceptance to a PhD program at Johns Hopkins University and my Google Anita Borg Memorial Scholarship to her. I am also thankful to her for organizing the "Semi-supervised Learning and Domain Adaptation" seminar which brought up extremely useful discussions and helped me to advance my knowledge of machine learning.

I wish to express my warm and sincere thank to Dr. Gurdip Singh. He was always accessible and advised me a lot during his courses. I am grateful to him for our fruitful interactions, his constant support and guidance.

My warmest thanks are due to Dr. Yuriy Kondratenko and Dr. Oleksandr Trunov for their encouragements and helping me believe that I was able to complete this thesis.

I am thankful to my parents Valentina Volkova and Aleksandr Volkova, and my sister Anna Volkova. The most special thanks goes to Pavlo Bohutskyi for his extreme patience, unconditional support and love through all this process.

I also thank staff members in Computing and Information Sciences Department, especially Jodi Milliner, Roberta Hodges and Jan Herndon. It has been an honor for me to be part of the Computing and Information Sciences Department at Kansas State University.

Dedication

To the most important people in my life:
my parents Valentina Volkova and Oleksandr Volkov

Chapter 1

Introduction

1.1 Background

The large spread of infectious diseases has a great negative impact on society. While human infectious diseases can result in significant loss of life, animal diseases can cause major problems across the world because of the influence on the economy[78] and trade[48]. Moreover, animal diseases that are zoonotic in type can also cause loss of life in addition to economic crises and political instability.

To conform to national security regulations, officials need an efficient way to determine what threats can potentially affect the health and welfare of the citizens, especially in light of recently increased concerns about bioterrorism. For that purpose, Infectious Disease Informatics (IDI) studies tasks such as: data collection, sharing, management, modeling and analysis in the domain of emerging infectious diseases[18],[84].

1.2 Motivation

An enormous amount of data about animal infectious disease-related events is available online in both structured and unstructured formats. Structured data is presented to public in official reports by different organizations such as: state and federal laboratories, local health care providers, governmental agricultural or environmental agencies. In addition, a lot of unstructured information can be found in a variety of other contexts *e.g.,* news, e-

mails, blogs, which in contrast to the official reports is completely unorganized. In order to exploit this unstructured data, machine learning and text mining techniques can be used to recognize disease-related events, *e.g., "On 12 September 2007, a new foot-and-mouth disease outbreak was confirmed in Egham, Surrey"*. Such techniques could be part of automated systems that can detect, monitor and track responses to animal infectious disease outbreaks (defined as a set of events which are constrained in space and have temporal overlap)[23].

At the *Knowledge Discovery in Databases Laboratory*, we have developed an intelligent assistive framework for tracking animal infectious disease-related events. This project was funded by the *National Agricultural Biosecurity Center (NABC)*[1].

There are several subtasks that we have completed during our NABC project. The first subtask is web crawling for animal disease-related data collection. The second subtask is animal disease relevance document classification. The next subtask is search of the indexed collection. The fourth subtask is data analysis which includes entity extraction, animal disease-related event recognition and classification. The last subtask is visualization of the automatically extracted events on a map using *GoogleMaps*[2] and within a timeline using *SIMILE*[3].

Technically, this thesis is concerned with the main logic of the developed framework which includes several subtasks from the above listed such as:

- the classification of the disease-related documents collected from different domains;
- domain-specific entity extraction (animal disease names, viruses, disease serotypes);
- automated animal disease-related event recognition and classification from unstructured web data.

[1]NABC - `http://nabc.ksu.edu/content/`
[2]GoogleMaps API - `http://code.google.com/apis/maps/`
[3]SIMILE API - `http://www.simile-widgets.org/timeline/`

1.3 Problem Statement

Suppose we have a set of documents D and a collection of sources of information C (*e.g.,* news, web-pages, scientific papers, medical literature, e-mails *etc.*). Every document d_i belongs to only one source c_j (many-to-one relation). First, we need to classify documents from D into two classes such as: disease-related D^R and disease non-related D^{NR} documents, where $D^R \cup D^{NR} \equiv D$. *A disease-related web-document* is a document that reports at least one emergent or non-emergent animal disease-related event.

We aim to extract structured information about any animal disease-related events from each piece of d_i in each source c_j, where $d_i \in D^R$, $c_j \in C$. Since we consider documents from different sources, we are looking for several event types including:

- Type 1: *Emergent animal disease-related events*, that happened recently in a short period of time *e.g., "On Jun 2, 2010, a total of 35 individuals were infected with a matching strain of salmonella, serotype newport".*

- Type 2: *Non-emergent animal disease-related events*, that happened long time ago *e.g., "The US saw its latest FMD outbreak in Montebello, California in 1929".*

Moreover, we need to be able to recognized and eliminate *animal disease non-related events* of the following types:

- Type 3: *Disease-related events that are not related to outbreaks e.g., "A meeting on foot and mouth disease was held in Brussels on Oct 17, 2007"* or *e.g., "A team at Peking University in Beijing studies tissue taken from people killed by H5N1 in China".*

- Type 4: *Hypothetical animal disease-related events, e.g., "30 million people could die if a human-to-human strain of bird flu spreads over the nation".*

- Type 5: *Negation of the animal disease-related events, e.g., "Samples from the farm in Romania have revealed no case of bird flu".*

The structured information that we want to extract about animal disease-related events includes *domain-specific* and *domain-independent* Named Entities (NE) such as:

- disease names (*e.g., "foot and mouth disease", "rift valley fever")*;
- viruses (*e.g., "picornavirus")* and serotypes (*e.g., "Asia-1")*;
- species (*e.g., "sheep", "pigs", "cattle")*;
- locations of events specified at different levels of geo-granularity (*e.g., "United Kingdom", "eastern provinces of Shandong and Jiangsu, China")*;
- dates in different formats (*e.g., "last Tuesday", "two month ago").*

Finally, we need to classify automatically-extracted animal disease-related events into two categories such as:

- confirmed, *e.g., "On 9 Jun 2009, the farm's owner reported symptoms of FMD in more than 30 hogs";*
- suspected, *e.g., "RVF is suspected in Saudi Arabia in September 2000".*

1.4 Significance of the Study

The design and development of a framework for epidemiological analytics requires resolving several challenging research tasks: *disease-related document classification, domain-specific entity extraction, event recognition and classification* as discussed in details in Chapter 3.

Text classification methodology has been studied extensively (*e.g.,* 20-Newsgroups, Reuters-21578 are typical datasets). However, there are no such datasets and works related to text classification in the domain of veterinary medicine. Therefore, in Chapter 4 we investigate state-of-the-art machine learning techniques for text classification in the context of the veterinary medicine domain in order to facilitate the design of our framework.

Information extraction has been studied by numerous researches, but it remains to be a challenging problem. In Chapter 5 we suggest our novel approach for an automated domain-specific ontology construction by learning semantic relations between ontology concepts using syntactic pattern matching. We then apply our automatically-learned ontology for the domain-specific entity extraction. Moreover, we suggest an alternative methodology for the entity extraction that is based on sequence labeling by extracting syntactic features using a sliding window approach. Both these approaches outperform all existing systems that report information extraction or named entity recognition results in the domain of veterinary medicine.

Event recognition and classification has not been studied well, especially in the domain of veterinary medicine. Despite of some research projects related to political event recognition from unstructured web documents at The Cline Center for Democracy at the University of Illinois, to our knowledge, there are no state-of-the-art approaches or universally applied methodologies for event recognition. In Chapter 6 we suggest our novel approach for animal disease-related event recognition. Moreover, we apply our event recognition approach to extract structured information in the predictive epidemiology domain.

1.5 Outline

The rest of the thesis is organized as follows:

Chapter 2: We give an overview of the web resources that report infectious diseases outbreaks. We present systems which are manually maintained by state and federal governmental agencies and discuss automated animal disease surveillance web interfaces. We describe approaches for text categorization, state-of-the-art methodologies for entity and relation extraction, and animal disease-related event recognition.

Chapter 3: We present an overall description of the framework for epidemiological analytics and its main functionality including web-crawling, information extraction and event recognition components.

Chapter 4: We introduce our supervised framework for document relevance classification in the veterinary medicine domain. We discuss different feature representations for the documents collected from multiple sources and various machine learning algorithms. We perform an experiment for disease relevance document classification using different feature representations and classification algorithms.

Chapter 5: To address the lack of a veterinary medicine ontology, we first manually build a set ontologies and expand the initial ontology with semantic relationships (synonymic, hyponymic and causal) identified using syntactic patterns and part of speech tagging. We then show how to use these semantic relationships for expansion of the manually constructed ontology and automatically construct new ontology. We present an overview for the biomedical entity extraction task for the domain of the veterinary medicine using an animal disease example. We discuss the results of biomedical entity extraction using manually *vs.* automatically-constructed ontologies. Moreover, we suggest an alternative sequence modeling approach for the entity extraction. We extract syntactic features using sliding window approach. We then report our experimental results for the entity extraction approach based on the syntactic features.

Chapter 6: We present our novel sentence-based methodology for disease-related event recognition in the domain of veterinary medicine. We first discuss the entity recognition phase; we then describe the event sentence classification phase; finally, we demonstrate the event tuple generation and aggregation phase. The experimental results for our event recognition approach confirm the feasibility of the proposed approach. Moreover, we apply our event recognition approach to a specific classification task in the predictive epidemiology.

Chapter 7: We conclude with a summary, a list of contributions and future work directions.

Chapter 2

Related Work

2.1 Monitoring Systems in Veterinary Epidemiology

2.1.1 Manually-Supported Web Interfaces

There are several manually-supported web-system for animal disease outbreak monitoring and reporting at international level including:

- *The World Organization for Animal Health (OIE)*[1] is the one of the most important sources that report about animal health situations at international level using *The World Animal Health Information Database (WAHID) Interface*[2].

- *The World Health Organization (WHO)*[3] provides users with an interactive information mapping system - *The WHO Global Atlas of Infectious Diseases*[4].

- *The Animal Production and Health Division at Food and Agricultural Organization of United Nations*[5] allows monitoring infectious disease outbreaks within a map and timeline view using *The Emergency Prevention System (EMPRES) for Transboundary Animal and Plant Pests and Diseases*[6].

[1]OIE - `http://www.oie.int/eng/en_index.htm`
[2]WAHID Interface - `http://www.oie.int/wahis/public.php?page=home`
[3]WHO - `http://www.who.int/en`
[4]WHO Atlas Interface - `http://diseasemaps.usgs.gov/index.htm`
[5]FAO - `http://www.fao.org/ag/againfo/home/en/index.htm`
[6]EMPRES - `http://www.fao.org/EMPRES/default.html`

– *The Department for Environment Food and Rural Affairs (DEFRA)*[7] provides users with consistent information about animal health and welfare in United Kingdom.

Many systems monitor situation about animal disease outbreaks at the country and state level in the United States:

– *The U.S. Department of Agriculture (USDA)*[8] manages a data system for animal diseases (*e.g., foot and mouth disease, rift valley fever*);

– *The U.S. Geological Survey (USGS)* administers a database for wildlife diseases through its *National Wildlife Health Center (NWHC)*[9];

– *Centers for Disease Control and Prevention (CDC)*[10] provide users with data about infectious diseases;

– *Iowa State University Center for Food Security and Public Health (CFSPH)*[11] website supplies users with information about infectious animal diseases, vaccines, disease fact sheets, image databases for diseases, and other useful resources for producers and veterinarians.

Several biological portals that are manually curated by research agencies and universities are also available online:

– *Foot-and-mouth disease (FMD) BioPortal*[12] is developed for global FMD surveillance based on news monitoring and maintained by *FMD Surveillance and Modeling Laboratory at the University of California UC Davis. FMD BioPortal* uses crawlers that regularly collect FMD-related news from the Internet. Relevant news are stored in the database after keyword-based filtering from a large document collection[68].

[7]DEFRA - `http://www.defra.gov.uk`
[8]USDA - `http://www.usda.gov/wps/portal/usdahome`
[9]NWHC - `http://www.nwhc.usgs.gov`
[10]CDC - `http://www.cdc.gov`
[11]CFSPH - `http://www.cfsph.iastate.edu`
[12]FMD BioPortal - `https://fmdbioportal.ucdavis.edu`

– *BioSurveillance Portal* at the University of Arizona, maintained by its *Artificial Intelligence Laboratory*[13] is a web-based *Infectious Disease Informatics (IDI)* system that provides access to distributed health data for several major infectious diseases;

In addition, there are several specific online resources for highly pathogenic animal diseases, *e.g.,* the *Reference Laboratories Information System*[14] for the *OIE/FAO Foot-and-Mouth Disease Reference Laboratories Network*. The necessity of human analysis and manual/semi-automated maintenance is a major drawback of the above discussed online systems for animal disease outbreaks tracking.

2.1.2 Automated Web Services

The BioCaster Global Health Monitor[15] is an online web-based system for detecting and mapping infectious disease outbreaks from news[26]. The system follows 1500 RSS feeds hourly that deal with a taxonomy of 4300 named entities (50 disease names, 243 country names, 4025 province/city names, and latitudes and longitudes for all locations). It is able to provide information on about 40 infectious diseases at up to 25-30 locations per day. *BioCaster Global Health Monitor* provides functionality such as: multilingual information extraction from news limited to English, French, Spanish, Chinese, Thai, Vietnamese, Japanese; their classification of documents as topically relevant or not; and plotting events on a GoogleMap[57],[37].

HealthMap[16] aggregates articles from *GoogleNews*[17] and *ProMED-Mail*[18] portal. It is a manually maintained Internet-based system that publishes reports generated by public health experts. The system allows tracking infectious diseases and locations related to outbreaks. It covers 2300 locations and 1100 disease names and identifies between 20-30 outbreaks per day. Since *HealthMap* is manually supported system, it supports processing

[13]BioPortal - `http://biocomputingcorp.com/bpsystem.html`
[14]ReLaIS - `http://www.foot-and-mouth.org`
[15]BioCaster - `http://biocaster.nii.ac.jp/`
[16]HealthMap - `http://healthmap.org/en`
[17]GoogleNews - `http://news.google.com/`
[18]ProMED - `www.promedmail.org`

text in multiple languages such as: English, French, Spanish, Portuguese, Russian, Chinese, Arabic[28].

The information retrieval system *MedISys*[19], supported by the European Union is a part of the Europe Media Monitor (EMM)[20] product family, and was developed for searching web-based resources and producing quantitative summaries of the latest epidemics reports. This system includes the information extraction subsystem (the *Pattern-based Understanding and Learning System, PULS)*[21] that allows automated recognizing of the metadata and structured facts related to the disease outbreaks in text. *MedISys* currently collects an average 50000 news articles per day from about 1400 news portals from commercial news providers and from about 150 specialized Public Health sites. Moreover, *MedISys* allows data aggregation from multiple sources approximately on 43 languages about health-related topics such as: epidemics, nuclear, chemical/radiological, bio-terrorism, *etc.* The current ontology contains 2400 disease names, 400 organisms, 1500 political entities and over 70000 location names including towns, cities, provinces. During the information retrieval phase, the system performs real-time news clustering and filtering by matching 3000 patterns (*e.g.,* multi-word terms and their combinations), then classifies sources into 750 categories. During the information extraction phase, additional metadata is extracted such as: language, source country, download time, source site *etc.* from documents previously converted to Unicode[67].

The main advantage of *EpiSpider*[22] is the ability to combine emerging infectious disease data from *ProMED-Mail* with similar information from other sites *e.g., The Global Disaster Alert Coordinating System (GDACS)*[23]. In addition, *EpiSPIDER* extracts this information from the Central Intelligence Agency (CIA) Factbook[24] and the United Nations Human Development Report[25] sites.

[19]MedISys - `http://medusa.jrc.it/medisys/homeedition/all/home.html`
[20]EMM - `http://emm.jrc.it/overview.html`
[21]PULS - `http://sysdb.cs.helsinki.fi/puls/jrc/all`
[22]EpiSpider - `http://www.epispider.org/`
[23]GDACS - `www.gdacs.org`
[24]CIA - `https://www.cia.gov/library/publications/the-world-factbook/`
[25]UNDHDR - `http://hdr.undp.org/en`

The main differences between these abovementioned intelligent systems and our framework for epidemiological analytics include:

1. the system purpose - disease surveillance *vs.* research or epidemiological analytics;

2. targeted audience - public *vs.* domain experts and analysts;

3. processed data - news *vs.* medical literature, blogs, e-mails, scientific papers *etc.*.

Table 2.1: The comparison of automated and semi-automated systems for animal disease outbreak monitoring

	BioCaster	HealthMap	MedISys+PULS	KDD System
Year	2007	2007	2007	2010
Country	Japan	USA	European Union	USA
Mined Sources	1500 News Feeds	Google News, ProMED-Mail	1400 news portals + 150 Public Health sites	Customized predefined set of seeds by domain experts
Productivity	25-30 locations on 40 diseases per/day	20-30 outbreaks per day	50,000 news articles per day	Future Work: Set up the schedule for crawling
Supported Languages	English, French, Spanish, Chinese, Thai, Vietnamese, Japanese	English, French, Spanish, Portuguese, Russian, Chinese, Arabic	43 languages	Future work: "Wikification" for the multilingual IE/IR
Geographical Entities	243 countries, 4,025 sub-countries (provinces, cities)	2,300 locations	70,000 locations (towns, cities, provinces)	> million locations from NGA GEOnet Names Database
Domain-specific Entities	50 diseases (ontology with synonyms, symptoms)	1100 diseases	2400 animal + human disease names, 400 organisms, 1500 political entities	Automatically-constructed ontology with > 1000 animal diseases, viruses, serotypes

2.2 Document Classification

Given a set of documents D^{train} and a set of classes S, such as each document $d_i^{train} \in D$ is labeled with a class $s_j \in S$, and taking a test document d_i^{test} from D^{test} ($D^{train} \cup D^{test} = \emptyset$), we want to predict the label of d_i^{test}. This is called supervised classification task because the labels for the instances, in our case documents in the training set, are known[43],[15]. When the labels are not available for the instances, the task becomes unsupervised text classification, also called clustering[5],[8]. In the semi-supervised text classification only part of the instances are labeled, usually there is a small set of labeled data and a lot of unlabeled data[9],[10].

The text categorization task has been extensively studied previously[66],[81]. In the supervised learning framework each document is represented as a feature vector $< w_{1,d}, w_{2,d}, ..., w_{n,d} >$, where n is size of the vocabulary of the document collection. This representation is called "bag-of-words" representation. The feature representation[7],[6] can be binary (0/1), based on term frequency (TF) or term frequency-inverse document frequency (TF-IDF) as shown in Equations 2.1, 2.2, 2.3 respectively[2],[11].

$$w_{i,d} = \begin{cases} 0, & \text{if } c(d, t_i) = 0 \\ 1, & \text{if } c(d, t_i) > 0, \end{cases} \tag{2.1}$$

where $c(d, t_i)$ is the number of time term t_i occurs in document d.

$$w_{i,d} = \begin{cases} 0, & \text{if } c(d, t_i) = 0 \\ c(d, t_i), & \text{if } c(d, t_i) > 0, \end{cases} \tag{2.2}$$

where term frequency $c(d, t_i)$ can also be normalized by the total number of occurrences of the terms in the document $c(d, t)$.

The TF-IDF representation takes into account the importance of the single significant term, for instance the term occurring frequently in the document, but rarely in the rest of the collection is given high weight.

The TF-IDF representation of the features (words) in the document is denoted as:

$$w_{i,d} = tf_i \cdot \log \frac{N}{df_i} = \frac{\sum_{d_i \in D} c(d, t_i)}{\sum_{d_i \in D, i} c(d, t_i)} \cdot \log \frac{N}{d_i | d_i \in D, c(d, t_i) > 0|}, \tag{2.3}$$

where the first component represents the term frequency and the second component - inverse document frequency, where the *log* is used to dampen the effect of TF-IDF relative to TF.

The abovementioned "bag-of-words" representation is usually sufficient for text categorization. However, there are other more advanced approaches such as: n-gram model for text classification which shows that 2-grams and 3-grams improve the classification, but this is not true for longer n-grams[29], using noun phrases as terms[40], part-of-speech tagging (POS)[4], bag-of-concepts[65].

In supervised learning, a classifier builds the model using the training data D^{train}. There are many of classifiers that have been used for supervised document classification such as: Naive Bayes (NB)[16], Support Vector Machines (SVM)[14], k-Nearest Neighbors (kNN)[59] and others[3],[1].

In this work, we aim to evaluate the classification accuracy for our domain-specific documents collected from multiple domains (*e.g.,* news, e-mails, papers) when using different feature representations ("bag -of-words" unigrams, bigrams, term frequency) in Chapter 4.

2.3 Entity Extraction

Entity extraction, also called Named Entity Recognition (NER), is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories, such as:

- person names (*e.g., "Bill Ball", "Mr. Smith"*),
- organizations (*e.g., "Apple, "IBM Inc."*),
- locations (*e.g., "China", "New York"*),
- expressions of times (*e.g., "June 20 2010", "last month"*),

– quantities (*e.g., "13", "one thousand"*) *etc.*

The list of predefined categories can be extended to include the specific knowledge for the domain of veterinary medicine such as:

– animal diseases, synonyms and abbreviations (*e.g., "Brucellosis", "Bang's disease"*),

– disease serotypes and corresponding viruses (*e.g.,"B. melitensis"*),

– corresponding species (*e.g., "sheep", "goat"*).

Let us consider the NER task in details, existing methodologies and state-of-the-art approaches. For instance, given the unstructured sentence about animal disease outbreak: *"The US saw its latest FMD outbreak in Montebello, California in 1929 where 3,600 pigs were slaughtered",* the NER system produces the annotated output such as:

"The $US_{[LOC]}$ *saw its latest* $FMD_{[DIS]}$ *outbreak in* $Montebello_{[LOC]}$, $California_{[LOC]}$ *in* $1929_{[DT]}$ *where 3,600* $pigs_{[SP]}$ *were slaughtered"*, where $E_{[DIS]}$ denotes the disease name entity, $E_{[LOC]}$ - location entity, $E_{[DT]}$ - date entity and $E_{[SP]}$ - species entity.

Various methods have been used for the named entity recognition. The earliest approaches such as: gazetteer and regular expressions are still commonly used for the domain-specific entity extraction[82]. The limitations of using these approaches include continuous manual support. The dictionary look-up methods achieve high precision, but low recall due to the limited to the size of the dictionary.

Other approaches such as Hidden Markov Models (HMM)[85] and Conditional Random Fields (CRF)[39], based on automatically learned patterns, give much better results in comparison to dictionary look-up methods. The CRF approach is implemented in the CRF project[26]. Another toolkit called Learning Based Java (LBJ) [27] also achieves high accuracy for the NER task, similar to CRF[61].

[26]CRF Project - `http://crf.sourceforge.net/`
[27]LBJ - `http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=LBJ`

A comprehensive list of existing NER systems including Stanford NER System[28], CMU Lemur Toolkit[29], Open NLP[30] is summarized by William Hsu et al.[33].

In addition to general named entities, there are works in domain-specific biomedical entity extraction that deal with human diseases, gene and protein extraction: dictionary-based bio-entity name recognition in biomedical literature[82], protein name recognition using gazetteer[69], and gene-disease relation extraction[79]. All these methods are based on static dictionaries for entity extraction, that limit the recall of the system by the size of the dictionary. There is a more effective method based on conditional random fields that has been applied for identifying gene and protein mentions in text[46]. This approach requires annotated training corpora for learning, which is not available for the veterinary medicine domain yet.

