AUTOMATIC TEXT FILTERING USING LIMITED SUPERVISION LEARNING FOR EPIDEMIC INTELLIGENCE

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades

Doktorin der Naturwissenschaften

Dr. rer. nat.

genehmigte Dissertation von

M. Sc. Avaré Bonaparte Stewart

geboren am 13. Marz 1966, in New York, USA

Hannover, Deutschland, 2013

Referent: Prof. Dr. techn. Wolfgang Nejdl
Ko-Referent: Prof. Dr. Matthew Smith

ZUSAMMENFASSUNG

Für viele Anwendungen hat das Web die Art, wie wir Information sammeln, prozessieren und nutzen, grundlegend verändert. Eine besondere Rolle spielt diese Datenerfassung für "Epidemic Intelligence (EI)". EI benötigt die automatische Filterung von Nachrichten und Meldungen im Zusammenhang mit Erkrankungen, dies wird ermöglicht über Verfahren des "Supervised Learnings". Allerdings müssen den "Supervised Learners" ausreichend markiertes Trainingsmaterial zur Verfügung stehen um gute Resultate zu erzielen. Demgegenüber stehen der hohe Zeitbedarf und die Kosten für die Bereitstellung und Pflege eines geeigneten grovolumigen Datensatzes. Dies ist auch bekannt als das "Labeling Bottleneck Problem".

Diese Arbeit beschäftigt sich mit der Lösung des Labeling Bottleneck Problems für die Anwendunge in EI durch "Limited Supervision" (eingeschränkte berwachung?), d.h. durch die Benutzung alternative Methoden zur Meldungsfilterung um den Bearbeiter zu entlasten. Hierzu wurden die folgenden drei Verfahren entwickelt: a) "semi-supervised" Lernen mit "schwachen" Annotationen und corpora-übergreifendem Generationsprozess b) aktives Lernen mit Label-Auflösung c) überwachungsfreies Lernen mit Erkrankungsmeldung Clustern. Für jede dieser Herangehensweisen wurde zusätzlich die Effektivität aus der Sicht eines Domänenexperten gemessen, was bei vielen der gegenwärtig verfügbaren Systemen vernachlässigt wurde.

Erstens, beim Semi-Supervised Lernen wird die Frage nach der Ermittlung von hochwertigen Startpunkten für die Initialisierung eines Klassifizieres behandelt. Das dazu entwickelte "xLabel"-Verfahren benötigt 1) nur einen kurzen Text und 2) nur eine kleine Zahl von "weak labels" zur vollständigen Initialisierung. Diese "weak labels" wurden dabei automatisch aus zuverlässigen und leicht verfügbaren zusätzlichen Corpora generieren.

Als zweites wurde "Active Learning (AL)" verwendet, eine Methode zum Erzeugen eines lernfähigen Classifiers, um die Kosten und den Aufwand der manuellen Annotation der Trainingsdaten zu reduzieren. Die Qualität von Clusteringverfahren wie sie bei AL verwendet werden leidet wenn die Annotationen des Classifiers nicht abgeglichen sind mit den Active Learner Clustern ("label-cluster alignment problem"), oder der Lerner nicht den gegenseitige Ausschlu zwischen relevanten und irrelevanten Zielkonzepten behandeln kann. Für diese Probleme zeigt diese Arbeit mögliche Lösungen auf, die Zuordnung der "wahren" Labels für ungeklärte Instanzen durch einen semi kontrollierten Clusteringalgorithmus basierend auf "Partially Labeled Dirichlet Allocation (PLDA)". PLDA erlaubt nicht nur das Clustering und die Labels miteinander anzugleichen, sondern verfügt zusätzlich einen Inferenzmechanismus für die

Label wodurch eine Vielzahl der Labels automatisch aufgelöst werden können, ohne menschliches Zutun. Darüberhinaus kann durch das Ausnutzen der zugrundeliegenden Topicmodells des PLDA können überlappende Kontexte in den Seeds eliminiert werden und der Klassifier neu angelernt werden auf einen Seedset mit gröerer Dichotomie (Gegensätzlichkeit).