Furthermore, there are several emergency surveillance systems that perform automated extraction of animal disease names from web documents described in Section 2.1.2.

- *BioCaster* is limited to 50 animal diseases and uses manually constructed multilingual ontology[21]. It uses support vector machines to extract entities including animal diseases, synonyms[45], viruses and agents[57].

- *Pattern-based Understanding and Learning System (PULS)*[67] and *HealthMap*[28] extract as high as 2400 and 1100 disease names respectively (both human and animal diseases). They both are based on dictionary look-up approach and do not recognize any other disease related concepts such as causative viruses or disease serotypes.

In this work, we aim to perform entity extraction in the domain of veterinary epidemiology in Chapter 5. Our goal is to improve the accuracy of the domain-specific entity extraction, including animal disease names, their synonyms, abbreviations and corresponding viruses, in order to boost the accuracy of disease-related event recognition task.

[28]Stanford NER - `http://nlp.stanford.edu/ner/index.shtml`
[29]Lemur Toolkit - `http://www.lemurproject.org/`
[30]Open NLP - `http://opennlp.sourceforge.net/`

2.4 Relationship Extraction

Relation extraction detects and defines the semantic relationships between entities, for instance part-whole relation, metonymy, synonymy, hyponymy *etc.* We aim to learn the relations between domain-specific entities in order to increase the accuracy of the entity extraction task. Resources that can be used for increasing biomedical entity extraction results by discovering semantic relationships between entities, can be divided into several categories:

- structured domain-independent *e.g., WordNet*[31];
- structured domain-dependent *e.g., Unified Medical Language System (UMLS)*[32], *Word Health Organization International Classification of Diseases (ICD)*[33], *Systematized Nomenclature of Medicine - Clinical Terms (SNOMED)*[34];
- semistructured domain-independent *e.g., Wikipedia*[35].

Although, *WordNet* is a manually constructed lexical database with structured knowledge and *Wikipedia*, by contrast, is an unstructured source of knowledge, they both are not domain-specific, therefore, they do not include enough information about infectious animal diseases, their synonyms and viruses. Also, the other domain-specific resources mentioned above *UMLS*, *ICD* and *SNOMED* cannot be applied for biomedical entity extraction in the domain of veterinary medicine, because they consist of concepts related to both human and animal diseases. Therefore, a unified ontology in a veterinary medicine domain is needed.

The process of the ontology construction is very difficult, labor-intensive and time consuming. In order to reduce the cost of building ontologies, there are several ontology learning systems which allow to extract concepts and relations between concepts from text *e.g., OntoLearn*[49], *OntoMiner*[24] and many others that are discussed in[31]. However, such systems are

[31]WordNet - `http://wordnetweb.princeton.edu/perl/webwn`
[32]UMLS - `http://www.nlm.nih.gov/research/umls/`
[33]WHO ICD - `http://www.who.int/classifications/icd/en/`
[34]SNOMED - `http://www.nlm.nih.gov/research/umls/Snomed/`
[35]Wikipedia - `http://www.wikipedia.org/`

generally based upon shallow natural language processing techniques and, therefore, mainly extract concepts with taxonomic (*e.g.,* synonymic *"is-a"*) relations between them. The taxonomic relation discovery approaches have been addressed primarily within the biomedical field as there are very large text collections readily available *e.g. PubMed*.

Other systems for automated ontology construction, such as *Text-To-Onto*[42] and its successor *Text2Onto*[20], allow extracting also non-taxonomic (*e.g.,* hyponymic) relations between concepts using association rule-mining and predefined regular expressions. Their main drawback is that they cannot effectively extract domain-specific concepts, because they identify semantic relations based on part-of-speech tags only. However, Cimiano and Staab[19] demonstrated the effectiveness of their system for extracting general concepts including person and location named entities. They use taxonomic and non-taxonomic patterns for semantic relation discovery between concepts, as a preliminary step for entity classification. It is sufficient to mention other related works about the extraction of semantic relations from web[13] and from bioscience text[62],[63]. In addition, Wang and Cohen[76] suggested a set expansion approach of domain non-specific named entities using the web and compared it with *GoogleSets* and *BayesianSets*[30].

By contrast with many ontology learning systems that use shallow parsing, *Concept Tuple-based Ontology Learning (CRCTOL)* performs full-text parsing using statistical and rule-based syntactic analysis of documents. It, thus, allows constructing richer ontologies in terms of the range and number of semantic relationships present in the ontology[35].

We suggest an approach for automated construction of a domain-specific ontology in Chapter 5, in contrast to other systems that construct general concept ontologies[20],[42],[31],[35], and use these ontologies to extract veterinary medicine entities. Similar to other systems[49],[24], we use a semantic relation extraction approach for automated ontology expansion, but by applying a comprehensive set of syntactic patterns and part of speech tagging, we capture non-taxonomic relations between concepts in addition to taxonomic relations.

2.5 Event Recognition

Event recognition is a detection of the events on sentence or document level where specific types of entities participate in. As described in Chapter 1 Section 1.3, we need to be able to discriminate several types of sentences that include:

- Type 1: *Emergent animal disease-related event* (emergent outbreak reports for animal disease surveillance);
- Type 2: *Non-emergent animal disease-related event* (past outbreak reports);
- Type 3: *Other disease-related events that are not related to outbreaks* (national and international level meetings/conferences about animal infectious disease surveillance);
- Type 4, 5: *Negation or hypothetical speculations of the animal disease-related events* (speculations in the scientific literature and other publications about animal infectious disease spread).

For instance, let us consider a sentence that we are particularly interested in: *"As of 21 Jun 2010, the total number of PCR-confirmed outbreaks (cases) in Miyazaki, Japan was 291"*. This sentence includes facts about confirmed animal disease-related event Type 1, which is an emergent outbreak report.

The event recognition task is very challenging because event may not be directly expressed in the sentence. For example, let us consider the sample paragraph: *"Foot-and-mouth disease killed 15 hog on farm in Taiwan. Outbreak was reported on 9 June"*.

It is clear that the sample paragraph requires coreference resolution, *e.g.,* identify all noun phrases that refer to the same object. There are several machine learning and other approaches that have been applied for the coreference resolution task including[56],[22] as well as existing Baltimore Anaphora Resolution Toolkit (BART)[70] and Illinois Coreference Resolution System [36].

[36] Illinois Coreference Resolution System - `http://l2r.cs.uiuc.edu/~cogcomp/coref_demo.php`

Moreover, let us discuss several systems presented in Section 2.1.2 for disease-related event detection that extract diseases and locations from text.

- *BioCaster* is an online ontology-based system for detecting and mapping infectious disease outbreaks from news[26]. Their approach for event detection is based on searching for disease-location pairs and calculating their frequency in the document and in the collection[37]. The methodology for deriving synonyms for disease-related verbs that are part of events <*disease, verb, location*> is similar to our approach. However, *BioCaster* does not provide assistance with classification of extracted events as confirmed or suspected.

- *Pattern-based Understanding and Learning System (PULS)* allows extracting metadata and structured facts related to animal disease outbreaks using pattern matching approach[67]. Similar to other systems, it does not classify extracted events and does not report anything about past outbreaks, which is important, for instance, for the predictive epidemiology domain.

- *HealthMap* is a manually supported web system, therefore it does not automatically extract events from the unstructured text. *HealthMap* crawls data from Google News and *ProMED-Mail* portal and provides reports about disease outbreaks to the public[28].As *HealthMap* processes only news articles and e-mails, it uses the date of publication as the reported date of the event.

Our approach addresses the limitations of the abovementioned systems. In Chapter 6 we propose a sentence-based approach for automated extraction of disease-related event tuples, which include disease, date, location, species entities and confirmation status. Moreover, we also classify automatically extracted event tuples into two categories such as: suspected or confirmed.

2.6 Summary and Discussion

In this Chapter we discussed the existing systems for monitoring animal disease outbreaks, related work for text categorization, entity and relation extraction, event recognition:

- We presented the broad overview of the existing systems for monitoring animal disease-related events. We discussed manually-supported web-services and their limitations *vs.* automated systems. Moreover, at a high level, we compared the automated systems with our framework for epidemiological analytics by several criteria including approaches to geo-location and domain-specific entity extraction, supported languages, productivity and mined sources.

- We briefly discussed related work to the most commonly used approaches for text categorization including supervised, unsupervised and semi-supervised learning. We introduced different representation of the document in terms of features such as: "bag-of-words" unigrams, bigrams, term frequency, TF-IDF, binary *etc.*

- We mentioned state-of-the-art methodologies for entity extraction such as: Hidden Markov Models and Conditional Random Fields. In addition to general named entity extraction (*e.g.,* geo-locations, dates, organizations), we concentrated on the related works for domain-specific entity recognition. We reviewed existing disease surveillance systems that also perform domain-specific entity extraction.

- We mentioned structured, unstructured and semi-structured sources for relation extraction between entities together with automated ontology learning systems.

- Finally, we reviewed the related work to the event recognition and existing approaches to coreference resolution. Furthermore, we discussed several methodologies for event recognition that existing surveillance systems apply.

Chapter 3

Framework for Epidemiological Analytics

3.1 System Overview

3.1.1 System Functionality

Taking into account the forensic, predictive and normative aspects of the system, we define its main purpose as capturing all possible breakdowns in communication channels between state, national and international levels of animal disease management. We target our intelligent tool for animal disease-related event detection at several groups of end-users:

- Research and Public Health Communities (*e.g.*, labs);
- Health Care Providers (*e.g.*, regional hospitals);
- Governmental Agencies (*e.g.*, CDC).

Users access system components using a web interface, search crawled documents, retrieve relevant information from the data storage, perform domain-specific entity extraction, recognize animal disease related events and visualize them on the map and within timeline, as shown in Figure 3.1.

The users of the system are provided with basic information retrieval and extraction functionality for detection, prevention and management of infectious animal disease event related information, including:

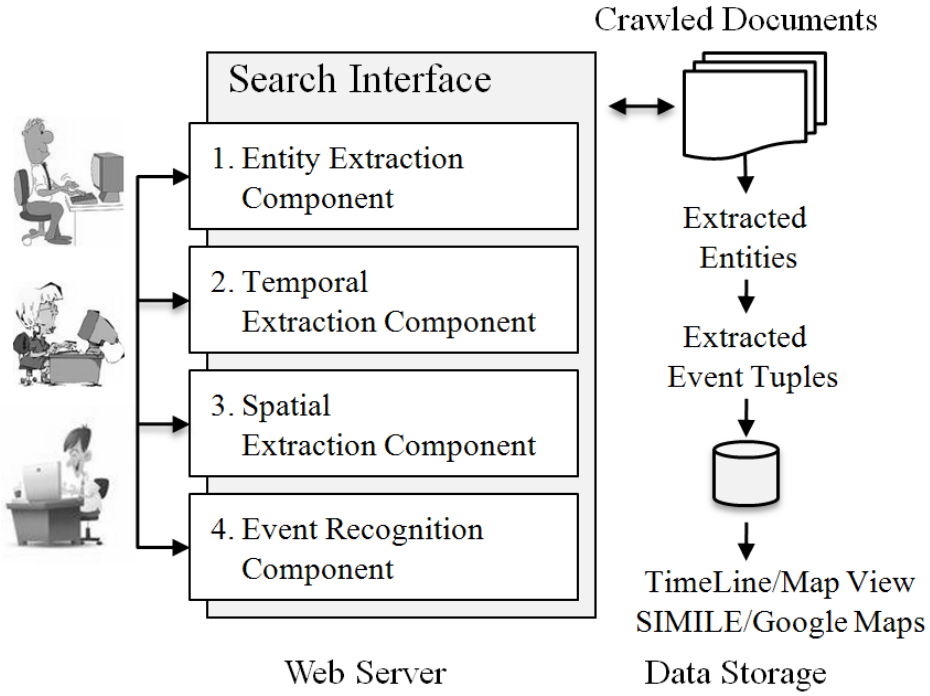


Figure 3.1: An overview of the epidemiological framework functionality

1. data **collection** using crawler components;

2. information **sharing** through the web interface;

3. query-based **search** using a Lucene-based[1] ranking component;

4. data **analysis** using entity extraction and event recognition components;

5. event **visualization** on a map (*GoogleMaps*) and within a timeline (*SIMILE*).

Algorithm 1 explains the information retrieval functionality listed above including data collection, sharing and search.

3.1.2 Data Collection using Web Crawling

For data collection we periodically crawl the web using Heritrix[2] crawler with a customized set of seeds (*e.g.,* ProMed-Mail, DEFRA *etc.*) and terms (infectious animal disease names

[1]Lucene Search Engine API - `http://lucene.apache.org/java/docs/`
[2]Heritrix Crawler - `http://crawler.archive.org/`

Algorithm 1 Information Retrieval Functionality (1 - 3)

```
Input: Set S of seeds s_p ∈ S and set T of terms t_i ∈ T, set of topics K.
Output: collection D of documents d_j, set of documents R^q relevant to
query q, and R^q ⊂ D.

doCrawl(S, T);
[D → K] = classifyDocsByTopics(D);
i = indexDocuments(D);
   if q ∈ {Disease} then
     [R^dis] = searchByDisease(dis, D);
      elseif q ∈ {Location} then
        [R^loc] = searchByLocation(loc, D)
        else
          [R^q] = searchByKeyword(q, D);
  end;
end.
```

from the ontology). Figure 3.2 shows that, by contrast with systems which use only news sources and do not digest refereed articles, we do not focus on specific sources.

After crawling, we perform an additional processing of web pages for entity extraction using domain-specific and domain-independent knowledge. Towards this goal, Weninger[77] developed a text-to-tag ratio-based method for content extraction from web pages.

Then, we perform document classification, as animal disease-related or non-related, using Naïve Bayes Classifier[50]. Finally, within the set of documents that are classified as disease-relevant, we allow users to perform search by disease and/or location entities in addition to general query-based keyword search.

3.1.3 Entity Extraction

After collecting the data, we are focused on an entity extraction task that is an automatic extraction of structured information about animal disease-related events from unstructured crawled web documents. More precisely, we seek to locate and classify atomic elements in text into predefined categories as shown in Figure 3.3:

– disease names (*e.g., "foot-and-mouth disease"*);

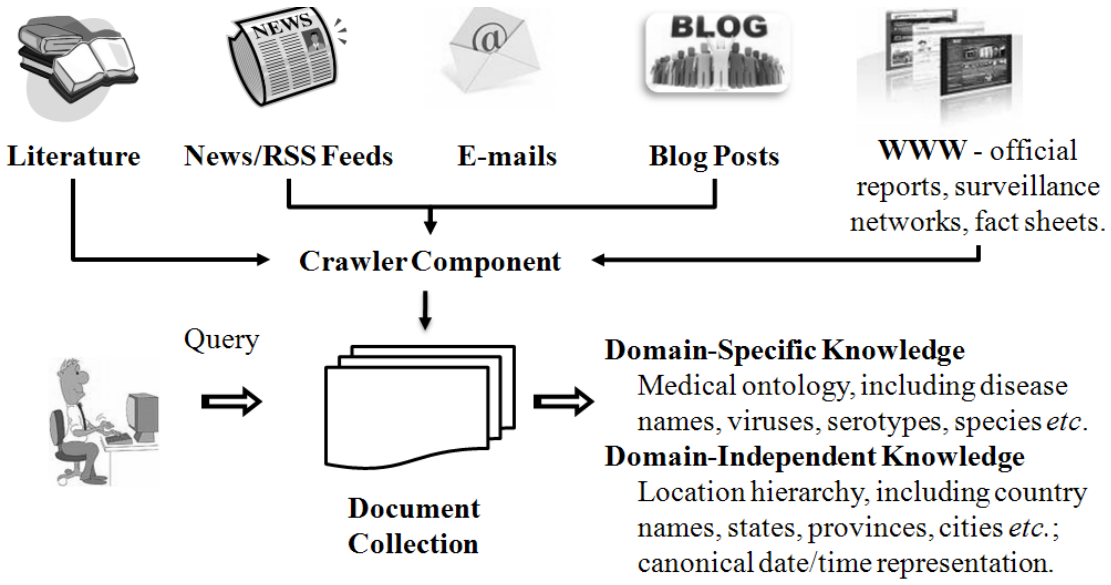


Figure 3.2: Crawled web documents from different domains

– viruses (*e.g., "FMDV"*), serotypes (*e.g. "SAT-1"*) - N/A;

– species (*e.g., "cattle"*) and quantities - N/A;

– locations (*e.g., "China"*);

– dates (*e.g., "Friday, Dec 13"*);

– organizations (*e.g., "Agriculture Ministry"*).

We developed several tagging tools including disease and species extractors[3] for automated domain-specific entity extraction. For animal disease extraction, we constructed an initial ontology O_{INIT} for the complete set of diseases and viruses using publicly available lists of animal disease names such as: CFSPH[4], DEFRA[5], OIE[6], Wikipedia[7].

For boosting animal disease extraction results, we enrich semantically and extend our initial ontology O_{INIT} by extracting semantic relations (including synonymic, hyponymic

[3]KDD DSEx - `http://fingolfin.user.cis.ksu.edu:8080/diseaseextractor/`
[4]CFSPH - `http://www.cfsph.iastate.edu/diseaseinfo/animaldiseaseindex.htm`
[5]DEFRA - `http://www.defra.gov.uk/foodfarm/farmanimal/diseases/atoz/`
[6]OIE - `http://www.oie.int/eng/maladies/en_alpha.htm`
[7]Wikipedia - `http://en.wikipedia.org/wiki/Animal_diseases`

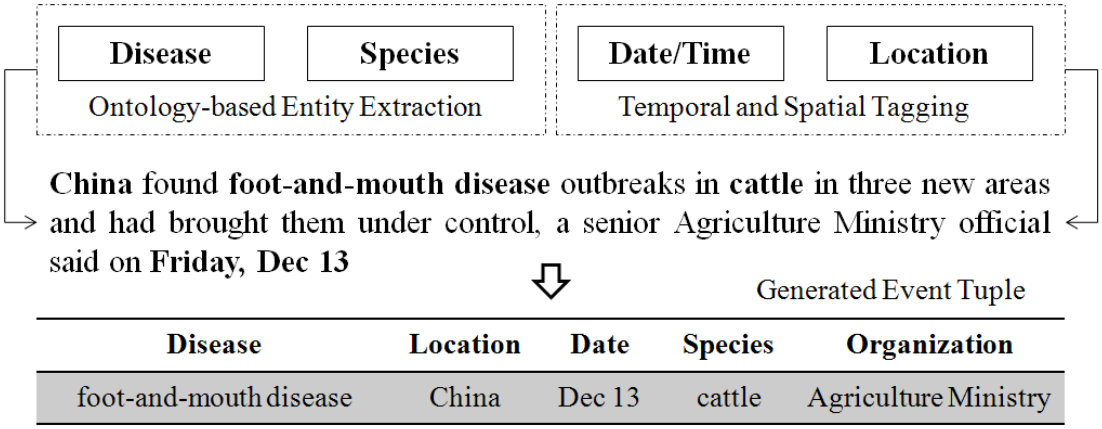


Figure 3.3: Information extraction component functionality for tagging disease names, species, dates, locations and organizations

and causative) between concepts. For semantic relation extraction approach we use syntactic pattern matching in combination with Part-of-Speech (POS) tagging[8]. For example, if we know a disease D and do not know the virus that causes it, we can learn the right-hand side patterns of relationship entailed in the text, such as E_i where "D *is caused by* E_i" [19,76].

For location named entity extraction, we used Stanford Named Entity Recognition (NER) tool[9]. It is based on conditional random fields approach developed by Lafferty[39]. Moreover, we refer to GEOnet Names Database (GNS)[10] for location disambiguation and getting latitude/longitude values. For date extraction, we perform pattern matching using regular expression-based rules. For species extraction we use pattern matching on a stemmed dictionary of animal names from Wikipedia.

3.1.4 Animal Disease-Related Event Recognition

The event recognition functionality is based on the entity extraction component which is described using an example in Figure 3.3. As can be seen, the extracted entities can be possibly augmented in event tuple in form $[disease, location, date, species]$, where the main event descriptors are disease, date, location and species. Additionally, we can extract organization

[8]NLTK POS Tagger - `http://www.nltk.org/`
[9]Stanford NER - `http://nlp.stanford.edu/ner/index.shtml`
[10]*GNS - http://earth-info.nga.mil/gns/html/*

that reports an outbreak.

We describe how the entity extractors discussed in Section 3.1.3 produce an event tuple for an example sentence in Figure 3.4. Initially, each document is tokenized into sentences; then disease, location, species and dates taggers are applied in addition to a confirmation status extractor which relies on the set of specific verbs for event recognition. For example, the sentence "*Foot and mouth disease is[V] a highly pathogenic animal disease"* is not disease related event, and by using constrained sets of a confirmation status verbs, we are able to eliminate this sentence.

Finally, the extracted events are visualized on the map using *GoogleMaps* and within a timeline using *SIMILE*. We summarize the entity extraction, event recognition and visualization functionality of the system in Algorithm 2.

Algorithm 2 Information Extraction, Event Recognition and Visualization Functionality (4 - 5)

```
Input:  Set of documents R^q ⊂ D relevant to q
Output:  Set of events E with attributes e_i = [dis, loc, dat, sp] on timeline/map.

foreach document d_j ∈ R^q do
   [dis, loc, dat, sp] = extractEntity(d_j);
   e_i = generateEventTuple([dis, loc, dat, sp]);
[E*] = eventAugmentation(E);
doVisualization(E*);
end.
```

3.2 Summary and Discussion

In this Chapter we presented an overview of the framework for epidemiological analytics. We described the main functionality of the system including data collection, information retrieval, information extraction, event recognition and aggregation, visualization.

During the development of the framework for epidemiological analytics we encountered several opened research questions including:

– managing the contextual specificity of blogosphere[55];

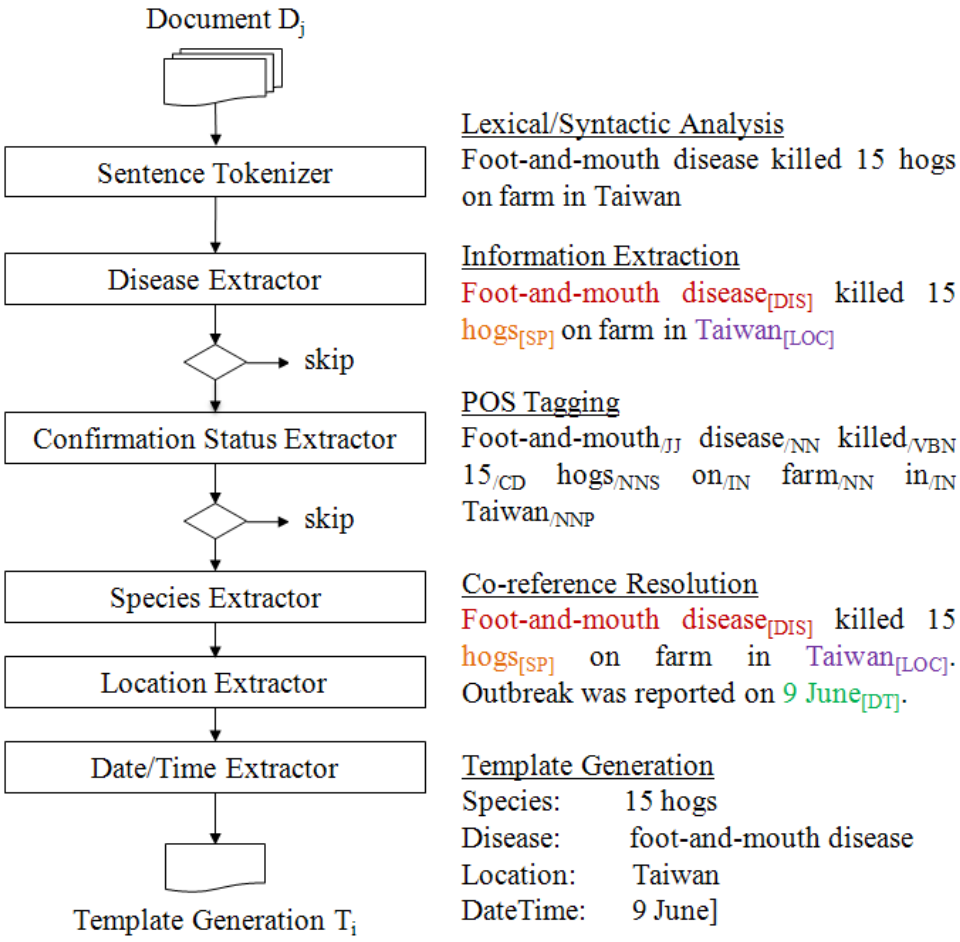


Figure 3.4: Event recognition component functionality for event tuple generation by extraction entities and confirmation status verbs

– processing biomedical literature[47],[34],[41];

– mining news content vs. official health reports[58].

Similarly to systems described in Section 2.1.2, we applied existing approaches for data collection using web crawling[17] and text classification[50]. However, for entity extraction, we proposed an ontology-based extraction method and semantic relation learning approach for ontology expansion[73,75]. For event recognition, we performed event tuple generation using extracted entities such as: disease, location, date, species together with the confirmation status verb (in contrast to the "disease-location" pairs used in other systems).

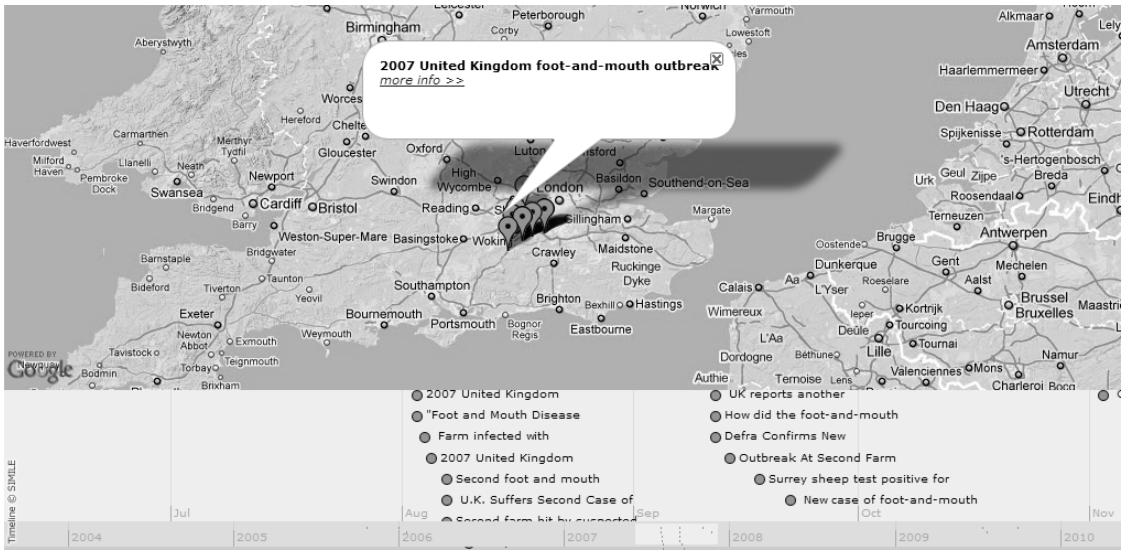


Figure 3.5: Temporal and spatial visualization of the extracted animal disease outbreak related events

Consequently, in comparison to other systems which are designed for mining news and have no functionality for past outbreak tracking (see Table 2.1 - Mined sources), perform ontology-based information extraction for limited number of domain-specific entities (see Table 2.1 - Domain-specific Entities), require manual moderation phase (*HealthMap*), limited with geo-entity extraction (see Table 2.1 - Geographical Entities) and have no timeline visualization (*BioCaster*), our system:

- performs focused crawling of different sources (books, research papers, blogs, governmental sources, *etc.*);
- uses semantic relationship learning approach (including synonymic, hyponymic, causal relationships) for automated-ontology expansion for domain-specific entity extraction (*e.g.,* diseases, synonyms, corresponding viruses)[75];
- recognizes geo-entities using CRF approach[33] and disambiguates them using GNServer;
- extracts animal disease-related events with more descriptive event attributes such as: species, dates, event confirmation status[72], in contrast to "disease-location" pairs;

– supports timeline representation of extracted events in *SIMILE* in addition to visualized events on *GoogleMaps*.

The main limitation of our system is the ability to process web documents only in English whereas the other systems process document in multiple languages as discussed in Table 2.1.

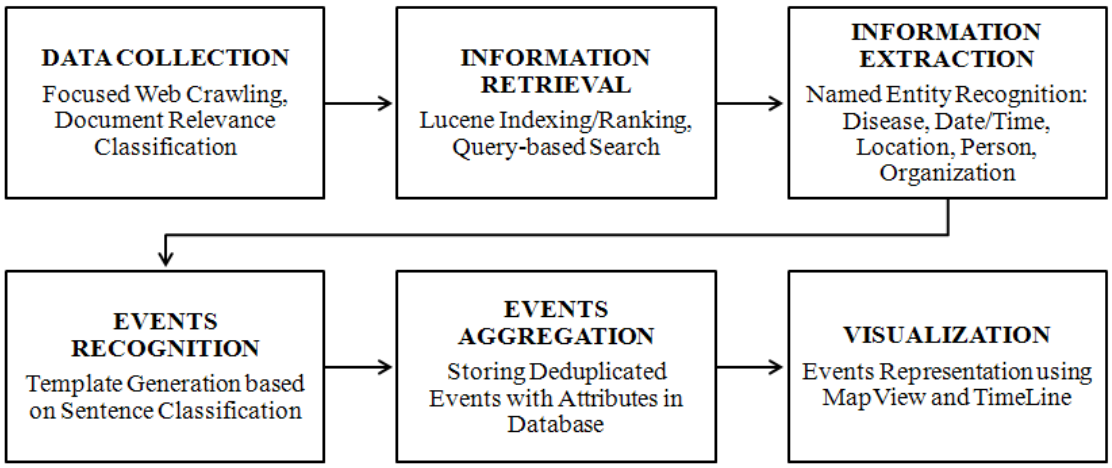


Figure 3.6: Main functional components of the epidemiological analytics framework

Finally, in Figure 3.6 we present main functional components of our framework for epidemiological analytics.

Chapter 4

Disease-Related Document Classification

4.1 Supervised Framework for Text Categorization

After the document collection by focused crawling, using the predefined set of seeds (*e.g., WHO, DEFRA*) and the set of terms (*e.g.,* animal disease names), we noticed that a lot of documents are disease non-relevant documents. Therefore, it was necessary to perform additional *disease-related document classification* task in order to leave only disease-related documents for the next *"entity extraction"* and *"event recognition"* phases and eliminate disease non-related documents. In Figure 4.1 we present the supervised learning framework for disease relevance document classification.

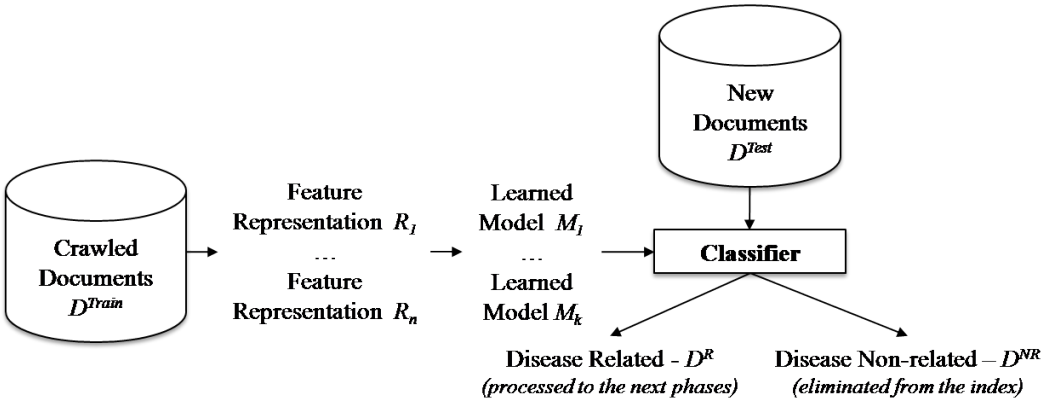


Figure 4.1: The supervised learning framework for disease relevance document classification

As shown in Figure 4.1 our goal is to categorize documents from the original collection D into two sub-collections such as: disease-related D^R and disease non-related D^{NR} documents using different feature representations $R_1 \dots R_n$ and classification algorithms for learning different models $M_1 \dots M_k$.

There are several approaches for obtaining feature representation of the collection D as described in Chapter 2, Section 2.2. The "bag-of-words" is the most commonly used approach for text categorization, where each word from the vocabulary V for the collection D is a feature. Moreover, prior to feature extraction, it is essential to remove all stop words and convert all words to a small case. As a result, each document from the collection D is represented as a point in m-dimensional feature space, where m is the sized of the vocabulary for a specific collection.

We considered two sets of features for the collection representation in the feature space such as: *comprehensive feature representation* and *simplified feature representation*. The comprehensive set of features includes several subtypes such as:

- R_1 - "bag-of-words" representation using binary counts as defined in Equation 2.1;
- R_2 - "bag-of-words" representation using one-gram words and their frequency as defined in Equation 2.2;
- R_3 - "bag-of-words" representation using word bigrams and bigram frequency (similar to the previous, but we counted the bigram frequency instead on unigram frequency).

We extracted a set of comprehensive features using *Mallet Toolkit*[1]. As a result, we ended up with three different feature representations for our collection D. The size of the feature vectors R_1, R_2 is equal to the size of the unigram vocabulary $|V^{uni}| = 28908$ tokens. The size of the feature vector R_3 is equal to the size of the bigram vocabulary $|V^{bi}| = 99108$ vector components.

[1]Mallet Toolkit - `http://mallet.cs.umass.edu/`

For simplified feature representation we extracted domain-specific noun and verb keywords from each document in the collection D. We present the domain-specific noun and verb keywords in Table 4.1. The simplified set of features includes several subtypes such as:

- R_4 - noun and verb keywords represented as binary counts (presence or absence 1/0 of noun and verb keyword in the document) as shown in Equation 2.1;
- R_5 - noun and verb keywords represented as a normalized term frequency (number of times the keyword appears in the document divided into the total number of tokens in the document) as shown in Equation 2.2.

As a result of the simplified feature extraction for our collection D, we obtained two feature vectors R_4 and R_5. The size of feature vectors R_4, R_5 is equal to two features for each document $|R_4| = 2$, $|R_5| = 2$, which is significantly smaller compared to the size of the feature vectors $|R_1| = 28908$, $|R_2| = 28908$, $|R_3| = 99108$. Moreover, it is crucial for the performance of the supervised learning framework.

In Figure 4.2 we show the collection representation in terms of simplified features R_5. Each document $d_i \in D$ is a point in the two-dimensional space, where the first dimension is a *NounNorm* feature and the second - *VerbNorm* feature. As can be seen, this feature space is not easy to separate using a linear classifier. Therefore, we decided to use many different classifiers in order to learn the best model from this data and then, select the inducer that demonstrates the best performance.

Table 4.1: Noun and verb keywords for simplified feature extraction for document relevance classification

Noun Keywords	Verb Keywords
Virus	Infected
Disease	Confirmed
Outbreak	Reported
Fever	Died
Illness	
Symptoms	

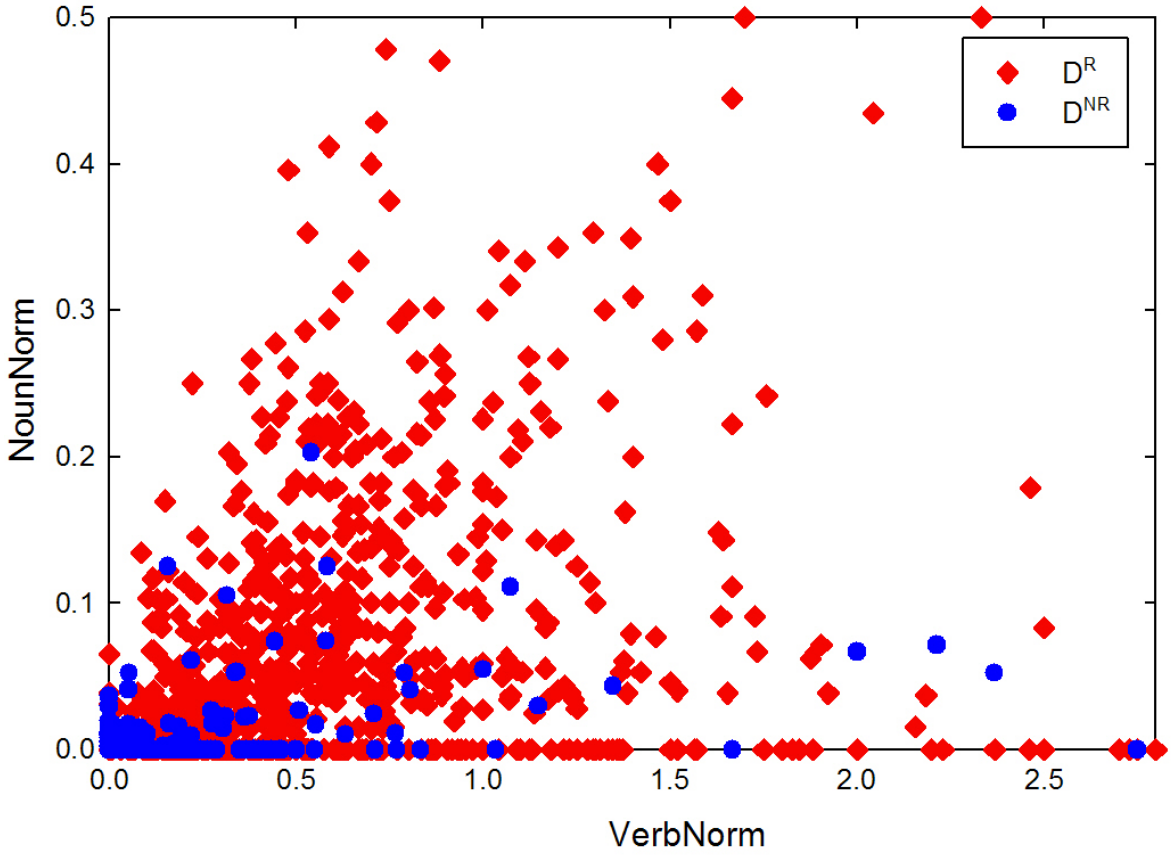


Figure 4.2: The feature space in terms of the noun and verb keyword normalized frequency

The model learning is the next stage in our supervised learning framework after the feature representation. For the purpose of learning different models $M_1 \dots M_3$, we used *Mallet* and learned two classifiers from R_1, R_2 and R_3 representations such as:

1. Naive Bayes learner that classifies a new instance x based on a tuple of attribute values $x = [x_1, x_2, \dots, x_n]$ into one of the classes $c_j \in C$:

$$\begin{aligned} c_{MAP} &= \arg\max_{\mathbf{c}_j \in C} P(c_j | x_1, x_2, \dots, x_n) \\ &= \arg\max_{\mathbf{c}_j \in C} \frac{P(x_1, x_2, \dots, x_n | c_j) P(c_j)}{p(x_1, x_2, \dots, x_n)} \\ &= \arg\max_{\mathbf{c}_j \in C} P(x_1, x_2, \dots, x_n | c_j) P(c_j), \end{aligned} \tag{4.1}$$

where $P(c_j)$ can be estimated from the frequency of classes in the training examples and $P(x_1, x_2, \ldots, x_n|c_j)$ requires the independence assumption which assumes that the probability of observing the conjunction of attributes $x_1, x_2, \ldots, x_n$ is equal to the product of the individual probabilities $P(x_i|c_j)$.

The Bayesian approach has several problems: first, because each document d_i is represented in the high-dimensional feature space, it is difficult to estimate $P(d_i|c_j)$ when the training collection size is small; second, it is dangerous to add such features as phrases, part-of-speech tags to the Naive Bayes Multinomial feature representation, because these features may be highly correlated with the original features.

2. Maximum Entropy classifier is an alternative probabilistic framework, that considers each class to be equally likely and separates the decision boundaries by maximizing the entropy of the model distribution $P(c_j|d_i)$.

For the purpose of learning models M_4, M_5, we used *Weka Software*[2] and selected several classifiers for learning from R_4 and R_5 representations such as:

1. Lazy: IB1, IBk (k-nearest neighbor learner, k=2), KStar use a simple distance measure to find the training instance closest to the given test instance, and predicts the same class as this training instance;

2. Meta: AdaBoost iteratively calls the set of weak classifiers and increases the weights of incorrectly classified examples from previous iteration, so the new classifier focuses more on those examples;

3. Trees: J48, RandomForest build the decision tree(s) from a set of labeled training data using the concept of information entropy;

4. Rules: ZeroR predicts the mean for a numeric class or the mode for a nominal class;

5. Bayes: Naive Bayes, Naive Bayes Multinomial as shown in Equation 4.1;

[2]Weka Data Mining Software - `http://www.cs.waikato.ac.nz/ml/weka/`

6. Functions: Logistic predicts the probability of occurrence of an event by fitting data to a logit function logistic curve; MultiLayer Perceptron and RBFNetwork use k-means clustering algorithm to provide the basis functions and learn either a logistic regression for a discrete class problems or linear regression for a numeric class problems[3].

After we learn different models $M_1 \dots M_k$ using different classifiers and different feature representations $R_1 \dots R_n$, we need to select only one classifier C_i that demonstrates the best performance as well as corresponding feature representation R_j in order to classify new web-documents D^{Test}. As a result of the disease-related document classification phase, we process all disease-related documents into the next phases such as *"entity extraction"* and *"event recognition"* as well as remove all disease non-related documents from the index.

4.1.1 Experimental Design and Results: Experiment A

For the disease-related document classification in the supervised framework, we performed a separate focused crawl where the list of terms included two specific animal infectious diseases such as *"foot and mouth disease"* and *"rift valley fever"*. More precisely, we narrowed the original list of terms that included all animal infectious diseases from the ontology to several specific terms of interest such as:

$$Terms = [foot\ and\ mouth\ disease,\ FMD,\ rift\ valley\ fever,\ RVF].$$

We collected 1500 documents both related and non-related for *FMD* and *RVF* diseases. Then, we manually labeled each document as *disease related* D^R or *disease non-related* D^{NR}. After labeling, we had 813 related and 752 non-related documents in the sample collection D. Next, we applied all abovementioned inducers for learning different models $M_1, M_2, \dots, M_k$ while experimenting with different feature representations $R_1, R_2, \dots, R_n$ for the documents. We show *Weka* attributes for the feature representations R_4 and R_5 in Table 4.2.

[3]WekaDocs - `http://weka.sourceforge.net/doc/weka/classifiers/functions/`

Table 4.2: Simplified feature representations for the document collection D: binary *vs.* normalized keyword frequency features

Binary Representation R_4	Normalized Keyword Frequency Representation R_5
@attribute noun NUMERIC	@attribute nounNorm NUMERIC
@attribute verb NUMERIC	@attribute verbNorm NUMERIC
@attribute class {0,1}	@attribute class {0,1}

We run the first set of experiments using *Weka* in order to learn different models $M_1, M_2, \ldots, M_i$ from simplified feature representations R_4 and R_5 for each of the classifiers using 10-fold cross validation. We present the results for animal disease-related document classification in terms of precision, recall, F-measure and Area Under Curve (AUC).

In Table 4.3 we show the results obtained using simplified binary counts as a feature representation R_4 for all abovementioned classifiers (lazy, meta, trees, rules, Bayes, functions). Similarly, in Table 4.4, we present the results obtained using simplified normalized keyword frequency as a feature representation R_5. Based on the performance results from Table 4.3 and Table 4.4, the normalized domain-specific keyword frequency R_5 is better feature representation than just binary counts R_4.

Table 4.3: The results for classifiers trained on the simplified binary features, tested with 10-fold cross validation

Learning algorithm	AUC	F-Measure	Precision	Recall
IB1	0.78	0.77	0.81	0.77
IBk	0.90	0.85	0.94	0.77
KStar	0.90	0.85	0.94	0.77
AdaBoost	0.90	0.85	0.94	0.77
J48	0.86	0.85	0.94	0.77
RandomForest	0.90	0.85	0.94	0.77
NaiveBayes	0.90	0.85	0.94	0.77
NBMulti	0.66	0.21	0.32	0.16
SimpleLogistic	0.86	0.85	0.94	0.77
Logistic	0.90	0.85	0.94	0.77
MultiLayerPerceptron	0.90	0.85	0.94	0.77
RBFNetwork	0.90	0.85	0.94	0.77

Table 4.4: The results for classifiers trained on simplified keyword frequency features, tested with 10-fold cross validation

Learning algorithm	AUC	F-Measure	Precision	Recall
IB1	0.87	0.85	0.86	0.85
IBk	0.93	0.90	0.91	0.89
KStar	0.95	0.90	0.90	0.90
AdaBoost	0.94	0.90	0.92	0.88
J48	0.92	0.90	0.91	0.90
RandomForest	0.94	0.89	0.89	0.90
NaiveBayes	0.94	0.83	0.74	0.95
NBMulti	0.51	0.21	0.26	0.18
SimpleLogistic	0.94	0.87	0.82	0.93
Logistic	0.94	0.87	0.82	0.93
MultiLayerPerceptron	0.94	0.89	0.89	0.89
RBFNetwork	0.93	0.87	0.84	0.90

All inducers demonstrate comparatively equal performance in terms of F-measure except *ZeroR* and *Multinomial Naive Bayes*. This can be explained by the fact that rule-base classifier such as *ZeroR* predicts mean for a numeric class and since we have a small collection it is very difficult to make an accurate prediction. Bayesian classifier such as *Multinomial Naive Bayes* is not able to make an accurate prediction using limited number of simplified features exacted from limited amount of training data.

Next, we run the second set of experiments using *Mallet* in order to learn different generative models $M_{i+1}, \ldots, M_k$ from comprehensive "bag-of-words" feature representations R_1, R_2, R_3. We learn *Naive Bayes* and *MaxEnt* classifiers and use 10-fold cross validation for testing. We report results in terms of accuracy and compare them with the results obtained using simplified features in Figure 4.3 for binary feature representation and in Figure 4.4 for term frequency as a feature representation. As can be seen, the more features we use the better accuracy we can get. The best accuracy is 0.97 obtained using *Naive Bayes* classifier and comprehensive bigram feature representation.