Letztendlich, "unsupervised Learning" mit Clustern kann auch als die Lösung für das "Labeling Bottleneck Problem", bezogen auf das Filtern der Texte, dienen. Eines der Hauptprobleme mit dem ungesteuerten Clustern ist das die Erkennung von Erkrankungsmeldungen durch generative Modelle zu sehr komplexen Lösungen führen. Durch die Anzahl der potentiellen Kluster (oder latenten Topics) stellt diese Komplexität eine signifikante Herausforderung für den Epidemiologen dar. Darüberhinaus mu jedesmal aufs neue die Signifikanz und Bedeutung eines entstehenden Musters bewertet werden, denn diese Muster sind nicht a-priori annotiert. Damit solche automatischen Methoden gute Resultate für den menschlichen Benutzer liefern, wurde eine benutzerbezogenes Verfahren gewählt welches sich auf zwei Punkte konzentriert: eine Bewertung der Cluster-Qualität und der Art ihrer Darstellung, beides der Schwerpunkt dieser Arbeit.

Letztendlich ist eine Schlufolgerung dieser Arbeit, da die Verwendung von Techniken mit "limited Supervision" (eingeschränkter Kontrolle) ein weiterer Schritt in Richtung besserer Unterstützung der Benutzer des World of Web Science ist, nicht nur für EI, sondern auch für andere Domänen.

SCHLAGWORTE

Semi-Supervised Learning, Active Learning, Unsupervised Learning,User Assessment

ABSTRACT

The Web has redefined the way we gather, process, and use information; and is capable of supporting a wide range of intelligence gathering tasks, in many domains. One such domain is Epidemic Intelligence (EI). EI requires techniques for automatically filtering disease reporting mentions, and is carried out by using supervised learning. One of the disadvantages of supervised learners is that they only do well, if given enough labeled training data. However, acquiring large volumes of data to build and maintain a classifier is an expensive and time-consuming process. This is known as the label bottleneck problem.

In this thesis, we tackle the label bottleneck problem for the domain of EI, using **limited supervision** approaches to learning - i.e, alternative ways of filtering disease reporting mentions that mitigate and/or avoid undue burden on an annotator. We develop three approaches that use limited supervision, namely: (1) semi-supervised learning with weak labeling and cross-corpora bootstrapping; 2) active learning with label resolution, and 3) unsupervised learning of disease reporting clusters. For each approach, we additionally measure its effectiveness from a domain expert's point of view, which is disproportionately, overlooked in state-of-the-art systems.

First, in Semi-supervised learning we tackle the question of obtaining quality seeds for bootstrapping a classifier. In our xLabel approach,we do so using semi-supervised classification that: 1) utilizes short text; and 2) is completely initialized with small amounts of ***weak labels*** that have been automatically acquired from highly reliable, and widely available, auxiliary corpora.

Second, Active learning (AL) is a methodology for building a trainable classifier that attempts to reduce the cost, or burden of manually labeling training data. Clustering approaches commonly used in AL suffers when: the classifier labels themselves are not aligned with the active learner clusters (label-cluster alignment problem); or when the learner is unable to handle the mutual exclusion between relevant and irrelevant target concepts. In our work, we tackle these problems, and facilitate the assessment of a true label for a dubious instance with a semisupervised clustering based on a Partially Labeled Dirichlet Allocation. PLDA not only allows us to align clusters with the labels, but also affords an inference mechanism with respect to the labels, so that we are able to automatically resolve many labels, without human intervention. Moreover, by exploiting the background topic model capabilities of a PLDA, we are also able to eliminate the overlapping context among the seeds in a principled way; and retrain a classifier with a more dichotomous seed set.

Finally, unsupervised learning, with clusters can also be considered as a

means of tackling the label bottleneck problem with respect to text filtering. One of the main problems with unsupervised clustering is that detecting disease reporting mentions using generative models can lead to very complex results. This complexity poses a significant challenge for an epidemic investigator, given the number of potential clusters (or latent topics). Additionally, since the pattern is not labeled apriori, the significance and meaning of the pattern must be interpreted. In order to ensure that the unsupervised methods produce results that are of value for the human users, we consider a user-centric approach which emphasizes both: an assessment of the cluster quality, and their representations.

Overall the implication for our work is that adopting **limited supervision** techniques, not only for EI, but also other domains as well, will help bring us another step closer to better supporting the information needs of users in the world of Web Science.