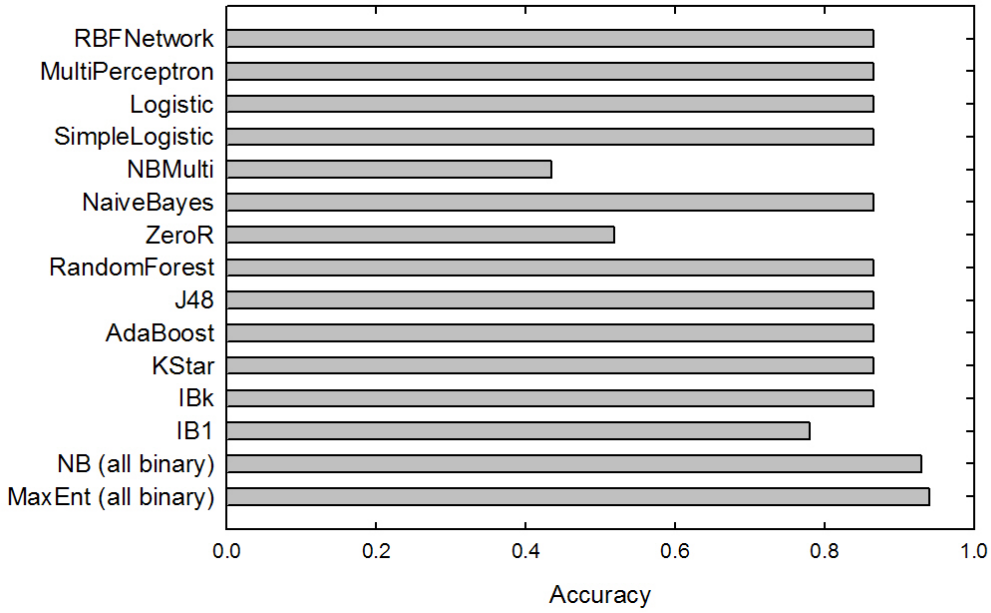


Figure 4.3: The accuracy for classifiers trained on binary features, tested with 10-fold cross validation

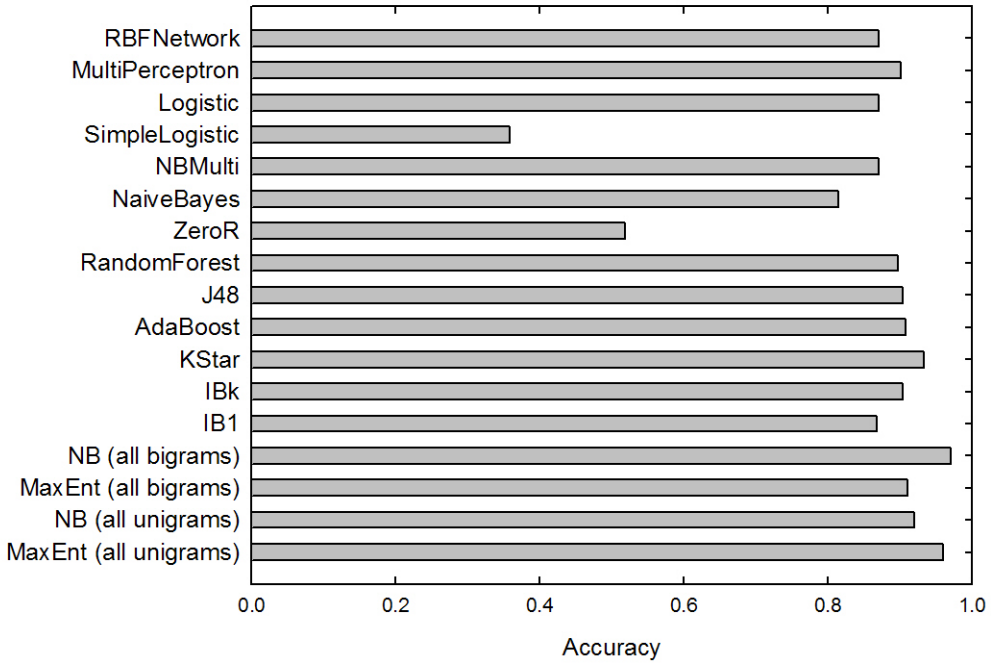


Figure 4.4: The accuracy for classifiers trained on the unigrams/bigrams/keyword frequency features, tested with 10-fold cross validation

4.2 Summary and Discussion

In this Chapter we described disease relevance document classification task in supervised framework. We considered both discriminative and generative approaches for learning different models $M_1, M_2, \ldots, M_k$ as well as different feature representations $R_1, R_2, \ldots, R_n$ for this binary classification problem.

Taking into account the limited size of the experimental collection D and the limitations of each learning approach, we summarize that:

- "bag-of-words" representation for each document in the collection gives higher accuracy compared to simplified feature representation; however, using the "bag-of-words" representation for big collection may be crucial because of the learning time;

- generative approaches for classification - *e.g., Naive Bayes* together with comprehensive feature representation R_3 - *e.g., bigrams*, give the highest accuracy - 0.97; we also report accuracy as high as 0.96 and 0.94 for the *MaxEnt* classifier using unigram "bag-of words" representation R_2 and comprehensive binary counts as feature representation R_1 respectively;

- normalized keyword frequency as a simplified feature representation R_5 gives better results compared to simplified binary keyword counts R_4;

- *rule-base classifiers (J48, RandomForest), functions (Logistic, MultiLayer Perceptron, RBF Network), meta (AdaBoost)* and *lazy (IBk, k=2* and *KStar)* perform almost equally in terms of F-measure values range $[0.85 - 0.90]$ and AUC $[0.90 - 0.94]$;

- *Naive Bayes Multinomial* is too sensitive to the number of features, therefore it performs poorly when we are using simplified feature representations;

- *KStar* classifier demonstrates the highest value of AUC when we are using simplified keyword frequency as features;

– *IB1* shows the worst performance in comparison to other lazy inducers (*IBk, k=2* and *KStar*);

– several classifiers such as: *Multinomial Naive Bayes* and *ZeroR* show the worst performance because these inducers are not able to learn accurate models from the existing feature representation.

Finally, let us compare document categorization performance of the existing disease surveillance systems *BioCaster* and *MedISys* with our results:

– *BioCaster* gold standard corpus includes 1000 articles as a training data. The reported accuracy is 84.4% obtained using *Naive Bayes* classifier and 10-fold cross validation with a "bag-of-words" feature representation in combination with named entity frequency and their roles such as: case, therapeutic, and transmission.

The list of the named entities includes: *anatomy, symptom, disease, virus, person, organization, location etc.* The reported accuracy using *disease* named entities as features is as high as 76.2% compared to the accuracy of 81.8% using *person* named entities as features for classification[25].

In addition, they report an increase in accuracy by using n-grams, semantic tag-based features and Chi-squared feature selection:

up to 94.8% using *Naive Bayes* - F-score 0.93, Precision 0.89, Recall 0.97;

up to 92.1% using *SVM* - F-score 0.89, Precision 0.88, Recall 0.90.

– *MedIsys* system applies the document categorization functionality before the information extraction by clustering the collected articles from more than 1500 news portals and 150 health care resources into 750 categories in real time every ten minutes. However, they are using the predefined set of 30 000 patterns (multi-word terms and their combinations in 43 languages) for the article clustering. For the comparison of the upcoming articles and further clustering, the Vector Space Representation (VSR) and the cosine similarity are applied.[67]

Table 4.5: The comparison of document classification performance of existing surveillance systems *vs.* our results

System	Features	Classifier	Accuracy	F-Measure	Precision	Recall
BioCaster (1000 docs)	Raw text, all NEs, roles	NB	0.84	-	0.75	**1**
	N-grams, semantic tag-based features, Chi-squared feature selection	NB	0.95	0.93	0.89	**0.97**
		SVM	0.92	0.89	0.88	0.90
OurResults (1578 docs)	Keywords normalized frequency	NB	0.81	0.83	0.79	0.95
	All Bigrams	NB	**0.97**	0.92	0.88	0.96
		MaxtEnt	0.91	**0.95**	0.93	**0.97**
	All Unigrams	NB	0.92	0.92	0.87	**0.97**
		MaxtEnt	**0.97**	**0.96**	**0.95**	0.96
	All Binary	NB	0.93	0.92	0.89	0.96
		MaxtEnt	0.94	**0.95**	0.93	**0.97**

As can be seen from Table 4.5, the size of the training collection and different feature representations allow us to achieve document classification result comparable to *BioCaster*. Unfortunately, we are unable to compare our results with *MedISys* system because first, authors do not report the document classification accuracy and second, it is different clustering task (multiple-class categorization *vs.* binary classification, unsupervised *vs.* supervised approach), therefore it would be unfair comparison in any case.

In addition, there is *FMD BioPortal* that also reports disease-related/non-related document classification results. They use "bag-of-words" feature representation, noun phrases and named entities as features together with different machine learning algorithms such as: K-Nearest Neighbor, Naive Bayes, SVM[84]. The highest precision is 77.04% in compared to our 95% as shown in Table 4.5.

Chapter 5

Domain-Specific Entity Extraction

5.1 Ontology-based Entity Extraction

5.1.1 Manual Ontology Construction

We manually construct an initial ontology O_{INIT} using lists of diseases retrieved from publicly available domain-specific dictionaries such as: CFSPH[1], DEFRA[2], OIE[3], Wikipedia[4]. After manual merging and deduplication of the abovementioned disease lists, we have 429 concepts in the initial ontology O_{INIT}. Next, we manually discover and update this ontology with sets of synonyms and abbreviations. The size of the manually-updated ontology with synonyms is $|O_S| = 581$ concepts, with abbreviations is $|O_A| = 453$ concepts and with both is $|O_{S+A}| = 605$ concepts. The initial manually-constructed ontology O_{INIT} is expanded with semantic relationships extracted as described in the next section.

5.1.2 Automated Relationship Extraction

Our relationship extraction approach is based on discovering semantic relationships between concepts in the collection by using rule-based syntactic pattern matching and part-of-speech (POS) tagging. We look for taxonomic and non-taxonomic linguistic relationships between entities using the initial ontology and raw data from the veterinary medicine domain. There

[1]CFSPH - `http://www.cfsph.iastate.edu/diseaseinfo/animaldiseaseindex.htm`
[2]DEFRA - `http://www.defra.gov.uk/foodfarm/farmanimal/diseases/atoz/`
[3]OIE - `http://www.oie.int/eng/maladies/en_alpha.htm`
[4]Wikipedia - `http://en.wikipedia.org/wiki/Animal_diseases`

are several relationships that we are interested in, such as:

1. Synonymic relationships of the form "E_1 *is a kind of* E_2", *e.g.,* E_1 = *"swine influenza" is a kind of* E_2 = *"swine fever"*, where E_1 and E_2 are synonyms - different words with identical or very similar meanings.

2. Hyponymic relationships of the form "E_1 *and* E_1 *are diseases*", *e.g.,* E_1 = *"anthrax"*, E_2 = *"yellow fever"* are diseases, where E_1 and E_2 are hyponyms (words that are conceptually included within the definition of another word - their hypernym *disease*, but not synonyms).

3. Causal relationships that capture causative dependencies between diseases and viruses such as "E_1 *is caused by* E_2", *e.g.,* E_1 = *"Ovine epididymitis" is caused by* E_2 = *"Brucella ovis"*.

We present syntactic patterns in Table 5.1 for synonymic, hyponymic and causal relationship discovery from text in the domain of veterinary medicine. We use the following notation:

- C_{GEN} corresponds to general "*disease*" concept,
- C_{INIT} represents the concept from the initial ontology,
- C_L represents the learned concept added to new ontology (add C_L learned concept to new ontology O_R if it is not present in the initial ontology O_{INIT}),
- "/" represents a flexible substring within a pattern,
- C_i, C_j correspond to the concepts,
- hyponymic$_{G \to S}$ represents the relationship with learning from general concept to specific,

Table 5.1: The subset of syntactic patterns for semantic relationship extraction between domain-specific concepts

Relationship Type	C_{INIT}	Relationship Pattern	C_L
Synonymic	C_i	"is a" "is a kind of" "and/, | , " "/, /also known as " "/, /is also called "	C_j
Hyponymic$_{G\leftarrow S}$	C_{GEN}	"such as/: | :" "*e.g.,* | for example" "/, for instance /," "including | ()" "/, especially /,"	C_i and/or/, C_j
Hyponymic$_{G\rightarrow S}$	C_{GEN}	"and|or other" "/, and|or C_j are"	C_i
Causal	C_i	"is caused by" "causes"	C_j

– hyponymic$_{G\leftarrow S}$ denotes the relationship which is read from right to left using the set of rules.

Let us consider several examples of the patterns that we consider in our approach for:

– synonymic relationship - *"foot and mouth disease is also called FMD"*,

– hyponymic$_{G\rightarrow S}$ relationship - *"diseases, for instance baylisascariasis and typeworm"*,

– hyponymic$_{G\leftarrow S}$ relationship - *"west nile virus is an animal infectious disease"*,

– causal relationship - *"lyme disease is caused by borrelia burgdorferi sencu lato, borrelia garinii"*.

As can be seen through these examples, the relationship extraction phase can be used to improve the descriptiveness of the ontology by including domain-specific semantic relationships between concepts.

5.1.3 Automated Ontology Construction

We construct new ontology O_R using the initial ontology O_{INIT} and semantic relationships extracted by applying syntactic patterns described in Table 5.1. In addition, we use POS tagging[5] to extract n-gram concepts *e.g., "swine vesicular disease"*. The resulting ontology O_R will contain automatically extracted disease synonyms, abbreviations and viruses.

More precisely, we start with the canonical disease name *"foot-and-mouth disease"* taken from the initial ontology and after processing the sentence "*Foot-and-mouth disease*, *FMD* or *hoof-and-mouth disease (Aphtae epizooticae)* is a highly contagious and sometimes fatal viral disease", we update the ontology O_R with *"foot-and-mouth disease"* $\xrightarrow{is\ a\ kind\ of}$ *"hoof-and-mouth disease"* $\xrightarrow{is\ a\ kind\ of}$ *"aphtae epizooticae"* $\xrightarrow{abbrev.}$ *"FMD"* $\xrightarrow{is\ a}$ *disease*, where $\xrightarrow{is\ a\ kind\ of}$, $\xrightarrow{abbrev.}$ denote synonymic relationships between concepts, $\xrightarrow{is\ a}$ denotes hyponymic$_{G \to S}$ relationships.

After processing the next sentence "*FMD* is caused by *foot-and-mouth disease virus (FMDV)*", we extract a causal relationship between concepts and update the ontology O_R with *"foot-and-mouth disease"* $\xrightarrow{is\ caused\ by}$ *"foot-and-mouth disease virus"* by associating *"FMD"* with its canonical disease name from the initial ontology O_{INIT} and relating *"foot-and-mouth disease virus"* with its synonym *"foot-and-mouth disease virus"*$\xrightarrow{is\ a\ kind\ of}$ *FMDV*.

From the sentence "Pandemic Strain of *Foot-and-Mouth Disease Virus Serotype O*" we extracted serotype of the disease and updated the ontology O_R with *"foot-and-mouth disease virus"*$\xrightarrow{has\ serotype}$*serotype O*.

5.1.4 Entity Extraction

We define the biomedical entity extraction task as the automated extraction of structured information related to animal diseases from unstructured web documents. This task requires the development of an extractor for tagging entities such as: animal disease names (*e.g.,*

[5]NLTK POS Tagger - `http://www.nltk.org/`

"Brucellosis"), their synonyms (*e.g., "Malta fever", "Undulant fever", "Bang's disease", "Gibraltar fever"*), viruses or other causative agents (*e.g., "Brucella abortus", "Brucella canis"*) and serotypes (*e.g., "A+M-", "A-M+", "A+M+"*).

We used an ontology-based pattern matching approach to design a biomedical entity extractor DSEx[6] that takes raw web documents as input and returns a set of attributes for the matching concepts as output.

In Figure 5.1, we show the attributes that the entity extractor outputs. Let us consider the sentence: "Species infecting domestic livestock are *B. melitensis*$_{DS}$ (goats and sheep, see *Brucella melitensis*$_{DS}$), *B. suis*$_{DS}$ (pigs, see *Swine brucellosis*$_{DS}$), *B. abortus*$_{DS}$ (cattle and bison), *B. ovis*$_{DS}$ (sheep), and *B. canis*$_{DS}$ (dogs)", where tag $_{DS}$ corresponds to animal disease names. The attributes extracted for the first entity in this sentence are: *[41 - 54, B. melitensis, 13, Brucellosis,* {*Malta fever, Undulant fever, Brucella*}*, 1].*

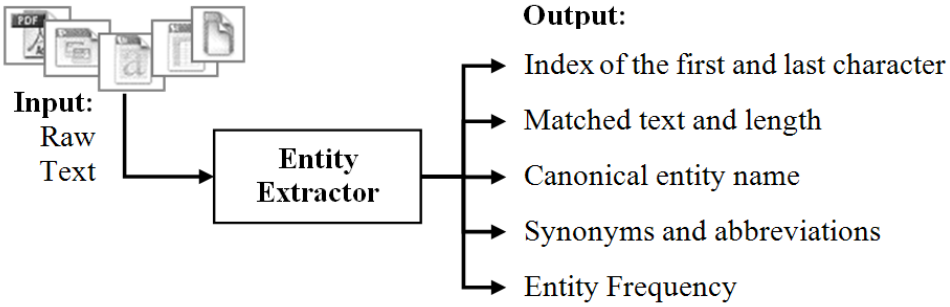


Figure 5.1: The output from the entity extractor

As can be seen from the example above, there are several subtasks of the entity extraction task[44]. The first is *terminology extraction*, which identifies specific relevant concepts named in documents based on the ontology (*e.g.,* diseases, viruses, serotypes). For example, we extract one disease term from the sentence: "Epidemics of *foot-and-mouth disease*$_{DS}$ have resulted in the slaughter of millions of animals".

The second subtask is the *segmentation task*, which means finding the starting and ending character positions of the named entities, for example: "*African swine fever virus*$_{VR,\ 1-25}$

[6]KDD DSEx - `http://fingolfin.user.cis.ksu.edu:8080/diseaseextractor/`

($ASFV_{VR,\ 28-31}$*)* is the causative agent of *African swine fever*$_{DS,\ 60-78}$".

The next subtask is the *association extraction task*, which we consider as a separate prerequisite task for the automated ontology construction in Section 5.1.2. It looks for phrases indicating relationships between entities and matches them against the set of patterns from Table 5.1 for inferring associations between diseases, their synonyms and abbreviations (*e.g., "avian influenza" is a kind of "bird flu" is a "H5N1"*) or disease and the causative virus (*e.g., "Brucellosis"* is caused by *"Bacillus abortus"*).

The *normalization subtask* matches all disease names to their canonical versions based on the constructed ontology. For example in the sentence: "*Tick fever*$_{DS}$ is a significant disease of cattle in Australia with up to 7 million animals potentially at risk", the extractor relates *"Tick fever"* with its canonical disease name *"Babesiosis"*.

Algorithm 3 Biomedical ontology-based entity extraction and semantic relationship discovery using syntactic patterns

Input: Two document collections D_1 and D_2, initial ontology O_{INIT} and other manually-constructed ontologies O_S, O_A, O_{S+A}, sets of patterns from Table 5.1
Output: Automatically-constructed ontologies O_R, O_G, sets of entities obtained using $\{E_{INIT}\}$, $\{E_S\}$, $\{E_A\}$, $\{E_{S+A}\}$, $\{E_R\}$ and $\{E_G\}$

```
for all d_j ∈ D_1 do
    R_i ⇐ ExtractRelation(O_INIT, D_1);
    O_R ⇐ ConstructOntology(O_INIT, R_i);
end for
for all {C_i} ∈ O_INIT do
    O_G ⇐ ConstructOntology({C_i}, GoogleSets);
end for
for all d_j ∈ D_2 do
    for all O_i ∈ {O_INIT, O_S, O_A, O_S+A, O_R, O_G} do
        {E_i} ⇐ ExtractEntity({O_i});
    end for
end for
```

Algorithm 3 shows the overview of the whole biomedical ontology-based entity extraction process. In the first *"for"* loop the initial ontology O_{INIT} is expanded using semantic relationships. We denote the resulting ontology as O_R. Alternatively, in the second *"for"* loop the initial ontology O_{INIT} is expanded using the *GoogleSets* approach, which is an

example of set expansion technique also applied to named entity recognition task[76]. The limitation of using *GoogleSets* expansion approach is the absence of any explicitly defined relationships between newly-discovered concepts and concepts from the initial set (*e.g., foot-and-mouth disease* and *FMDV* are not related). We denote the ontology automatically-constructed using *GoogleSets* by O_G.

After expanding the initial ontology using the two approaches described above, we perform entity extraction at third *"for"* loop using manually-constructed ontologies - O_{INIT}, O_A, O_S, O_{S+A} and automatically built ontologies O_R and O_G. To summarize, the objective of the entity extraction task is to resolve domain-specific terminology extraction, segmentation and normalization subtasks as described above.

5.1.5 Experimental Design and Results: Experiment B

For ontology-based biomedical entity extraction in the domain of veterinary medicine, we aim to extract entities that match at least one concept in the ontology such as a disease or one of its synonyms, abbreviations, causative viruses or disease serotypes. We compared results for domain-specific biomedical entity extraction from different ontologies as summarized in Figure 5.2:

- first, we used the manually-constructed ontologies O_{INIT}, O_S, O_A, O_{S+A};
- second, we used the ontology O_R obtained based on semantic relationship extraction approach;
- third, we used the new ontology O_G based on *GoogleSets* expansion approach .

To compare and evaluate the ontologies that we designed, we retrieved 2000 domain-specific web documents using *Google*, including pdfs that report animal disease outbreaks. Next, we sampled 200 documents where the distribution of the domain-specific entities is sufficient enough for the accurate evaluation of the proposed approach (*e.g.,* number of the disease names is more that 5 for each document). To avoid any bias in terms of overlap

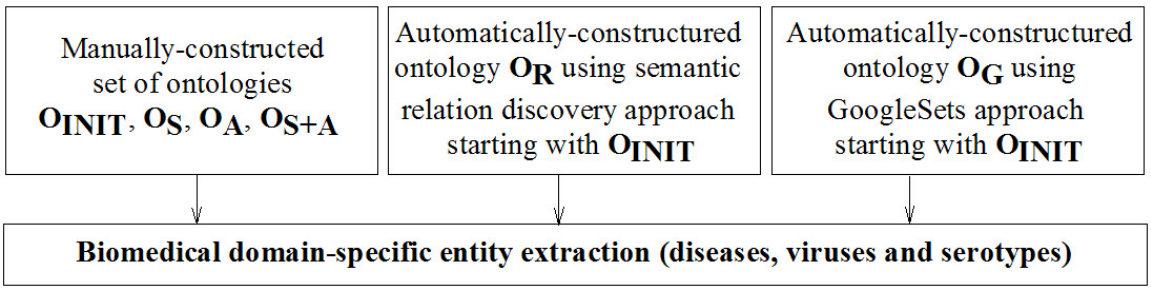


Figure 5.2: Summary of the ontologies used for entity extraction

between learning and validation data, we used first 100 documents to construct the ontology O_R. The other 100 documents were used to evaluate the entity extraction results obtained with all ontologies. The size of the collection used for evaluation of the extraction results is constrained by the effort required for manual annotation of the domain-specific entities. However, the number of documents that are used for new concept learning and automated-ontology construction should be potentially increased.