KEYWORDS

Semi-Supervised Learning, Active Learning, Unsupervised Learning, User Assessment

FOREWORD

For Abíróla

ACKNOWLEDGMENTS

This thesis has been a work requiring personal tenacity and commitment. It has also, unquestionably, required the support of many others; whom I would like to thank. First and foremost, I would like to thank Professor Wolfgang Nejdl, a visionary with indelible leadership. He has provided me with consistent support and a nurturing environment that allowed me to realize my first accepted funded proposal (upon which this work is based); and learn to conduct research. To him, I am immeasurable grateful. I also would like to thank my colleagues at L3S for their collaboration, particularly those within the M-eco Team. They have all been tangible examples, and from them I have learned a great deal. I also would like to thank the epidemiologists who warmly welcomed us into their domain, and provided valuable feedback for this work. Finally, I would like to thank Jens Muuss and Family for their encouragement each step of the way. Above all, I thank Idanel Bonaparte, who never allowed me to give up.

Contents

List of Figures

Introduction

1.1 Motivation: Epidemic Intelligence Scenario

Today's Web proves to be one of the largest repositories of available information for networked computer users. To better understand the potential of this phenomena, Web Science, has evolved as a multidisciplinary area of research, devoted to the socio-technical aspects of human and computer information processing. Individuals influence, and are influenced by information that is available on the Web. The Web has redefined the way we gather, process, and use information; and is capable of supporting a wide range of intelligence gathering tasks, in many domains. One such domain is Epidemic Intelligence (EI). In EI, a number of disciplines come together to help health officials monitor potential public health threats, by harnessing information about disease reporting mentions from unstructured textual on the web [PCKC06].

An overview of the EI pipeline used in the M-eco system is shown in Figure 1.1).[1] The data processing pipeline of the system is triggered by the arrival of different types of textual documents, such as: RSS news feeds, blogs, and microblogs from a `Media Stream`. We operate the system in *near-real time*, in which EI domain experts can expect to get current and relevant information at rate of four to six times a day. Importantly, not all documents are relevant to EI experts: thus, the first task in realizing an EI system, is in filtering out irrelevant textual mentions from the documents in the stream. Moreover, not all portions of a documents are of interest to the experts. Therefore, depending upon the type of text, we seek to detect the relevant portions of segmented documents (at the sentence level), and use these portions for: ***i)*** downstream analysis; and ***ii)*** presentation to the expert during there investigations. Taken together, our document collection consists of short (or sparse) text, ranging in length from a dozen words to a few sentences [PNH08].

Filtering of documents (or document segments) is accomplished by relying upon supervised classifiers [Zha08, NSC10a, vEHV+10a]. A supervised classifier learns to model the relationship between an observed variable (instance) and a target variable

[1] This EI system was development in the context of an European Union funded project, M-Eco http://www.meco-project.eu/, which was principally envisioned by this author.

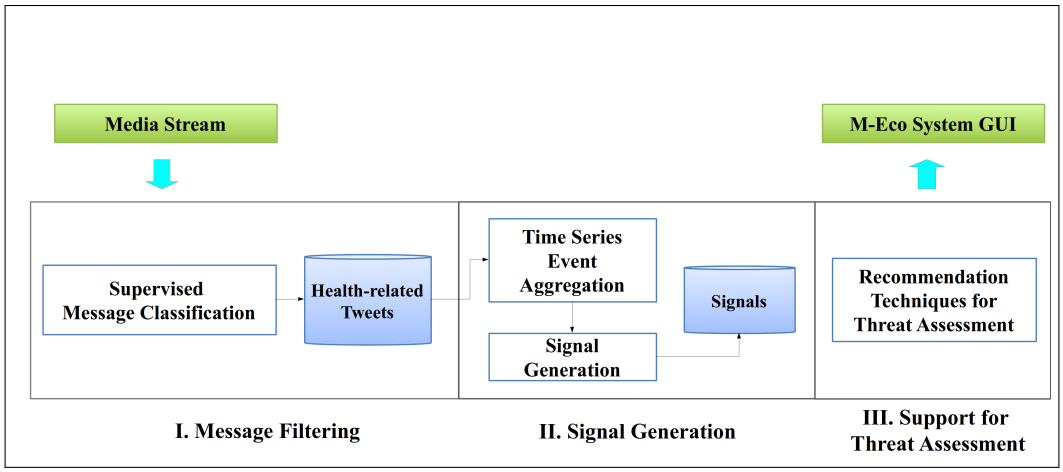


Figure 1.1 Overview of a M-eco Epidemic Intelligence System illustrating disease reporting messaging filtering for micro-blog text (or tweets).

(label or relevance judgment). The learned model is then used to perform inferencing, i.e., predict whether an unseen, future instance is a relevant disease reporting mention, or not.

We define a disease reporting mention with respect to the presence of selected entities types, and the roles they have within a specific scope, or segment, of a document. The predefined EI entity types of interest are: *Time* for temporal expressions; *Medical Condition* for infectious diseases, symptoms or their pathogens; *Location* for a city, state, or country; and *Victim* for an organism known to be affected by the medical condition. The EI system does not strive to detect all types of diseases, but only infectious, or communicable ones. For this purpose, a list of terms consisting of infectious diseases, their synonyms, pathogens and symptoms, which are provided by the domain experts, is used. All documents are annotated with these types of entities, if they are present within the document, and are used by the supervised classifier as features for representing the document's content.

The presence of EI entities types is a useful criteria for determining the relevance of a document for the task of EI. However, the presence of these entities alone is not be enough to help the classifier discriminate between relevant and non-relevant documents, thus depending upon the sparsity of the document, semantic feature types, which help to discriminate the role of the entities, are often used [Zha08, NSC10b, yZhL09, CCD09, CDKC09]. By eliminating those documents (document segments) that are unlikely to be relevant for the task, the supervised classifier, in essence, reduces the number (and portion) of documents that an investigator must examine in order to assess a public health threat.

However, even after message filtering, investigators are still typically inundated with the volume of text that they must examine in order to determine the extent to which the information constitutes a threat to public health. Thus, successive

stages of the proposed EI system (`Signal Generation` and `Support for Threat Assessment`) are intended to tackle the problem of information overload, and help users effectively digest the information and gather intelligence. During `Signal Generation:` outbreak warnings (or signals) are created from relevant short text (sentences or micro-blogs text) that has been previously filtering in the `Message Filtering` stage and aggregated according to counts of the common entity tuples they contain. Then, this time series data is used as input to biosurveillance algorithms for signal generation [KRSN12, SDA12]. A signal is a temporal anomaly generated from the counts of time series data that occur when an infectious disease or death is above an expected level, for a particular time and place. A signal consists of: ***i)*** an event surrogate, ***ii)*** a threshold value for which a temporal anomaly flag is raised if the time series count exceeds the threshold, for the given time window, and ***iii)*** a set of aggregated tweets which contributed towards temporal anomaly. At a minimum, a disease and temporal entity are required.

M-eco offers the functionality of *signal-based* retrieval, that is, returning signals as results of a given query instead of only documents. Once the desired signals are obtained, the user is able to access the original document associated to each of them. Having signals as a basic unit of information allows a user to perform a focused indexing of only the tweets relevant to a particular signal. Figure 1.2 shows the M-eco user interface along with a brief description of its main panels.

`Query Input`. The interface is designed to facilitate exploration, and allows users to find and analyze all signals generated by M-eco. It offers a free-text search field where arbitrary queries can be input. Such queries could represent medical conditions, locations, or any other relevant keywords of interest.

`Query Results`. The system also offers the functionality to sort within the signals loaded. If users wants to sort based on any of the columns of the result table, they can click on the name of the column and the system sorts the records in ascending (or descending) order. The user can access detailed information about the signals, as well as the corresponding documents, by clicking on the medical condition link in the Query Results list.

`Geo-located Signals`. Besides the result table, the system also displays a map with the locations of the signals loaded at the moment. If the user selects a marker on the map, then the system displays a box with information about the signal. The map visualization also offers controls to adjust the map type and zoom levels.

`Faceted Filter`: In order to help users manage the large amounts of data generated by the system, the search component incorporates *filters* to restrict the subset of results to a specific criteria. The following filters are supported: time range; medical condition, location pairs; medical condition only; location only; surveillance algorithm.