As a result of using manually and automatically-constructed ontologies, we performed domain-specific entity extraction task and obtained sets of entities $\{E_1, E_2 \dots E_n\}$ and their attributes for each document $D_i \in C$ in the collection, as described in Figure 5.1.

In Figure 5.3, we report results for the different ontologies we used in terms of precision and recall, where precision represents the number of correctly extracted entities divided by the total number of extracted entities and recall/sensitivity represents the number of correctly extracted entities divided by total number of existing correct entities in the collection. Points from left to right represent the values obtained using: manually constructed ontology O_{INIT} - 429 concepts, ontology with manually-collected synonyms and abbreviations O_{S+A} - 605 concepts, ontology O_G learned using *GoogleSets* expansion approach - 754 concepts, ontology O_R constructed using semantic relationship extraction - 772 concepts.

As expected, an increase in precision and recall is achieved when switching from the manually-constructed initial ontology O_{INIT} to an ontology which is also manually built, but enriched with synonyms and abbreviations O_{S+A}. Furthermore, the precision and recall values obtained using the automatically-constructed ontologies O_R and O_G are higher com-

pared to the values obtained using the manually-constructed ontologies. As can be seen, the ontology O_R that is built using the semantic relationship extraction approach achieves the highest precision value of 0.84 and recall value 0.77 compared to manually-constructed ontology O_{INIT} recall 0.25 and precision 0.54.

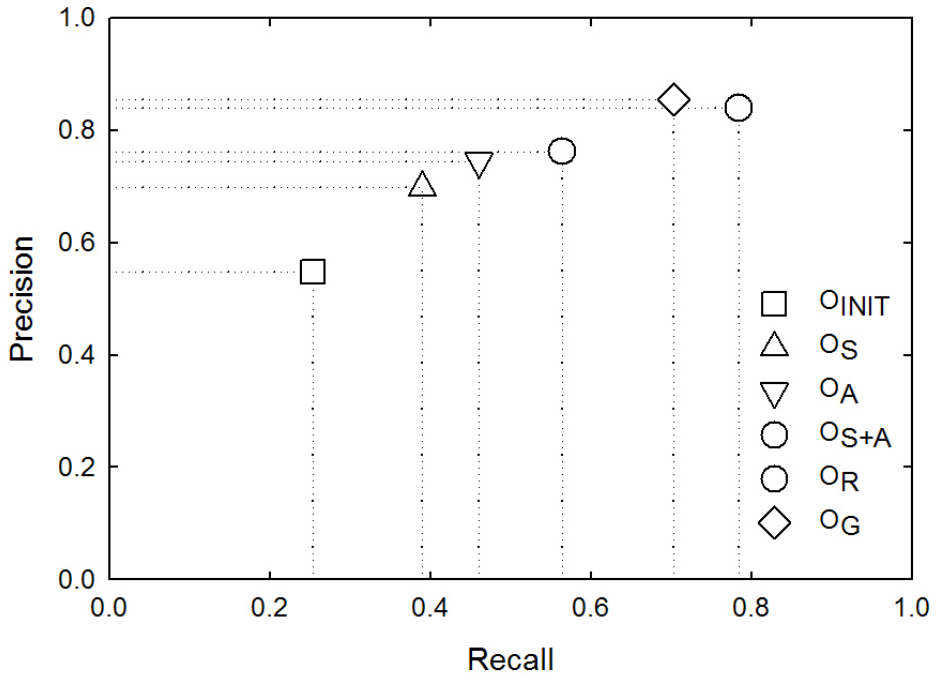


Figure 5.3: Entity extraction results using different ontologies

Figure 5.4 presents the ROC curves corresponding to the entity extraction results obtained using different ontologies. As can be seen, the results obtained using manually-constructed ontologies O_{INIT}, O_S, O_A, O_{S+A} are inferior compared to the results obtained using automatically-constructed ontologies O_R and O_G.

In Figure 5.5 we report F-score values obtained by using different ontologies for entity extraction as a function of the ontology size: (1) initial ontology O_{INIT}, (2) O_S with synonyms, (3) O_A with abbreviations, (4) O_{S+A} with synonyms and abbreviations, (5) *Google-Sets* for O_G and (6) and relationship extraction for O_R. As we have seen, F-score values increase with transitions from O_{INIT} to O_{S+A} through O_S and O_A. The results obtained

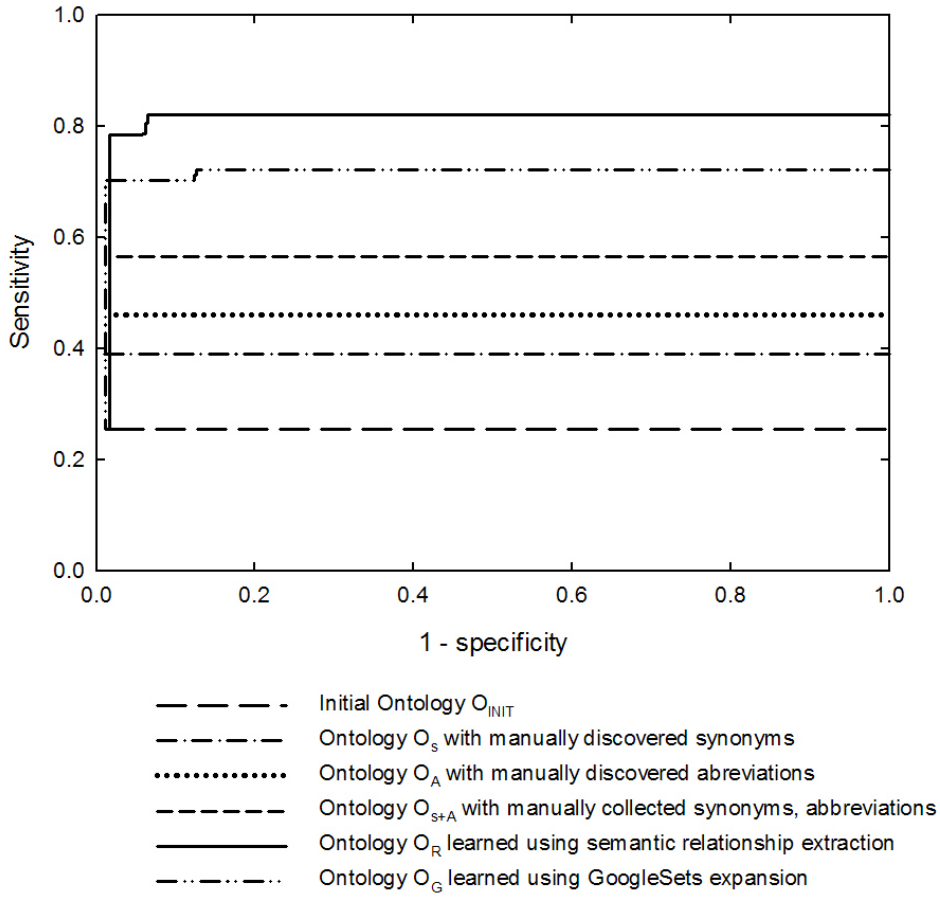


Figure 5.4: ROC curves for manually *vs.* automatically-constructed ontologies

using automatically-constructed ontologies O_R and O_G are much higher in comparison to the results obtained using manually-constructed semantic ontologies. However, when the size of the automatically-constructed ontologies O_R and O_G increases, we can see the drop in F-score. It means that we started to add spurious entities and relationships to the ontologies O_R and O_G. For example, the lowest F-score for the ontology $|O_R| = 1287$ concepts equals 0.63 compared to the highest 0.8 when $|O_R| = 773$ concepts. Similarly, the lowest F-score for the ontology $|O_G| = 1238$ concepts equals 0.43 compared to the highest 0.75 when $|O_R| = 775$ concepts.

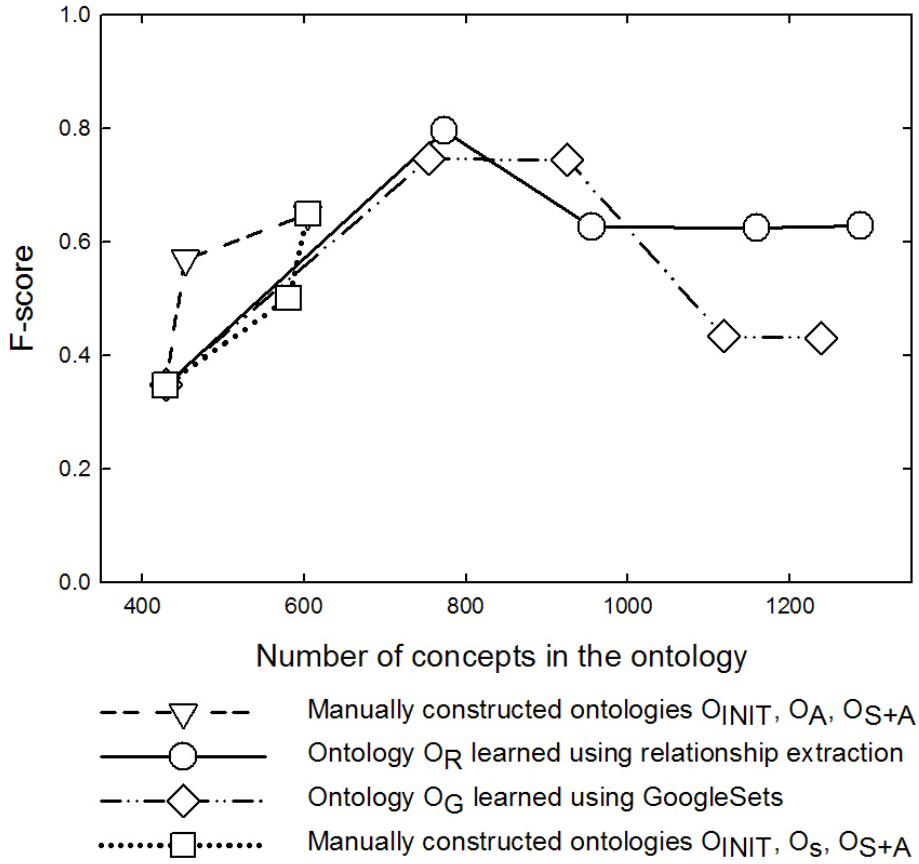


Figure 5.5: F-score values as a function of the ontology size

All results show that enriching the ontology by discovering additional concepts using relationship extraction or *GoogleSets* expansion approaches, brings new domain-specific knowledge and, therefore, allows boosting domain-specific biomedical entity extraction results. However, the concepts that are newly added to the ontology may add noise if they are based on spurious relationships. For instance, results obtained using *GoogleSets* expansion approach for discovering disease synonyms or causative viruses, contain many irrelevant concepts and do not capture any relationship between them explicitly, in comparison to the semantic relationship extraction approach.

5.2 Entity Extraction using Syntactic Features

5.2.1 Sequence Labeling and Syntactic Feature Extraction

As we discussed in Chapter 2 Section 2.3, there are other than ontology-based approaches for the entity extraction and named entity recognition including HMMs - generative, global constraint model and CRFs - discriminative, global constraint model. CRFs, very popular sequence labeling approach[39], has been successfully applied to the domain-specific entity extraction such as: protein, DNA, RNA recognition in medical literature[46].

In this section we propose our novel domain-specific entity extraction sequence labeling approach that is based on syntactic feature extraction using a sliding window approach. More precisely, for each word w_i in the document D, where D is a sequence of connected words/components $w_1, w_2, ..., w_i, ..., w_n$, we consider the set of syntactic features including:

- POS tag (numeric word-level feature);
- capitalization (binary word-level feature);
- capitalization inside (binary word-level feature for identifying abbreviations);
- position in the sentence (numeric document-level feature);
- position in the document (numeric document-level feature);
- frequency (numeric document-level feature);

Next, we generate corresponding feature vectors $F_1, F_2, ..., F_k$ using sliding window technique in order to take into account existing dependencies between words w_{i-1}, w_i, w_{i+1} in the sequence when, for instance $z = 1$. The customizable size of the sliding window allows generating more specific feature vector, when z is small, or more general feature vector, when z is big, for each component w_i.

As shown in Figure 5.6, our goal is to represent any document as a sequence of connected components and extract corresponding feature vectors $F_1, F_2, ..., F_k$ for each component $w_1, w_2, ..., w_i, ..., w_n$ in the sequence using different sliding window, for instance $z = 3$.

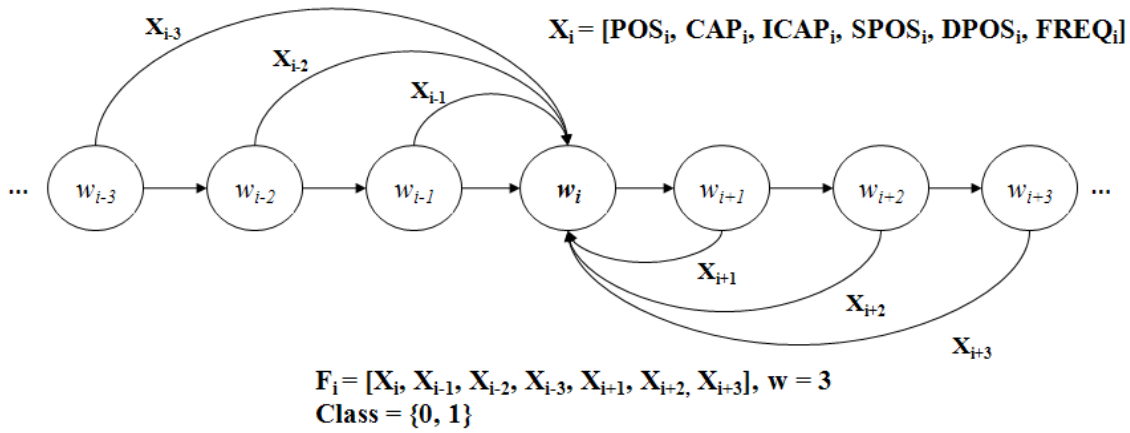


Figure 5.6: Syntactic features extraction approach using sliding window with size $z = 3$

For the target word/component w_i in the sequence $w_{i-z}, \ldots w_i, \ldots w_{i+z}$ within a window $z = 3$, we extract corresponding component features $X_{i-z}, \ldots, X_i, \ldots X_{i+z}$:

$$\begin{cases} X_{i-z} = [POS_{i-z}, CAP_{i-z}, ICAP_{i-z}, SPOS_{i-z}, DPOS_{i-z}, FREQ_{i-z}] \\ \ldots \\ X_i = [POS_i, CAP_i, ICAP_i, SPOS_i, DPOS_i, FREQ_i] \\ \ldots \\ X_{i+z} = [POS_{i+z}, CAP_{i+z}, ICAP_{i+z}, SPOS_{i+z}, DPOS_{i+z}, FREQ_{i+z}] \end{cases} \tag{5.1}$$

We then generate a feature vector F_i for the target component w_i in the sequence $w_{i-z}, \ldots w_i, \ldots w_{i+z}$:

$$F_i = [X_i, \ldots, X_{i-z}, \ldots, X_{i+z}] \tag{5.2}$$

Using the descriptions from the Equation 5.1, we can rewrite the last notation as:

$$\begin{aligned} F_i = & [POS_i, CAP_i, ICAP_i, SPOS_i, DPOS_i, FREQ_i, \ldots \\ & POS_{i-z}, CAP_{i-z}, ICAP_{i-z}, SPOS_{i-z}, DPOS_{i-z}, FREQ_{i-z}, \ldots \\ & POS_{i+z}, CAP_{i+z}, ICAP_{i+z}, SPOS_{i+z}, DPOS_{i+z}, FREQ_{i+z}] \end{aligned} \tag{5.3}$$

Based on our sequence labeling approach, we are able to represent documents as a sequence of positively or negatively labeled components ("1" for disease, "0" otherwise), and generate the set of feature vectors $F_1, F_2, ..., F_k$ for each component w_i in each sequence within a different window $w_{i-z}, \ldots w_i, \ldots w_{i+z}$.

Let us consider an example document D with only one sentence - *"Severe disease in dairy cattle caused by Salmonella Newport"*.

In Figures 5.7 and 5.8 we show the examples of syntactic feature extraction for negatively (Class=0) and positively (Class=1) labeled examples respectively.

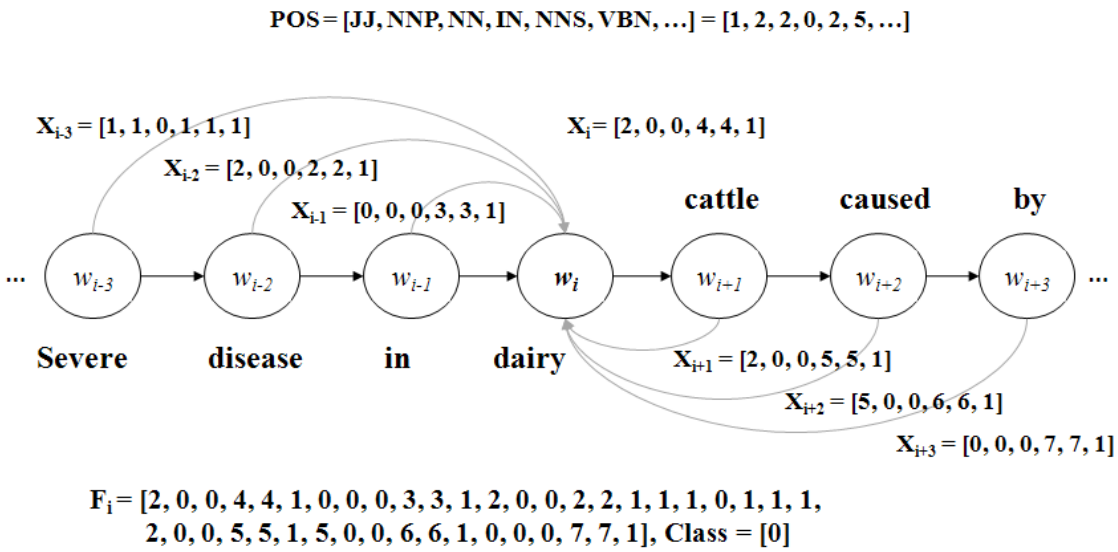


Figure 5.7: An example of the syntactic feature extraction for negatively labeled example $w_i =$ "*dairy*" within a window size $z = 3$

More precisely, we extract corresponding syntactic feature component vector X_i for target component w_i as well as corresponding component vectors $\ldots X_{i-z} \ldots X_{i+z}$ for the adjacent components $w_{i-z} \ldots w_i$ and $w_i \ldots w_{i+z}$ within a predefined window as shown in Equation 5.1. In our example, the window size is $z = 3$, so we are looking for 3 components/words before the target component and 3 components after the target component w_i in the sequence. Next, we generate the complete feature vector F_i for the labeled target component w_i in accordance to Equations 5.2 and 5.3.

If there are empty components before or after the target word w_i, we are unable to extract corresponding feature vectors $\ldots X_{i-z} \ldots X_{i+z}$. In this case, we fill in the missing values in the feature vector F_i with -1 (as shown in Figure 5.8 for positively labeled example, which is the one word before the last word in the document).

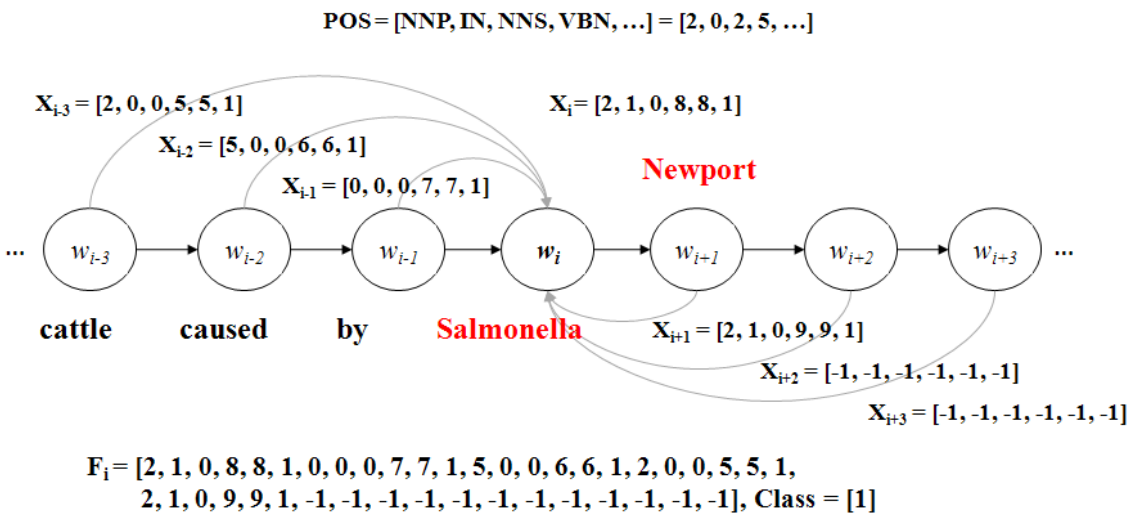


Figure 5.8: An example of the syntactic feature extraction for positively labeled example w_i = "*Salmonella*" within a window size $z = 3$

5.2.2 Experimental Design and Results: Experiment C

For the evaluation of the proposed domain-specific entity extraction approach based on the sequence labeling using syntactic features and sliding window, we used the same set of 100 documents from Experiment B. We did not convert words into lower case, because we were looking for the capitalization feature. However, we removed all stop words because they can not bring any meaningful knowledge to the data, but increase the size of the dataset dramatically. As a result, we extracted a set of feature vectors F_i for a different window size as presented in Table 5.2.

We used several learners including *Naive Bayes, J48, Random Forest, Logistic, OneR, AdaBoostM1* and reported averaged results over 3 runs for each of them. Since we had approx. 202977 examples in out dataset, we used 80% of the data for training and the remaining 20% for testing (approx. 160000 training and approx. 40000 testing examples).

We report the results of disease entity extraction in terms of precision, recall, and Area Under Curve (AUC) in Table 5.3. We do not report accuracy because the data set is unbalanced (approx. 8570 positive examples *vs.* approx. 194430 negative examples).

Table 5.2: An experimental set up for the syntactic feature extraction with different sliding window size

Window Size	Feature Vector Size	Description
$z = 0$	$F = [7 \times 202977]$	Only target word w_i
$z = 1$	$F = [13 \times 202977]$	w_i + 1 word after || before
	$F = [19 \times 202977]$	w_i + 1 word before && after
$z = 3$	$F = [28 \times 202977]$	w_i + 3 words after || before
	$F = [49 \times 202977]$	w_i + 3 words before && after
$z = 5$	$F = [37 \times 202977]$	w_i + 5 words after || before
	$F = [67 \times 202977]$	w_i + 5 words before && after
$z = 7$	$F = [49 \times 202977]$	w_i + 7 words after || before
	$F = [91 \times 202977]$	w_i + 7 words before && after

Let us briefly summarize the highest precision, recall and AUC values obtained using different feature sets and different classifiers from Table 5.3:

- the highest values of recall are equal to 1 for all feature representations obtained using *AdaBoostM1* inducer;
- the highest values of precision for all feature representations are 0.968 obtained using *Random Forest* learner when the feature vector F_i is small; however, when the size of the feature vector F_i increases, the highest precision values are obtained using *J48*.
- the highest values of AUC are obtained using *Random Forest* classifier an are in range [0.691..0.782].