Promoted by their interest in a signal, experts can further explore the underlying text of a signal, that has been filtered by the `Message Filtering` stage, in order

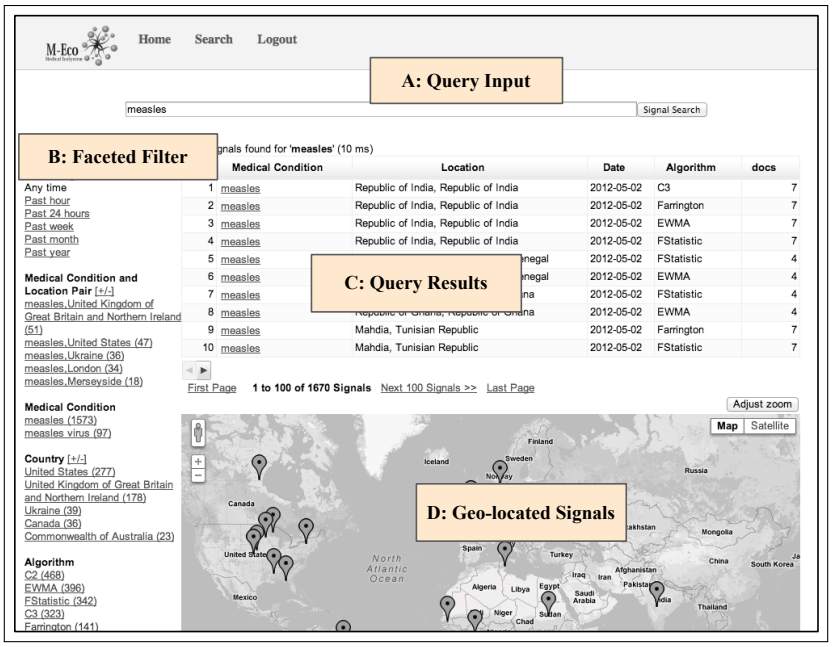


Figure 1.2 Filtered disease reporting mentions converted to time series data; and aggregated into views (signals) for browsing. *A. Query Input*, *B. Faceted Filter*: options for filtering signal search results by signal meta-data, *C. Query results*: resulting set of signals, and *D. Geo-located Signals*: a map for visualizing signals' geo-location.

to better understand the nature of a potential threat (Figure 1.3). The word cloud, along with its accompanying short text snippet, helps officials to get a quick overview of an incident, by summarizing its content. The `Signal Generation` and `Support for Threat Assessment` stages are beyond the scope of this thesis.

In the aforementioned scenario, the time series data used for anomaly detection, and the associated underlying documents, should be free of noise. That is, we would like to filter out documents for which the relevant entity types are present; but the role that these entities have, is not considered to be relevant in the context of a disease reporting mention. Automatic message filtering, and the use of limited amounts of labels to construct such a filter, is the focus of this thesis. In the remainder of this work, we restrict our discussion to Stage I. `Message Filtering`, and its relevance for the EI investigation task.

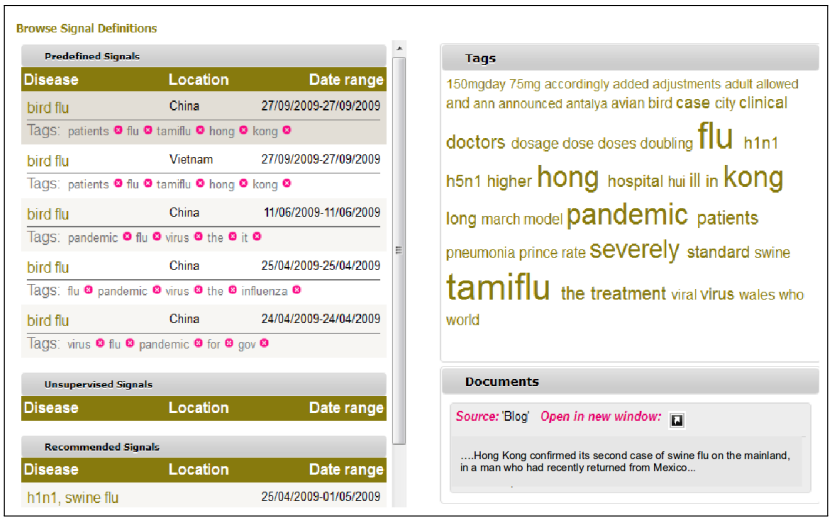


Figure 1.3 Zooming in on a selected signal shows summary views containing a word cloud and a short text snippet from a blog that have been obtained from a disease reporting message filter.