In Figures 5.9 and 5.10 we report F-measure values obtained using different learners *vs.* the size of the feature vectors F_i, respectively. As can be seen for the figures, F-measure values decrease for the *Naive Bayes* classifier when the length of the feature vector F_i increases.

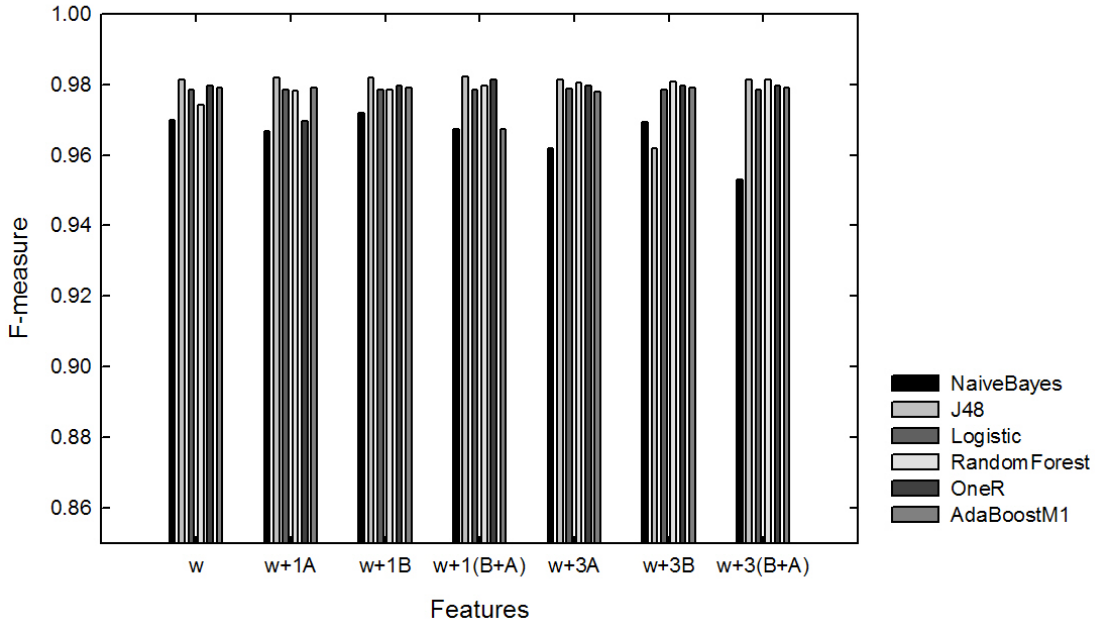


Figure 5.9: The results of disease entity recognition using syntactic features in terms of F-measure, $z = [1..3]$

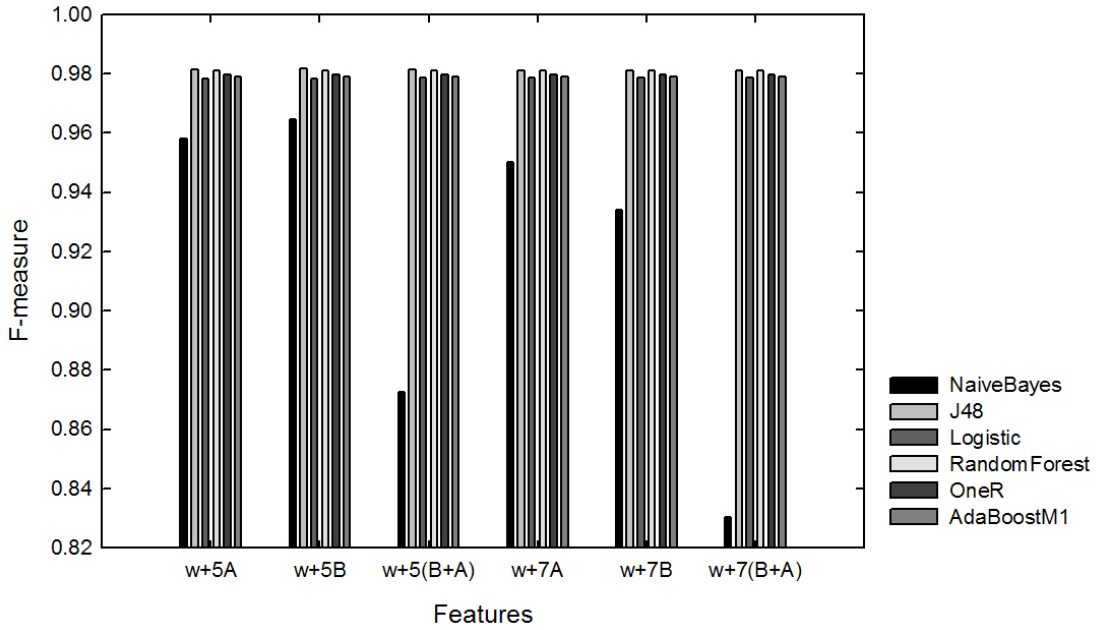


Figure 5.10: The results of disease entity recognition using syntactic features in terms of F-measure, $z = [5..7]$

Table 5.3: The results (from the top to the bottom in the cell: precision, recall, AUC) for classifiers trained on different features F_i

Features	Naive Bayes	J48	Logistic	Random Forest	OneR	AdaBoostM1
w_i	0.960	0.965	0.958	**0.968**	0.960	0.958
	0.984	0.997	0.999	0.980	0.999	**1**
	0.739	0.706	0.738	0.691	0.525	**0.757**
$w_i \wedge w_{i+1}$	0.960	0.966	0.958	**0.968**	0.960	0.958
	0.975	0.997	0.999	0.988	0.999	**1**
	0.7	0.714	0.738	**0.771**	0.525	0.758
$w_i \wedge w_{i-1}$	0.961	0.966	0.958	**0.968**	0.960	0.958
	0.984	0.997	0.999	0.988	0.999	**1**
	0.706	0.714	0.739	**0.773**	0.525	0.759
$w_i \wedge w_{i-1} \wedge w_{i+1}$	0.961	0.967	0.958	**0.968**	0.960	0.958
	0.974	0.996	0.999	0.993	0.992	**1**
	0.685	0.730	0.739	**0.782**	0.525	0.759
$w_i \wedge w_{i+1} \wedge$	0.921	**0.966**	0.959	**0.966**	0.960	0.958
$w_{i+2} \wedge w_{i+3}$	0.962	0.996	0.999	0.995	0.999	**1**
	0.674	0.7	0.737	**0.776**	0.525	0.758
$w_i \wedge w_{i-1} \wedge$	0.961	**0.966**	0.958	**0.966**	0.961	0.958
$w_{i-2} \wedge w_{i-3}$	0.978	0.996	0.999	0.995	0.999	**1**
	0.677	0.7	0.739	**0.773**	0.525	0.759
$w_i \wedge w_{i-1} \wedge \ldots$	0.963	**0.967**	0.958	0.965	0.9608	0.958
$w_{i-3} \wedge w_{i+1}$	0.943	0.995	0.999	0.998	0.999	**1**
$\cdots \wedge w_{i+3}$	0.661	0.772	0.739	**0.775**	0.525	0.758
$w_i \wedge w_{i+1} \wedge$	0.963	**0.967**	0.958	0.964	0.961	0.958
$\cdots \wedge w_{i+5}$	0.953	0.996	0.999	0.997	0.999	**1**
	0.662	0.715	0.734	**0.764**	0.527	0.761
$w_i \wedge w_{i-1} \wedge$	0.963	**0.967**	0.958	0.964	0.906	0.958
$\cdots \wedge w_{i-5}$	0.966	0.996	0.999	0.998	0.999	**1**
	0.660	0.751	0.736	**0.771**	0.527	0.761
$w_i \wedge w_{i-1} \wedge \ldots$	0.966	**0.968**	0.958	0.963	0.960	0.958
$w_{i-5} \wedge w_{i+1}$	0.796	0.994	0.999	0.999	0.999	**1**
$\cdots \wedge w_{i+5}$	0.647	0.751	0.753	0.**757**	0.527	0.761
$w_i \wedge w_{i+1} \wedge$	0.963	**0.967**	0.958	0.963	0.960	0.958
$\cdots \wedge w_{i+7}$	0.937	0.995	0.999	0.998	0.999	**1**
	0.655	0.740	0.734	**0.765**	0.527	0.713
$w_i \wedge w_{i-1} \wedge$	0.964	**0.967**	0.958	0.964	0.961	0.958
$\cdots \wedge w_{i-7}$	0.906	0.995	0.999	0.998	0.999	**1**
	0.651	**0.764**	0.737	0.762	0.527	0.761
$w_i \wedge w_{i-1} \wedge \ldots$	0.967	**0.968**	0.958	**0.968**	0.960	0.958
$w_{i-7} \wedge w_{i+1}$	0.728	0.993	0.999	0.994	0.999	**1**
$\cdots \wedge w_{i+7}$	0.639	**0.745**	0.735	**0.745**	0.527	0.761

5.3 Summary and Discussion

In this Chapter we have presented several approaches for the domain-specific entity extraction including animal disease names, their synonyms, abbreviation, corresponding viruses and disease serotypes. We applied two methods for the entity extraction: first, automated ontology construction approach by learning semantic relationships between concepts using syntactic patterns and second, sequence labeling approach based on syntactic feature extraction and different sliding window. In order to conclude, let us review our entity extraction results together with the results from the existing surveillance systems and related works:

- *BioCaster* named entity recognition system uses 200 news articles and achieves F-score as high as 76.97% for all named entity classes using Support Vector Machines and feature window -2/+1 including surface word, orthography, biomedical prefixes/suffixes, lemma, head noun and previous class predications[38].

  When we apply our automatically constructed ontology, we report the highest F-score 81.7%. Moreover, using our sequence labeling approach with sliding window and syntactic feature extraction, we achieve F-measure as high as 78.2% (*Random Forest* classifier and sliding window is w_i -1/+1). Based on our results, we can conclude that syntactic features bring additional knowledge and are more useful for entity extraction compared to other word-level features.

- The other paper that presents an assessment of disease named entity recognition on a corpus of annotated sentences, reports several methods including *UMLS Metathesaurus*[7] look-up, statistical approach based on term frequency, *MetaMap*[8] lexical look-up and several voting methods (vote 1 - when the disease is proposed by at least one method; vote 2 - when two methods agree; vote 3 - when all three methods agree)[36]. We present their highest results for disease named entity recognition in Table 5.4 for easier visual comparison with our results.

[7]UMLS Metathesaurus - `http://www.nlm.nih.gov/research/umls`
[8]MetaMap - `http://mmtx.nlm.nih.gov/`

Table 5.4: The disease named entity recognition results in terms of precision, recall and F-measure from[36] *vs.* our results

NER Approach	Precision	Recall	F-measure
UMLS Look-up	73.6	68.1	70.7
Term Frequency	58.8	76.3	66.4
MetaMap	77.0	54.3	63.7
Vote 1	56.2	87.0	68.3
Vote 2	76.9	69.8	73.2
Vote 3	89.2	42.3	57.4
Ontology-based	84.8	78.9	**81.7**
Syntactic Features	**96.8**	**99.3**	78.2

– Moreover, there are other works that report entity extraction results in the biomedical domain including protein name, DNA, RNA, cell type. For instance, Lee et. al. use orthographic features together with Support Vector Machines and report the highest F-score 77.9% during the identification phase and 66.9% during the classification phase. However, when they add their dictionary look-up post-processing phase, they are able to increase the classification results up to 79.9% on the identification phase and up to 66.5% on the classification phase[83]. Therefore, it is necessary to combine our syntactic features together with ontology look up for further boosting the entity extraction results in the domain of veterinary epidemiology.

Finally, there is a very useful survey on named entity recognition and classification by Nadeau et.al.[51] and a lecture about sequence models by Pereira[9]. It will be useful to apply this material during our future work in order to improve the results of our domain-specific entity extraction approaches by adding different feature representations, for instance:

– word-level features (punctuation, morphology - prefix/suffix, stem, function - n-grams);

– document-level and corpus features (anaphora/co-reference, meta information, corpus frequency - co-occurrences).

[9]Sequence Models - `http://www.cis.upenn.edu/~pereira/classes/CIS620/lectures/CRFs.pdf`

Chapter 6

Animal Disease-Related Event Recognition

6.1 Sentence-based Event Recognition Methodology

In this Section we describe in detail our methodology for identifying disease-related events and their associated confirmation status. The confirmation status refers to an event being suspected or confirmed. This information is important with respect to the action that might need to be taken. Our approach to the event recognition problem involves three main steps:

- first, we perform entity recognition from unstructured sources;
- next, we classify the sentences from which entities are extracted as being related to an event or not; furthermore, if they are related to an event we classify them as confirmed or suspected;
- finally, we combine entities within an event sentence into structured tuples.

Figure 6.1 illustrates these three steps through an example.

6.1.1 Entity Recognition

The entity recognition module in our system automatically extracts structured information related to animal diseases from unstructured web documents. To achieve this functionality

we associate meta-data in the form of ontologies with documents in our collection. Specifically, the meta-data consists of *domain-independent* location and time hierarchies (including names of countries, states, cities; and canonical dates) and a *domain-specific* medical ontology (including diseases, serotypes, and viruses). Based on these ontologies and pattern matching, we design specialized extractors that locate and classify atomic elements into predefined categories such as:

- disease names (*e.g., "foot and mouth disease", "rift valley fever")*;
- viruses (*e.g., "picornavirus")* and serotypes (*e.g., "Asia-1")*;
- species (*e.g., "sheep", "pigs", "cattle")*;
- locations of events specified at different levels of geo-granularity (*e.g., "United Kingdom", "eastern provinces of Shandong and Jiangsu, China")*;
- dates in different formats (*e.g., "last Tuesday", "two month ago").*

For the animal disease name recognition, we developed an Animal Disease Extractor (DSEx)[1], which relies on a medical ontology, automatically-enriched with synonyms and causative viruses[75]. For species extraction we use pattern matching on a stemmed dictionary of animal names from Wikipedia[2]. Furthermore, we used the Stanford NER[3] tool (which uses conditional random fields) together with NGA GEOnet Names Database (GNS)[4] for location recognition and set of regular expressions for date/time extraction.

The top panel in Figure 6.1 shows a paragraph where entities recognized by our extractors are highlighted. As an example, the output from our entity recognition module for the sentence *"Taiwan's TVBS television station reports that agricultural authorities confirmed foot-and-mouth disease on a hog farm in Taoyuan"* is shown below:

[1]KDD DSEx - `http://fingolfin.user.cis.ksu.edu:8080/diseaseextractor/`
[2]Species in Wikipedia - `http://en.wikipedia.org/wiki/List_of_animal_names`
[3]Stanford NER - `http://nlp.stanford.edu/ner/index.shtml`
[4]GNS - `http://earth-info.nga.mil/gns/html/`

– animal diseases - *"foot-and-mouth disease"* (recognized by the DSEx);

– locations - *"Taoyuan"* (recognized by the Location Extractor);

– species - *"hog"* (recognized by the Species Extractor).

More precisely, Figure 6.1 describes main steps of the proposed approach: first, entities are recognized using several extractors; second, the true event sentences are identified and classified as suspected or confirmed; next, instances from true event sentences are grouped together into potential event tuples; finally, instances of the same event are consolidated into one comprehensive tuple.

6.1.2 Event Sentence Classification

After the entities are recognized in a document, we next extract sentences that contain such entities and classify them as corresponding to true events or false positive events. True events should include a disease name together with a disease-related verb. Furthermore, these events are classified as confirmed or suspected using the Confirmation Status Extractor. This extractor relies on a restricted list of verbs that suggest confirmed events (*e.g., happened*) or suspected events (*e.g., catch*) and their synonyms identified using *GoogleSets* or *WordNet*[27]. For example, the following sentence is classified as corresponding to a confirmed event: *"On 9 Jun 2009, the farm's owner reported symptoms of FMD in more than 30 hogs."*

The initial list of verbs consists of single word verbs/nouns (*e.g., kill*) and verb/noun phrases (*e.g., strike out*). The first two columns in Table 6.1 show the number of initial verbs denoted as $IN\text{-}V$ and verb phrases denoted as $IN\text{-}VP$ for both suspected and confirmed categories. Columns 3 and 4 show similar numbers for the augmented list of verbs obtained using *GoogleSets* ($GS\text{-}V$, $GS\text{-}VP$ respectively), while columns 5, 6 show these numbers for *WordNet* ($WN\text{-}V$, $WN\text{-}VP$ respectively). The complete list of these indicative nouns/verbs is in Table A.2 in Appendix A. The list of verbs used to classify sentences as confirmed or suspected is also useful for eliminating frequent, but not event-related sentences such as: "*Foot and mouth disease is[V] a highly pathogenic animal disease*".

Table 6.1: Statistics about the restricted list of verb features for the event recognition

Status	IN-V	IN-VP	CS-V	GS-VP	WN-V	WN-VP
Suspected	7	1	55	2	37	10
Confirmed	7	1	55	13	48	9

The second step in Figure 6.1 shows more examples of potential event-related sentences and their classification. We first classify sentences as event-related - *"YES"* or event non-related - *"NO"*. We then classify event-related sentences as suspected or confirmed based on the restricted list of verbs/verb phrases represented in Table 6.1.

6.1.3 Event Tuple Generation

An *event* is an occurrence of a disease within a particular time and space range. We use four main event attributes to specify an event: disease name, date, location, species. In addition, as we extract events automatically from crawled web documents, we also include an attribute that specifies the confirmation status of an event. Thus, an event can be described as a tuple of the following form:

$$Event_i \stackrel{\text{def}}{=} < disease, date, location, species, status >, \tag{6.1}$$

where each attribute in the tuple is obtained with one of the extractors described in Section 6.1.1. The following tuple <*FMD, 9 Jun 2009, Taoyuan, hog, confirmed*> is an example of an event. Given the incomplete and the uncertain nature of the information available online, it is possible for events to have missing values, *e.g.,* $< disease, ?, location, species, ? > = < FMD, ?, Taoyuan, hog, ? >$, $< disease, date, ?, species, ? > = < FMD, 06/09/09, ?, hog, ? >$. For instance, news reports can contain information about disease-related events that happened in some location without a specific date or species being provided.

Furthermore, several sentences in a document can contain information about the same event and we aggregate the corresponding event tuples into a unique tuple based on the attributes available, as shown in the last step in Figure 6.1.

In addition to abovementioned main event attributes, there are other infinite set of

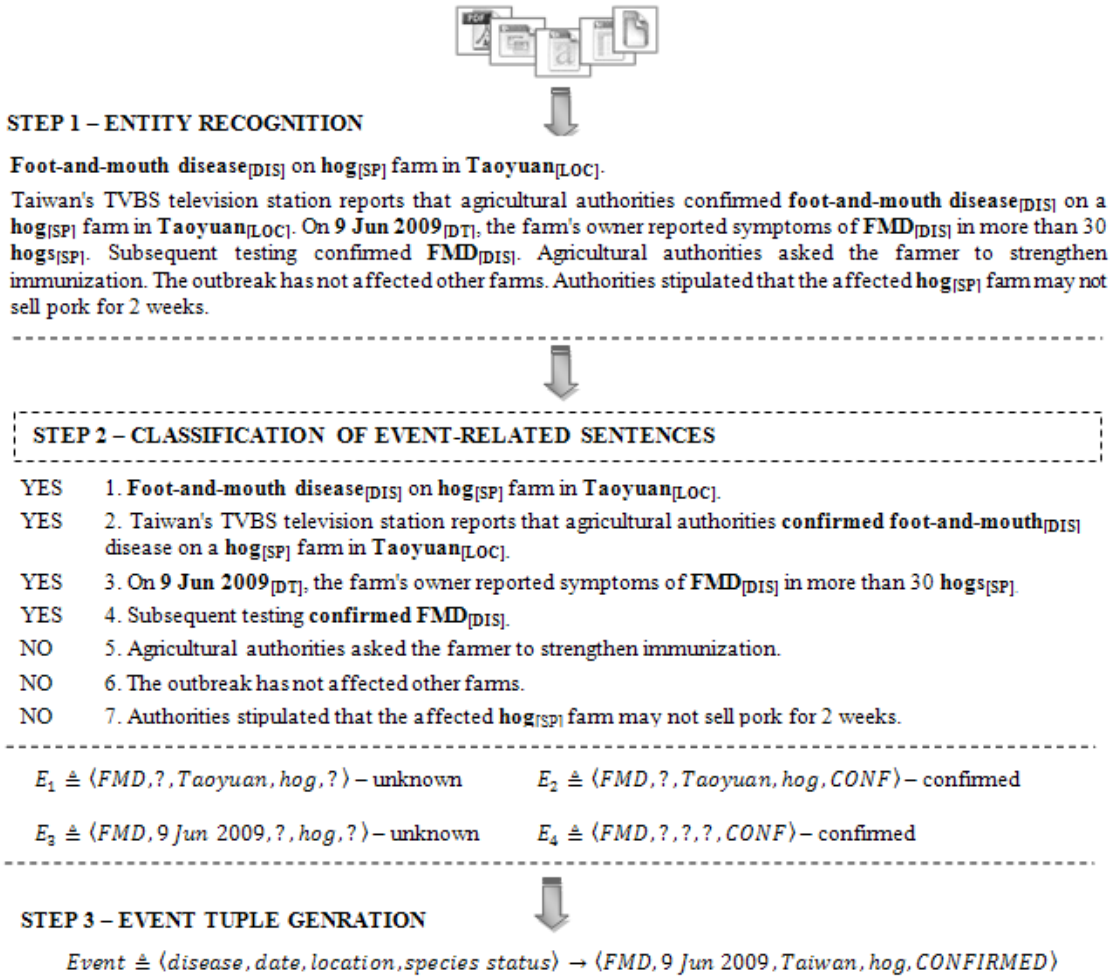


Figure 6.1: Description of the event recognition approach workflow through an example

attributes related to animal disease events that can possibly be extracted from web-pages such as: date of event report (*e.g.,* 12/18/2007), reporting source (*e.g.,* FMD Institute of Animal Health), morbidity and mortality (*e.g.,* number of animals infected and killed), damage, governmental response, *etc.*

$$Event_i \stackrel{\text{def}}{=} < what; when; where; who; reported\_date; reported\_source... >$$

Algorithm 4 summarizes the steps for entity recognition, event-related sentence classification and tuple generation.

Algorithm 4 Entity Recognition, Sentence Classification and Tuple Generation

```
Input:  Set of web documents D
Output:  Set of extracted events e_k ∈ E for each document d_j ∈ D
   foreach document d_j ∈ D do
    S = TokenizeToSentences(d_j);
     foreach sentence s_i ∈ S do
       disease = ExtractDiseaseEntities(s_i);
       if disease ≠ ∅ then
          status = ExtractConfirmationStatus(s_i);
          if status ≠ ∅ then
             date = ExtractDateEntities(s_i);
             location = ExtractLocationEntities(s_i);
             species = ExtractSpeciesEntites(s_i);
         else
             skip sentence s_i;
          end if;
       else
          skip sentence s_i;
       end if;
     end for;
     E = GenerateTuples(disease, date, location, species, status);
     e_k = AggregateTuples(E);
   end for.
```

6.1.4 Experimental Design and Results: Experiment D

We used the existing *DUCView Pyramid* scoring tool[52] to score automatically generated event tuples and evaluate our approach. Pyramid scoring is a technique for evaluating summarization results, which was introduced in[53] and relies on multiple summaries to assign the significance weights to summarization content units (*i.e.*, entities)[54].