1.2 Label Bottleneck Problem of Supervised Learning

The task of automatic filtering using a supervised classifier is not limited to the domain of public health. In fact, it is important for any type of ongoing intelligence gathering from the Web, in general. One drawback of a supervised learning approach to text filtering is that it suffers from high initialization and maintenance costs, when it comes to building and maintaining a data set to train the classifier. This is the well-known ***label bottleneck problem***. A major reason for the label bottleneck problem is that supervised approaches operate under two major assumptions: ***i)*** high quality labeled, text is available for training a classifier and; ***ii)*** the source data on which the classifier is built, has the same feature space and distribution as the target data on which it is deployed.

For-hire, human intelligent task (HIT) workers within crowdsourcing platforms, such as Mechanical Turk [PD11a] can be employed to help acquire labeled data for EI. In fact, small amounts of labeled data using HIT workers is easily obtained. However, a different type of cost consideration must be taken into account for ongoing intelligence gathering, since large amounts of labels will be needed (perhaps even frequently) over a long duration - and monetary resources devoted to a crowdsourcing strategy are typically limited.

Finally, within the domain of EI the most prevalent approach to detecting ailment mentions from unstructured text is by using rule-based filtering [SFvdG+08a, YCB+99, SFvdG+08b]. A rule is a conditional of the form: *contextual pattern* $\rightarrow$

action. If the contextual pattern matches the appropriate parts of an input text, then the action part of the rule fires. A contextual pattern is intended to describe the context in which entities (disease, location, etc.) appear. Similar to the annotation problem of supervised learners, rule-based approaches face the challenge of also building (and maintaining) the pattern base for capturing the nuances within linguistic expressions, which can be infinite, even for a single task, such as detecting disease reporting mentions.

1.3 Contributions: Limited Supervision Learning for EI

The volume and types of Web data necessitate techniques for automatically filtering, such as supervised learners. However, *all existing EI systems that rely upon supervised learning assume that large volumes of labeled text are available to aid in constructing classifier models*. Unfortunately, this is far from the truth, in practice. Notably, approaches exist in other domains for tackling the label bottleneck problem, but these approaches have not yet made thier way into the domain of public health. From a socio-technical point of view, mechanisms must also be considered to help domain experts assess and judge the final quality of automated results.

In this thesis, we tackle the label bottleneck problem using **limited supervision approaches to learning - alternative ways of filtering disease mentions that mitigate and/or avoid undue burden on an annotator**. We seek to develop mechanisms that address the need to fully annotate training data for building a supervised learner within the domain of EI. We present three approaches that use limited supervision, namely: ***1)*** semi-supervised learning; and ***2)*** active learning, and ***3)*** unsupervised learning. For each approach, we additionally measure its effectiveness from a domain expert's point of view, which is disproportionately, overlooked in state-of-the-art systems. The contributions of this work are outlined below.

1.3.1 Semi-Supervised Learning with Weak Labels

Semi-supervised learning has been successfully used in many tasks to tackle the label bottleneck problem. Traditionally, a small set of high quality manually labeled seeds are assumed to be used for text level classification. In this work we address the classification task that: ***i)*** utilizes short text (a dozen words to a few sentences) [PNH08]; and ***ii)*** is completely initialized with small amounts of ***weak labels*** that have been automatically acquired from the short text of highly reliable, and widely available, auxiliary corpora (or EI knowledge bases).

1.3.2 Cross-Corpora Label Bootstrapping

In the absence of labels for a desired domain in EI, we show that the propagation of labels from an auxiliary domain is an effective way to overcome the label bottleneck problem. In this thesis, we apply a semi-supervised learner that has been constructed from an EI knowledge base, to the task of assigning a set of initial labels to vast amounts of completely unlabeled short text in a target domain. One of the main problems with semi-supervised learning is that they tend to suffer from a low recall (recall gated); and have a low accuracy. In our work we present solutions to tackle recall gating and over-fitting in our cross-corpora setting.