To perform the evaluation, we used Google to retrieve 100 documents related to two animal diseases: rift valley fever (RVF) and foot-and-mouth disease (FMD). We manually created two sets of summaries for each of the 100 documents and extracted entities corresponding to event tuples from each summary and each document. Then, we used the *DUCView* tool to compare automatically generated event tuples with entities from human summaries. As a result, the entities from event tuples are assigned weights in the range [0, 1]

where 1 represents the best recognition score and it means that entity from automatically-generated tuple is present in all summaries. The entity weights are used to calculate an aggregated score for event tuples. Specifically, the score for an event tuple described in Equation 6.1 is given by:

$$Score_i =< w_d disease, w_t date, w_l location, w_s species, w_c status > \quad (6.2)$$
$$subject\ to\ disease + status = 2$$

where $disease, \cdots, species$ take 0/1 values (entity present or not in the tuple) and a tuple is valid only if both *disease* and *status* are present. The resulting scores are reported as a measure of the accuracy of the proposed event tuple recognition and classification approach and shown in Table 6.2.

More precisely, we evaluate our event tuple recognition and classification approach by applying three lists of verbs and verb phrases for confirmation status extraction which are introduced in Table 6.1. Furthermore, we consider stemmed S *vs.* non-stemmed NS versions of these lists. The results for the non-stemmed version of the lists are shown in the first three columns of the Table 6.2 for the initial list, *GoogleSets* augmented list and *WordNet.* augmented list, respectively. Similarly, the results for the stemmed version are shown in the last three columns of the Table 6.2.

Table 6.2: Pyramid Event Score Distribution by Range

Score Range	IN-NS	GS-NS	WN-NS	IN-S	GS-S	WN-S
Low [0 - 0.3]	73%	43%	38%	19%	18%	13%
Medium [0.31 - 0.7]	18%	27%	29%	27%	30%	13%
High [0.71 - 1]	9%	30%	33%	54%	52%	74%
Average Score	**0.17**	**0.40**	**0.45**	**0.64**	**0.65**	**0.75**

As can be seen from the Table 6.2, the initial list of verbs results in many low score events which means that not many tuples can be extracted with high confidence using only these verbs. While the augmented lists, without stemming, give better results, only approximately one third of the events are scored with a high confidence for both *GoogleSets* and *WordNet.*

However, the scores increase significantly for all lists when stemming is used. The best results are obtained for the *WordNet*

Figures 6.3a, 6.3b and 6.3c show the comparative histograms of the event score distribution using the initial, automatically augmented with *GoogleSets* and *WordNet* lists, for both stemmed and non-stemmed versions of lists. As can be seen, more events are identified using the *WordNet* list and they have higher scores (many of them have the max score 1).

Figure 6.2 demonstrates the dependency between the quality and the size of the list of verbs and the average event recognition score.

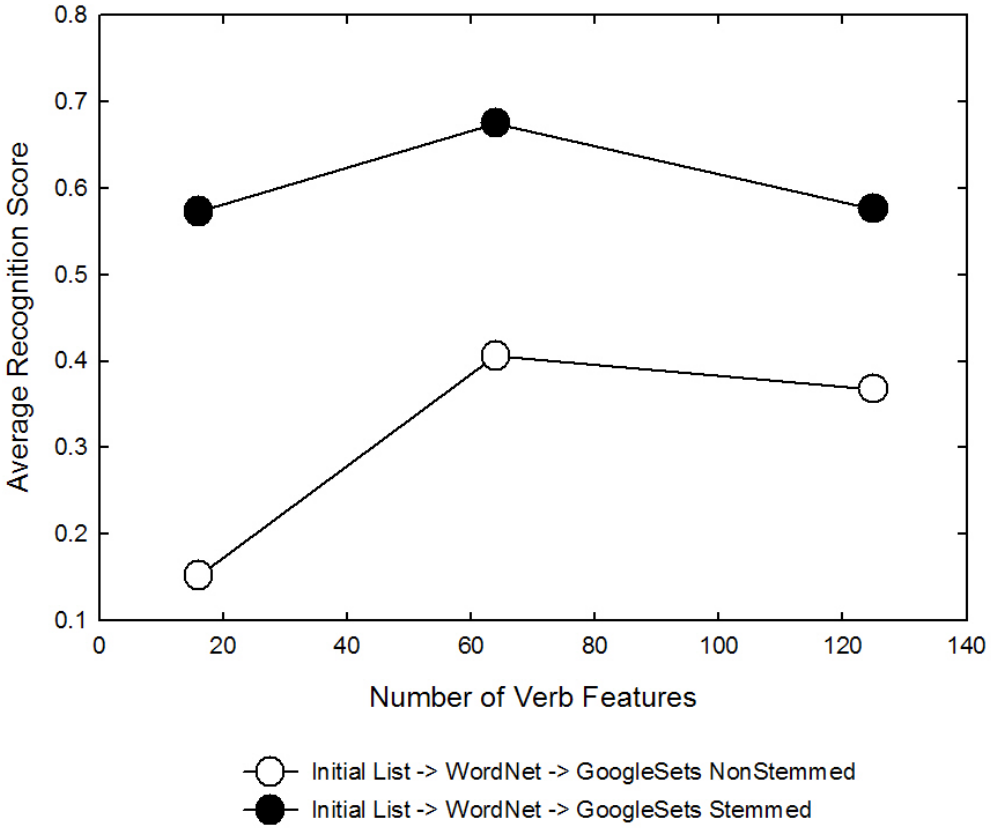
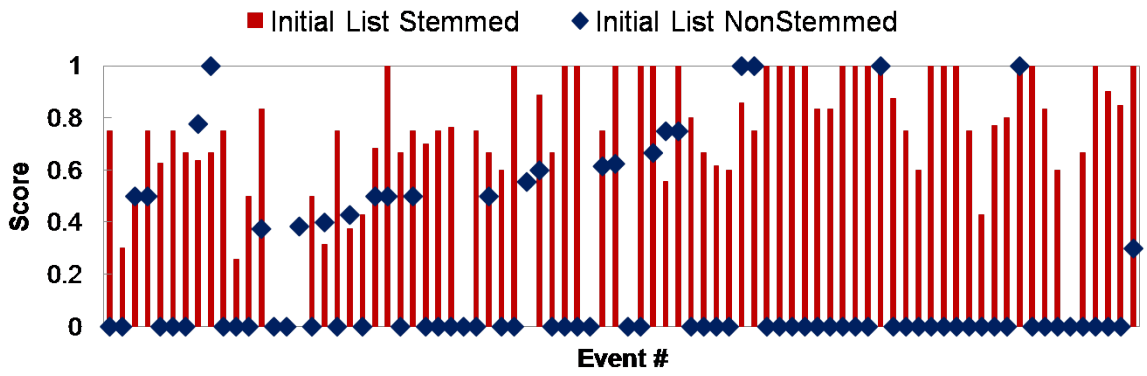
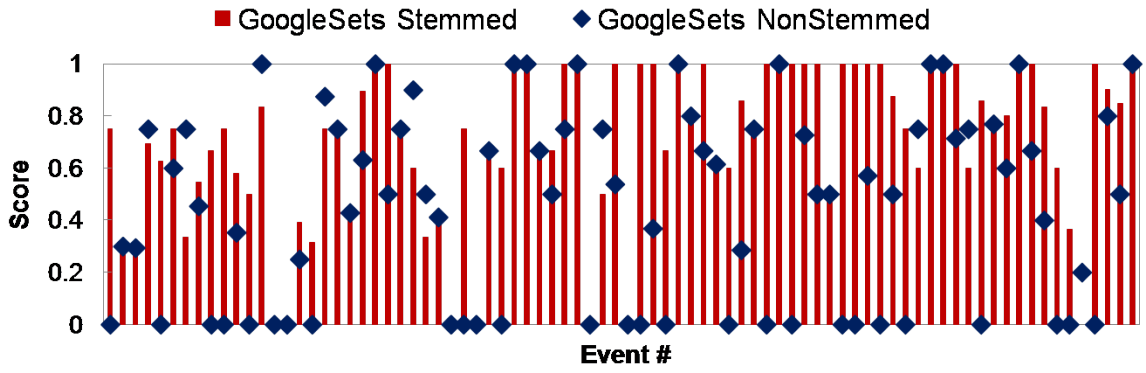


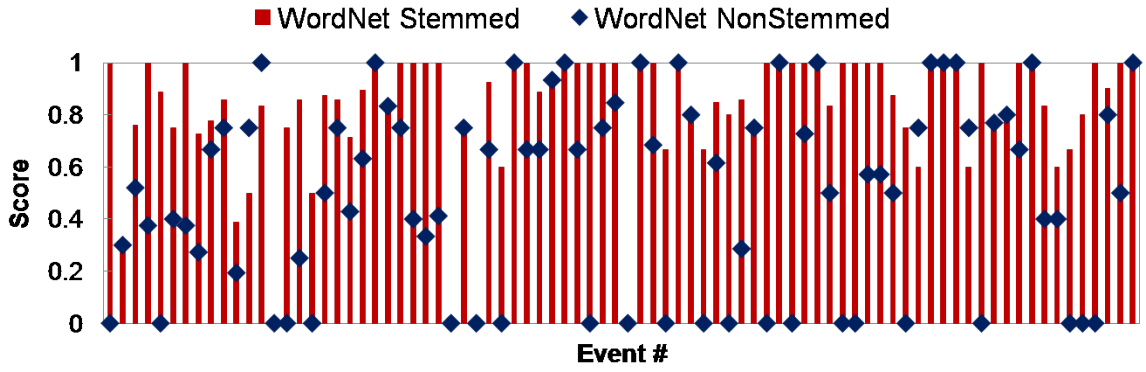
Figure 6.2: Event recognition score *vs.* the size of the list of stemmed/unstemmed n-grams



(a) Event score distribution using Initial list of verbs



(b) Event score distribution using augmented with *GoogleSet* list of verbs



(c) Event score distribution using augmented with *WordNet* list of verbs

Figure 6.3: Event recognition score histograms for different n-gram lists: initial list *vs.* list augmented using *GoogleSets vs.* list augmented using *WordNet*

6.2 Event Recognition for Predictive Epidemiology

The predictive epidemiology is an area for modeling of animal infectious disease spread and suggesting optimal mitigation strategies to control the impact on the society and environment. The input data for such spatial-temporal predictive models is usually secured or restricted in use. Such restrictions lead to the information incompleteness that, in turn, drops the accuracy of the model prediction. Since a plethora of animal disease-related data is available online, it can be used as an input for these predictive models.

In this Section we present ongoing research on the text mining project in the domain of epidemiology for animal disease event extraction and classification into predefined categories such as: susceptible, infected and recovered. We consider similar event attributes as described in Section 6.1 including animal disease names, dates, species with corresponding numbers and geo-referenced locations. Currently, we are interested in extracting data related to the foot-and-mouth disease outbreak in 2001 in United Kingdom [80] from ıProMed-Mail reports [5]. The relevant extracted information will be the input for the spatial-temporal predictive model suggested in [64].

6.2.1 Event Attribute Extraction

An outbreak is a collection of events that are connected by disease name and happened within *restricted space and time* range as formalized in Equation 6.3:

$$Outbreak_j \stackrel{\text{def}}{=} \{E_1, E_2, ...E_n\} or \sum_{i=1}^{n} E_i,$$

where $E_1, E_2, ...E_n$ are events that share the same disease name.

An event is an occurrence of a disease within a particular time and space range. Therefore, the modified main event attributes compared to Equation 6.1 are: disease, date, location, species, numbers, status as specified in Equation 6.3:

$$Event_i \stackrel{\text{def}}{=} < disease, date, location, species, number, status >, \tag{6.3}$$

[5] ProMed-Mail - `http://www.promedmail.org`

For event attribute extraction we need to use methods and tools presented in Section 6.1.1 together with modified species extractor that should correctly extract numbers related to species. We present examples of automatically extracted event attributes for each event status: susceptible, infected, recovered below:

1. *"The signs suggested the 27 pigs could be suffering from foot and mouth disease (FMD) in Island of Anglesey, Wales as reported on 2/21/2001"*

   – *<FMD; 2/21/2001; Island of Anglesey, Wales; 27 pigs; susceptible>*;

2. *"The UK Ministry of Agriculture confirmed on 2/20/2001 that 27 pigs found with vesicles in an abattoir near Brentwood, Essex, have Foot and Mouth Disease"*

   – *<FMD; 2/20/2001; Brentwood, Essex; 27 pigs; infected>*;

3. *"Almost 2000 cattle and more than 15 000 sheep, have been or are waiting to be slaughtered since the resurgence of the disease in Northumberland as reported on 8/31/2001"*

   – *<FMD; 8/31/2001; Northumberland; 2000 cattle, 15 000 sheep; recovered>*.

6.2.2 Event Sentence Status Classification

In order to classify disease-related sentences into three status categories: susceptible, infected or recovered applying our event recognition approach presented in Section 6.1.2 and described in Figure 6.1, we need to consider corresponding list of n-gram patterns for each of the category (similarly to the lists of nouns and verbs in Table A.2). The complete list of these indicative n-grams for susceptible, infected and recovered classes is presented in Table A.1 in Appendix A.

In addition, we present Algorithm 5 for SIR event recognition and classification which is a modified version of Algorithm 4 presented earlier. The main difference is that we are required to extract numbers for species; therefore, we consider species with numbers as a main attribute of the event tuple in current task.

Algorithm 5 Entity Recognition, SIR Sentence Classification and Tuple Generation*

```
Input: Set of documents D tokenized into sentences n_i ∈ N, list of n-grams
for SIR classification from Table A.1
Output: Events e_k ∈ E classified into three categories such as: suspected
or infected or recovered S ∨ I ∨ R
   foreach document d_j ∈ D do
    S = TokenizeToSentences(d_j);
     foreach sentence s_i ∈ S do
        status = ExtractSIRStatus(s_i);
        if status ≠ ∅ then
           species = ExtractSpeciesWithNumbers(s_i);
           if species ≠ ∅ then
              disease = ExtractDiseaseEntities(s_i) &&
              location = ExtractLocationEntities(s_i) &&
              date = ExtractDateEntites(s_i) &&
          else
              skip sentence s_i;
           end if;
        else
           skip sentence s_i;
        end if;
     end for;
     E = GenerateTuples(disease, date, location, species, status);
     e_k = AggregateTuples(E);
   end for.
```

As we mentioned above, we are interested in extracting data about foot-and-mouth disease outbreak that happened in United Kingdom in 2001. For that purpose, we collected 118 related reports from *ProMed-Mail* and applied our event recognition approach.

In Figure 6.4 we present our extracted results on the maps for each month separately. We use 3 colors for three event types, respectively: yellow for susceptible, red for infected and green for recovered state. Moreover, for each location we consider the last corresponding status at the end of the month, *e.g.,* Cumbia was in susceptible state at the beginning of June, then it was in infected state and, finally, it was in recovered state at the end of June.

The extracted results confirm the effectiveness of our event recognition and classification approach, however our results are limited to the information presented in the source.

(a) February (b) March (c) April

(d) May (e) June (f) August

(g) September (h) October (i) December

Figure 6.4: The spread of foot-and-mouth disease outbreak in UK, 2001

6.3 Summary and Discussion

In this Chapter we presented our novel sentence-based approach for disease-related event recognition. We first extracted event attributes and classified sentences as event related or non-related. We then classified event-related sentences into two categories such as: suspected and confirmed based on the predefined set of the indicative n-grams (nouns, verbs and noun and verb phrases). Finally, we generated comprehensive event tuple for every document in the experimental collection.

We applied our event recognition approach for predictive epidemiology domain. Our preliminary results show that automatically extracted data can be effectively used as an input to spatio-temporal models for prediction of the disease spread.

However, it is important to point out the limitations of the proposed sentence-based event recognition approach:

- the accuracy of the event recognition completely depends on the separate entity extraction accuracy; it means that if the accuracy of the location or species extraction is low then it influences the accuracy of the event recognition;
- the co-reference resolution was not implemented, we left it as future work;
- the event aggregation and deduplication requires more comprehensive heuristics and additional knowledge, for example co-reference resolution;
- the event tuple with missing values should be more accurately aggregated into more comprehensive event tuple with all attribute values.

In order to conclude, let us analyze our results for the event recognition task and the results reported by the other surveillance systems:

- *PULS* receives approximately 10000 documents from *MedISys* per month with 27% relevant documents (they report animal disease-related events) and the remaining 73%

does not report any events. Authors evaluated their event recognition functionality by collecting 100 English-language documents with 156 events.

The event recognition results are divided into four categories:

1. Number of correctly identified true events - *True Positive* value is 63 out of 156;
2. Disease names were detected 74 times by system, however there was no event to report (the disease name appeared in the drug, vaccines context);
3. Disease names appeared in other context including politics, sport - 19 misclassified events;
4. Diseases that should have been identified but were not - *False Negatives*, 13 events.

Based on the abovementioned results the reported precision value is 0.88, calculated as category one events and category two events divided by the total number of the events 63 + 74 = 137 out of 156.

Moreover, authors present the other experiment for estimating the number of *False Negatives* using other 200 documents. The system reported no events for all documents, however 14% of them included disease related events.

– *BioCaster* detects approximately 950 disease-location pairs per month as disease-related events from news including *Google News - 26.4%*, *Yahoo News - 30.3%*, *ProMed-Mail -18.6%* and 24.3% for others. The experiment includes 1 month period news. The reported results - 887/950 correct disease-location pairs and 93.4% precision.

We report our event recognition results in terms of accuracy in contrast to *MedISys* and *BioCaster* that report only precision. We insist that recall component is more important than precision for the event recognition. It is highly desirable to recognize all possible events in the documents, for instance if there are 3 events and we are able to recognize only 1 (but with high precision), then the approach is not efficient enough.

Chapter 7

Conclussions

7.1 Summary

Monitoring epidemic crises, caused by rapid spread of infectious animal diseases, can be facilitated by the plethora of information about disease-related events that is available online. Therefore, the ability to use this information to perform domain-specific entity recognition and event-related sentence classification, which in turn can support time and space visualization of automatically extracted events, is highly desirable.

For that purpose, in this thesis we precisely formulated and investigated several research problems such as:

- animal disease-related document classification;
- domain-specific entity extraction;
- animal disease-related event recognition.

For **disease-related document classification** we considered a framework, different feature representations for the documents and different machine learning classification algorithms. We evaluated the document relevance classification using binary features, keyword term frequency and the "bag-of-words" representation using word unigrams and bigrams. Our experimental results demonstrate the effectiveness of our text categorization component in the designed framework for epidemiological analytics.

For **domain-specific entity extraction**, we proposed two different techniques: an ontology-based approach and a sequence labeling methodology using syntactic features with sliding window.

- For our **ontology-based approach**, we used a semantic relationship extraction based on syntactic patterns and POS tagging to construct an ontology (containing animal diseases, their synonyms and viruses). We compared the automatically-constructed ontology obtained using our relationship extraction approach with an ontology constructed using *GoogleSets* expansion approach, which refers to expanding a given partial set of objects into a more complete set. We compared the entities extracted using all abovementioned ontologies in terms of precision and recall, and reported F-measure values as a function of the ontology size. The results show that our semantic relationship extraction approach brings new knowledge to an initial ontology and, therefore, boosts the domain-specific biomedical entity extraction results.

- For our **sequence labeling approach**, we exacted syntactic word-level features (capitalization, POS tagging, abbreviations, term frequency) using a sliding window approach. We evaluated our approach using different machine learning algorithms together with different feature representations and various window sizes. We reported results of the domain-specific entity extraction in terms of F-measure, precision and recall and compared them with the results from the other surveillance systems.

For **event recognition** we presented our novel sentence-based approach for animal disease event recognition and classification. Entity and confirmation status extraction methods are used to automatically generate structured summaries about domain-specific events in the form of tuples. Furthermore, we applied several lists of verbs for confirmation status extraction including *WordNet* and *GoogleSets*. We used *DUCView* tool[52] to calculate scores for automatically generated event tuples, which can be seen as a measure of accuracy of our approach. The highest accuracy was obtained using a *WordNet* augmented list of verbs.

7.2 Contribution

The major contributions of this thesis are presented in 3 strongly refereed papers and 2 poster presentations:

– **Computational Knowledge and Information Management in Veterinary Epidemiology**[74]

We have presented a system for animal disease outbreak analysis by automatically extracting relational information from online data. We aim to detect and map infectious disease outbreaks by extracting information from unstructured sources. The system crawls web sites and classifies pages by topical relevance. The information extraction component performs document analysis for animal disease related event recognition. The visualization component plots extracted events into *GoogleMaps* using spatial information and supports timeline representation of disease outbreaks in *SIMILE*.

– **Boosting Biomedical Entity Extraction by Using Syntactic Patterns for Semantic Relation Discovery**[73]

We have proposed a novel ontology-based biomedical entity extraction approach using semantic relations learning by syntactic pattern matching. We first, manually construct an ontology for extracting entities such as: animal disease names, viruses and serotypes. We then use an automated ontology expansion approach to extract semantic relations between concepts. Such relations include asserted synonymy, hyponymy and causality. The relations are extracted by using a set of syntactic patterns and part-of-speech tagging. The resulting ontology contains richer semantics compared to the manually-constructed ontology. We compare our approach for extracting synonyms, hyponyms and other disease-related concepts, with an approach where the ontology is expanded using *GoogleSets*, on the veterinary medicine entity extraction task. Experimental results show that our semantic relation extraction approach produces a significant increase in precision and recall as compared to the *GoogleSets* approach.

– **Named Entity Recognition and Tagging in the Domain of Epizootics**[75]

We have discussed the web-mining of animal-disease related information from published news articles and publicly-available postings. Previously, such kinds of tasks were performed mostly for human diseases related data. Meanwhile, our task is directly related to web-crawling for extraction animal disease related information. We define the domain-specific information extraction task from crawled unstructured data in the domain of veterinary medicine. This task is related to the development of several modules for tagging entities such as: animal disease names, species, vaccines, serotypes *etc.* The extraction technique is based on a pattern matching approach. The gazetteer is semi-automatically collected from official web-portals and manually enriched with synonymic and causative relationships between related ontology concepts.

– **Animal Disease Event Recognition and Classification**[72]

We have proposed a rule-based approach to the problem of extracting animal disease-related events from web documents. Our approach relies on the recognition of structured entity tuples, consisting of attributes, which describe events related to animal diseases. The event attributes that we consider include animal diseases, dates, species and geo-referenced locations. We perform disease names and species recognition using an automatically-constructed ontology, dates are extracted using regular expressions, while locations are extracted using a conditional random fields tool. The extracted events are further classified as confirmed or suspected based on semantic features, obtained from the *e.g., GoogleSets* and *WordNet*. Our preliminary results demonstrate the feasibility of the proposed approach.

– **Automated Event Extraction and Named Entity Recognition in the Domain of Veterinary Medicine**[71]

We will summarize our results in automated event extraction and classification, domain-specific named entity recognition by their comparison with other surveillance systems.

7.3 Future Work

Several future research directions are discussed below:

– **Domain-specific Entity Extraction**

First, there are several manually or semi-automatically constructed multilingual ontologies for the veterinary epidemiology implemented in *BioCaster*[21] and *MedISys*[12]. Therefore, we will work on the automated multilingual ontology construction for the domain of veterinary medicine using other semistructured sources *e.g., Wikipedia* for entity extraction[60] and entity disambiguation[32].

Second, there are several language independent set expansion approaches of general named entities using web[76] and biomedical entities[45]. Therefore, we plan to extend our semantic relationship extraction approach to other domains and other generalized named entities[19]. Moreover, we plan to enrich the ontology obtained using *GoogleSets* with relationships extracted using our automated-ontology construction approach[73].

– **Event Recognition**

We intend to apply a deeper syntactic analysis of the sentence[37] and part-of-speech tagging in addition to the list of verbs that we used. We plan to consider the negation words, modal words and tense *etc.* Moreover, it is necessary to integrate coreference resolution functionality[70] into our event recognition approach. We need to study how to deal with pronoun resolution when the event is reported in multiple sentences.

Bibliography

[1] Liang Chen A, Naoyuki Tokuda B, and Akira Nagai C, *A new differential lsi space-based probabilistic document classifier*, Information Processing Letters **88** (2003), no. 5, 203–212.

[2] Akiko Aizawa, *The feature quantity: an information-theoretic perspective of tfidf-like measures*, Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval (Athens, GR) (Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, eds.), ACM Press, New York, US, 2000, pp. 104–111.

[3] Khalid Al-Kofahi, Alex Tyrrell, Arun Vachher, Tim Travers, and Peter Jackson, *Combining multiple classifiers for text categorization*, Proceedings of CIKM-01, 10th ACM International Conference on Information and Knowledge Management (Atlanta, US) (Henrique Paques, Ling Liu, and David Grossman, eds.), ACM Press, New York, US, 2001, pp. 97–104.

[4] Akiko Azawa, *Linguistic techniques to improve the performance of automatic text categorization*, Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium (Tokyo, JP), 2001, pp. 307–314.

[5] Douglas Baker and Andrew K. McCallum, *Distributional clustering of words for text classification*, Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (Melbourne, AU) (Bruce Croft, Alistair Moffat, Cornelis J. Van Rijsbergen, Ross Wilkinson, and Justin Zobel, eds.), ACM Press, New York, US, 1998, pp. 96–103.

[6] Roberto Basili and Alessandro Moschitti, *A robust model for intelligent text classification*, Proceedings of ICTAI-01, 13th IEEE International Conference on Tools with

Artificial Intelligence (Dallas, US), IEEE Computer Society Press, Los Alamitos, US, 2001, pp. 265–272.

[7] Roberto Basili, Alessandro Moschitti, and Maria T. Pazienza, *An hybrid approach to optimize feature selection process in text classification*, Proceedings of AI*IA-01, 7th Congress of the Italian Association for Artificial Intelligence (Bari, IT) (Floriana Esposito, ed.), Springer Verlag, Heidelberg, DE, 2001, Published in the "Lecture Notes in Computer Science" series, number 2175, pp. 320–325.

[8] Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, and Yoad Winter, *Distributional word clusters vs. words for text categorization*, Journal of Machine Learning Research **3** (2003), 1183–1208.

[9] Mohamed Benkhalifa, Amine Bensaid, and Abdelhak Mouradi, *Text categorization using the semi-supervised fuzzy c-means algorithm*, Proceedings of NAFIPS-99, 18th International Conference of the North American Fuzzy Information Processing Society (New York, US), 1999, pp. 561–565.

[10] Mohamed Benkhalifa, Abdelhak Mouradi, and Houssaine Bouyakhf, *Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization*, International Journal of Intelligent Systems **16** (2001), no. 8, 929–947.

[11] Brigitte Bigi, *Using kullback-leibler distance for text categorization*, Proceedings of ECIR-03, 25th European Conference on Information Retrieval (Pisa, IT) (Fabrizio Sebastiani, ed.), Springer Verlag, 2003, pp. 305–319.

[12] Konstantin Bogatyrev and Roman Yangarber, *Use of deep syntax parsing in cross-language information extraction*, MLMTA, 2005, pp. 18–24.

[13] Razvan C. Bunescu and Raymond J. Mooney, *Learning to extract relations from the web using minimal supervision*, ACL, 2007.

[14] Lijuan Cai and Thomas Hofmann, *Hierarchical document categorization with support vector machines*, Proceedings of CIKM-04, 13th ACM International Conference on Information and Knowledge Management (Washington, US) (David A. Evans, Luis Gravano, Otthein Herzog, ChengXiang Zhai, and Marc Ronthaler, eds.), ACM Press, New York, US, 2004, pp. 78–87.

[15] Ana Cardoso-Cachopo and Arlindo L. Oliveira, *An empirical comparison of text categorization methods*, Proceedings of SPIRE-03, 10th International Symposium on String Processing and Information Retrieval, Springer Verlag, Heidelberg, DE, 2003, Published in the "Lecture Notes in Computer Science" series, number 2857, pp. 183–196.

[16] Kian M. Chai, Hwee T. Ng, and Hai L. Chieu, *Bayesian online classifiers for text classification and filtering*, Proceedings of SIGIR-02, 25th ACM International Conference on Research and Development in Information Retrieval (Tampere, FI) (Micheline Beaulieu, Ricardo Baeza-Yates, Sung Hyon Myaeng, and Kalervo Järvelin, eds.), ACM Press, New York, US, 2002, pp. 97–104.

[17] Soumen Chakrabarti, *Mining the web: Discovering knowledge from hypertext data*, 1st ed., Morgan Kaufmann, October 2002.

[18] Huang. Chen, Sherrylynne S. Fuller, and Charles P. Friedman, *Medical informatics: Knowledge management and data mining in biomedicine (integrated series in information systems)*, Springer, June 2005.

[19] Philipp Cimiano and Steffen Staab, *Learning by googling*, SIGKDD Explor. Newsl. **6** (2004), no. 2, 24–33.

[20] Philipp Cimiano and Johanna Völker, *Text2onto - a framework for ontology learning and datadriven change discovery*, 2nd European Semantic Web Conference, 2005.

[21] N. Collier, A. Kawazoe, Y. Tateisi, L. Jin, M. Shigematsu, D. Dien, R. Barrero, K. Takeuchi, and A. Kawtrakul, *A multilingual ontology for infectious disease surveil-*

lance: rationale, design and challenges, Language Resources and Evaluation **40** (2006), no. 3, 405–413.

[22] Aron Culotta, Michael Wick, Robert Hall, and Andrew Mccallum, *First-order probabilistic models for coreference resolution*, In Proceedings of HLT-NAACL 2007, 2007.

[23] Laurie Damianos, Jay Ponte, Steve Wohlever, Florence Reeder, David DAY, George Wilson, and Lynette Hirschman, *MiTAP, text and audio processing for bio-security: a case study*, Eighteenth national conference on Artificial intelligence (Menlo Park, CA, USA), American Association for Artificial Intelligence, 2002, pp. 807–814.

[24] Hasan Davulcu, Srinivas Vadrevu, Saravanakumar Nagarajan, and I. V. Ramakrishnan, *Ontominer: Bootstrapping and populating ontologies from domain-specific web sites*, IEEE Intelligent Systems **18** (2003), no. 5, 24–33.

[25] Son Doan, Ai Kawazoe, and Nigel Collier, *The role of roles in classifying annotated biomedical text*, BioNLP '07: Proceedings of the Workshop on BioNLP 2007 (Morristown, NJ, USA), Association for Computational Linguistics, 2007, pp. 17–24.

[26] Son Doan, QuocHung-Ngo, Ai Kawazoe, and Nigel Collier, *Global Health Monitor - a web-based system for detecting and mapping infectious diseases*, Proceedings of the International Conference on Natural Language Processing, 2008, pp. 951–956.

[27] Christiane Fellbaum, *Wordnet: An electronic lexical database*, Bradford Books, 1998, http://wordnet.princeton.edu/.

[28] Clark Freifeld, Kenneth D. Mandl, Ben Y. Reis, and John S. Brownstein, *Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports.*, J Am Med Inform Assoc (2007).

[29] Johannes Frnkranz, Tom Mitchell, and Ellen Riloff, *A case study in using linguistic*

phrases for text categorization on the www, In Working Notes of the AAAI/ICML Workshop on Learning for Text Categorization, AAAI Press, 1998, pp. 5–12.

[30] Zoubin Ghahramani and Katherine A. Heller, *Bayesian sets*, in Advances in Neural Information Processing Systems, 2005.

[31] Asunción Gómez-Pérez and David Manzano-Macho, *Deliverable 1.5: A survey of ontology learning methods and techniques*, Tech. report, IST Programme of the Commission of the European Communities, May 2003.

[32] Xianpei Han and Jun Zhao, *Named entity disambiguation by leveraging wikipedia semantic knowledge*, CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management (New York, NY, USA), ACM, 2009, pp. 215–224.

[33] William Hsu, Svitlana Volkova, Timothy Weninger, Jing Xia, and Wesam Elshamy, *Multimodal Information Exraction: The Predictive epidemiology Domain*, Tech. report, Kansas State University, 2009.

[34] Jing Jiang and Chengxiang Zhai, *An empirical study of tokenization strategies for biomedical information retrieval*, Inf. Retr. **10** (2007), no. 4-5, 341–363.

[35] Xing Jiang and Ah-Hwee Tan, *CRCTOL: A semantic-based domain ontology learning system*, J. Am. Soc. Inf. Sci. Technol. **61** (2010), no. 1, 150–168.

[36] Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann, *Assessment of disease named entity recognition on a corpus of annotated sentences.*, BMC bioinformatics **9 Suppl 3** (2008), no. Suppl 3.

[37] Ai Kawazoe, Hutchatai Chanlekha, Mika Shigematsu, and Nigel Collier, *Structuring an event ontology for disease outbreak detection*, BMC Bioinformatics **9 Suppl 3** (2008).

[38] Ai Kawazoe, Lihua Jin, Mika Shigematsu, Roberto Barrero, Kiyosu Taniguchi, and

Nigel Collier, *The development of a schema for the annotation of terms in the biocaster disease detecting/tracking system*, KR-MED, 2006.

[39] John Lafferty, Andrew McCallum, and Fernando Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, Proc. 18th International Conf. on Machine Learning (2001), 282–289.

[40] David D. Lewis, *An evaluation of phrasal and clustered representations on a text categorization task*, SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), ACM, 1992, pp. 37–50.

[41] Yue Lu, Hui Fang, and Chengxiang Zhai, *An empirical study of gene synonym query expansion in biomedical information retrieval*, Inf. Retr. **12** (2009), no. 1, 51–68.

[42] Alexander Maedche and Steffen Staab, *Mining ontologies from text*, EKAW '00: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management (London, UK), Springer-Verlag, 2000, pp. 189–202.

[43] Christopher Manning and Hinrich Schütze, *Foundations of statistical natural language processing*, pp. 575–608, The MIT Press, Cambridge, US, 1999.

[44] Andrew Mccallum, *Information extraction: Distilling structured data from unstructured text*, Queue **3** (2005), no. 9, 48–57.

[45] John Mccrae and Nigel Collier, *Synonym set extraction from the biomedical literature by lexical pattern discovery*, BMC Bioinformatics **9** (2008), 159+.

[46] Ryan Mcdonald and Fernando Pereira, *Identifying gene and protein mentions in text using conditional random fields*, BMC Bioinformatics **6** (2005), no. Suppl 1, S6+.

[47] Qiaozhu Mei, Dong Xin, Hong Cheng, Jiawei Han, and Chengxiang Zhai, *Semantic annotation of frequent patterns*, ACM Trans. Knowl. Discov. Data **1** (2007), no. 3, 11.

[48] N. M. M'ikanatha, D. D. Rohn, C. Robertson, C. G. Tan, J. H. Holmes, A. R. Kunselman, C. Polachek, and E. Lautenbach, *Use of the internet to enhance infectious disease surveillance and outbreak investigation*, Biosecurity and bioterrorism : biodefense strategy, practice, and science **4** (2006), no. 3, 293–300.

[49] Michele Missikoff, Roberto Navigli, and Paola Velardi, *The usable ontology: An environment for building and assessing a domain ontology*, In Proceedings of the International Semantic Web Conference (ISWC, Springer-Verlag, 2002, pp. 39–53.

[50] Tom M. Mitchell, *Machine learning*, McGraw-Hill Science/Engineering/Math, 1997.

[51] Nadeau, David, Sekine, and Satoshi, *A survey of named entity recognition and classification*, Linguisticae Investigationes **30** (2007), no. 1, 3–26.

[52] Ani Nenkova, *Pyramid Annotation Guide - DUC 2006*, `http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html_ducview`.

[53] Ani Nenkova, *Understanding the process of multi-document summarization: content selection, rewriting and evaluation*, Ph.D. thesis, Columbia University, New York, NY, USA, 2006, Adviser-Mckeown, Kathleen.

[54] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown, *The pyramid method: Incorporating human content selection variation in summarization evaluation*, ACM Trans. Speech Lang. Process. **4** (2007), no. 2.

[55] Amy Neustein, *Sequence package analysis: A new method for intelligent mining of patient dialog, blogs and help-line calls*, JOURNAL of Computers **2** (2007), no. 10.

[56] Vincent Ng and Claire Cardie, *Improving machine learning approaches to coreference resolution*, ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (Morristown, NJ, USA), Association for Computational Linguistics, 2002, pp. 104–111.

[57] Collier Nigel, Doan Son, Kawazoe Ai, Goodwin Reiko Matsuda, Conway Mike, Tateno Yoshio, Ngo Quoc-Hung, Dien Dinh, Kawtrakul Asanee, Takeuchi Koichi, Shigematsu Mika, and Taniguchi Kiyosu, *Biocaster: detecting public health rumors with a web-based text mining system*, Bioinformatics **24** (2008), no. 24, 2940–2941.

[58] John P.Woodall, *Official versus unofficial outbreak reporting through the internet*, International JOURNAL of Medical Informatics **47** (1997), 31–34.

[59] Imad Rahal and William Perrizo, *An optimized approach for knn text categorization using p-trees*, Proceedings of SAC-04, 19th ACM Symposium on Applied Computing (Nicosia, CY), 2004, pp. 613–617.

[60] Alexander E. Richman and Patrick Schone, *Mining wiki resources for multilingual named entity recognition*, Proceedings of ACL-08: HLT (Columbus, Ohio), Association for Computational Linguistics, June 2008, pp. 1–9.

[61] Nick Rizzolo and Dan Roth, *Learning Based Java for Rapid Development of NLP Systems*, Proceedings of the International Conference on Language Resources and Evaluation (LREC) (Valletta, Malta), May 2010.

[62] Barbara Rosario, *Extraction of semantic relations from bioscience text*, Ph.D. thesis, University of California at Berkeley, Berkeley, CA, USA, 2005, Adviser-Hearst, Marti.

[63] Barbara Rosario and Marti A. Hearst, *Classifying semantic relations in bioscience texts*, In Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics, 2004, pp. 431–438.

[64] Sohini Roy Chowdhury, Caterina Scoglio, and William Hsu, *Evolution and control strategies of the foot and mouth disease epidemic on a weighted contact network*, Proceedings of EPIDEMICS2, Elsevier, Second International Conference on Infectious Diseases Dynamics, 2009, p. Abstract: P2.07.

[65] Magnus Sahlgren and Rickard Cöster, *Using bag-of-concepts to improve the performance of support vector machines in text categorization*, COLING '04: Proceedings of the 20th international conference on Computational Linguistics (Morristown, NJ, USA), Association for Computational Linguistics, 2004, p. 487.

[66] Fabrizio Sebastiani, *Machine learning in automated text categorization*, ACM Comput. Surv. **34** (2002), no. 1, 1–47.

[67] Ralf Steinberger, Flavio Fuart, Erik Groot, Clive Best, Peter Etter, and Roman Yangarber, *Text mining from the web for medical intelligence*, Mining Massive Data Sets for Security (2008).

[68] Mark Thurmond, Andres Perez, Chunju Tseng, Hsinchun Chen, and Daniel Zeng, *Global foot-and-mouth disease surveillance using bioportal*, Intelligence and Security Informatics: Biosurveillance, Springer Berlin/Heidelberg, 2009, pp. 169–179.

[69] Yoshimasa Tsuruoka and Jun'ichi Tsujii, *Boosting precision and recall of dictionary-based protein name recognition*, Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine (Morristown, NJ, USA), Association for Computational Linguistics, 2003, pp. 41–48.

[70] Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti, *Bart: a modular toolkit for coreference resolution*, HLT '08: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (Morristown, NJ, USA), Association for Computational Linguistics, 2008, pp. 9–12.

[71] Svitlana Volkova, *Automated event extraction and named entity recognition in the domain of veterinary medicine*, 2010, Poster presentation at Grace Hopper for Women in Computing Celebration, ACM Student Research Competition (SRC), To appear.

[72] Svitlana Volkova, Doina Caragea, William Hsu, and Swathi Bujuru, *Animal disease event recognition and classification*, CEUR Workshop Proceedings of Web Science and Information Exchange in the Medical Workshop, WWW Conference, 2010.

[73] Svitlana Volkova, Doina Caragea, William Hsu, John Drouhard, and Landon Fowles, *Boosting biomedical entity extraction by using syntactic patterns for semantic relation discovery*, n Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Inteligent Agent Technology, Inpress, 2010.

[74] Svitlana Volkova and William Hsu, *Computational knowledge and information management in veterinary epidemiology*, In Proceedings of IEEE International Conference on Intelligence and Security Informatics, 2010.

[75] Svitlana Volkova, William Hsu, and Doina Caragea, *Named entity recognition and tagging in the domain of epizootics*, 2009, Poster presentation at Women in Machine Learning Workshop (WiML'09).

[76] Richard C. Wang and William W. Cohen, *Language-independent set expansion of named entities using the web*, Data Mining, IEEE International Conference on (2007), 342–350.

[77] Tim Weninger and William H. Hsu, *Text extraction from the web via text-to-tag ratio*, DEXA Workshops, IEEE Computer Society, September 2008, pp. 23–28.

[78] Mary E. Wilson, *Travel and the Emergence of Infectious Diseases*, Journal of Agromedicine, vol. 3, 1996, pp. 51–66.

[79] Hong woo Chun, Yoshimasa Tsuruoka, Jin dong Kim, Rie Shiba, Naoki Nagata, and Teruyoshi Hishiki, *Extraction of gene-disease relations from medline using domain dictionaries and machine learning*, Proc. PSB 2006, 2006, pp. 4–15.

[80] Mark Woolhouse, *Foot-and-mouth disease in the uk: What should we do next time?*, JOURNAL of Applied Microbiology **94** (2003), no. s1, 126–130.

[81] Yiming Yang and Xin Liu, *A re-examination of text categorization methods*, SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), ACM, 1999, pp. 42–49.

[82] Zhihao Yang, Hongfei Lin, and Yanpeng Li, *Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature*, Computational Biology and Chemistry **32** (2008), no. 4, 287 – 291.

[83] Ki-Joong Lee Young-Sook, Ki joong Lee, Young sook Hwang, and Hae chang Rim, *Two-phase biomedical ne recognition based on svms*, Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, 2003, pp. 33–40.

[84] Yulei Zhang, Yan Dang, Yi-Da Chen, Hsinchun Chen, Mark Thurmond, Chwan-Chuen King, Daniel Dajun Zeng, and Catherine A. Larson, *Bioportal infectious disease informatics research: disease surveillance and situational awareness*, dg.o '08: Proceedings of the 2008 International Conference on Digital Government Research, Digital Government Society of North America, 2008, pp. 393–394.

[85] GuoDong Zhou and Jian Su, *Named entity recognition using an hmm-based chunk tagger*, ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (Morristown, NJ, USA), Association for Computational Linguistics, 2002, pp. 473–480.

Appendix A

Event Indicative N-grams: Complete Lists by Classes

In Tables A.1 and A.2 we present the complete list of indicative n-grams (nouns/verbs, noun/verb phrases) for susceptible, infected, recovered classes and suspected, confirmed classes respectively.

Table A.1: The complete lists of verb and noun n-grams for susceptible, infected and recovered classes

Status	Corresponding Verb/Noun N-gram Patterns
Susceptible	susceptible, be taken in, fall for, give in, impression, inclined, influence, liable, movable, non-resisted, open, predisposed, prone, ready, receipted, response, sensitized, sensitive, subject, vulnerable, head counted, agriculture censured, density, suspect, populated, healthy, at risk, exposed;
Infected	contaminated, disease, sick, infect, diseased ridden, plague ridden, plagued, affect, influenced, morbid, report, diagnosed, infection, emit, virus produced;
Recovered	recovered, removed, disposed, euthanasia administered, cull, ip cul, infect premiseculed, dc cull, danger direct contact cul, dead, clean, death, mortal, cp cull contiguous premiseculed, discard, eliminated, excised, withdrawn, isolated, separated, retrieved, regain, recuperated, heal, cure, slaughter, evaded, electrocuted, burn, buried, incinerated, destroyed, end.

Table A.2: The complete lists of verb and noun n-grams for confirmed and suspected classes

List Name	Status	Corresponding Verb/Noun N-gram Patterns
Initial	Confirmed	confirm, infect, strike, outbreak, tested positive, detected, diagnosed, diseased;
	Suspected	spread, catch, threat, danger, risk of infection, warned, subject, suspect;
GoogleSets	Confirmed	confirm, open, close, select, search, review, buy, alert, prompt, reserve, set time out, quote, clear timeout, fetch, write, set interval, prepare, delete, print, describe, execute, scroll to, scroll by, move to, add, clear, save, back, infect, the exchange of, strike, ball, strike looking, fouled off the pitch, outbreak, tested positive, tested negative, those affected, at risk, detected, request, reset, not enabled, not detected, diagnosed, facilitated, represented, assessed, clarified, collected, advocated, assisted, guided, supported, demonstrated, referred, familiarized, educated, provided, arranged, ensured, diseased, remedy, truth, given, ill, dead, morbid;
	Suspected	spread, catch, try, finally, throw, threat, risk, hazard, danger, warning, caution, risk, hazard, peril, note, jeopardy, para, flammable, threat, poison, endanger, corrosive, attention, notice, endangerment, chance, menace, imperil, tip, hint, error, important, section, hazardous, safety, imperilment, jeopardize, threaten, combustible, signs, explosive, gamble, pitfall, fatal, emergency, death, caustic, toxic, harmful, chapter, warned, mobile, phone, blocked, suspect, risk of infection, susceptibility to infection;
WordNet	Confirmed	stroked, affected, affirmed, beared out, buried, burned, cleaned, contaminated, corroborated, cp cull, contigu premis, cull, culled, dc cull, danger, direct contact, cull, dead, death, destroyed, detected, diagnosed, discarded, diseased, diseas-ridden, disposed, electrocuted, eliminated, emitted, ended, eraded, euthanasia administrated, excised, healed, infected, infested, influenced, ip cull, infect premis cull, isolated, killed, morbided, mortality, outbreak, plagu, plagu-ridden, reasserted, recovered, recuperated, regained, removed, reported, retrieved, separated, sick, slaughtered, substantiated, support, sustain, test posit for, product, withdrawn;
	Suspected	spread, catch, threated, danger, risk of infection, warned, predicted, alerted, scared, re-emerged, agriculture census, at risk, be taken in, believed, believed like, density, expected, exposed, fall for, give in, guess, head count, health, herd counted, imagined, impression, inclined, influenced, liable, movable, nonresistant, opened, populated, predisposed, prone, reading, receipted, responded, sensiled, sensited, subjected, supposed, surmised, suspected.