1.3.3 Semi-Supervised Active Learning with Label Resolution

Clearly there are cases for which even a well chosen EI auxiliary source is not suitable for handling the label bottleneck. One such example is when crossing the boundary between short text to the sparse text of Twitter micro-blogs. The corpora may no longer be compatible enough to support the propagation of labels (due to grammatically incorrect text; very limited context; lingo or metaphorical usage in Twitter). In such cases, we consider an active learning approach to handling the label bottleneck. The assumption of active learning is that if the learner is allowed to take part is selecting the more informative instances, it will ultimately lead to a learner that is supplied with as little training data as possible, for a desired optimal performance. Active learning comes at the expense of an oracle assessing the true label of dubious instances, so it is important that as few labels as possible are presented to the oracle. Clustering has been successfully used in many active learning strategies to help reduce the number of requests (queries) needed. However, approaches that are based on clustering can suffer when: ***i)*** no obvious clustering exists; ***ii)*** clusterings exist, but are at an unknown granularities; ***iii)*** the classifier labels themselves are not aligned with the active learner clusters (label-cluster alignment problem) [Das11]. In this thesis, we seek to address the label bottleneck problem with an active learner that is label-cluster aware. In doing so, we are able to mitigate the number of human annotations that are required for resolving an uncertain label for instances that stem from a non-separable context between the relevant and non-relevant training seeds of a binary classifier (the mutual exclusion problem). We address the mutual exclusion problem in semi-supervised active learning (SSL-AL) by exploiting Partially Labeled Latent Dirichlet Allocation (PLDA). As a type of semi-supervised clustering, PLDA is not only capable of constructing per-label clusters (label-aligned clustering); but is also capable of modeling an overlapping context among the training data (as a set of background clusters). Armed with such a model, we are able to eliminate the overlapping context among the seeds and retrain a classifier with a more dichotomous seed set. To the best of our knowledge, no previous cluster-based approach to SSL-AL has

employed the use of PLDAs for supporting label resolution in this way.

1.3.4 Unsupervised Learning of Disease Reporting Mentions

Unsupervised learning, specifically generative topic models, have also been extensively used as a means to understand overarching patterns in the data without relying upon labels at all. Notably, with the exception of one recent work by Paul et al., [PD11a] little work has otherwise been done in using unsupervised clustering to detect disease reporting mentions for EI. In addition to the fact that an oracle need not provide labels, another advantage of an unsupervised approach is that it has the potential of detecting public health related events that were not explicitly under surveillance.

1.3.5 Expert Interpretation and Assessment

We notice with the exception of a few systems [vEHV+10b, DKCC08], most supervised learning approaches do not employ the assessment of the domain experts to judge the final quality of the results - even fewer, for unsupervised systems [SS11]. Expert interpretation is especially crucial for clusterings since, their results may be difficult to interpret. In this work, we also report on the usefulness of disease reporting clusters that have been obtained from a generative topic model, from the perspective of domain experts. The goal is to offer much needed insights into how such an approach could be more beneficial and widely accepted as a viable technique for text filtering in EI.

1.4 Structure of This Work

This thesis is organized as follows: in Chapter 2 we first present an overview of limited supervision approaches that can serve as an alternative to supervised learning. We then proceed by providing the reader with a deeper insight into what constitutes an disease reporting mention within short text; and how can it be represented as set of features for a trainable classifier, in Chapter 3. We then present our approach to handling the label bottleneck using semi-supervised learning with weak labels acquired from EI-Knowledge Bases, to bootstrapping the short text of blogs and news, in a cross-copora setting.

In Chapter 4, once again we begin with a characterization of disease reporting mentions, but this time for sparse text, which is significantly different from the short text presented in Chapter 3. We then present our solution to handling the label bottleneck problem with Active Learning with Label Resolution. In Chapter 5, our unsupervised learning of disease reporting mentions for EI is presented for handling the label bottleneck problem. In each of the chapters describing our approach (Chapters 3, 4, and 5) we report on the usefulness of our results from the perspective of

domain experts, and provide an outlook for motivating the work in the chapter that follows. Finally in Chapter 6 we conclude by first summarizing the work done in this thesis; then provide several scenarios intended to show - in a more global context - how the various results presented in this thesis could be exploited to support an EI system. The thesis concludes by outlining directions for future work.

1.5 List of Supporting Publications

A number of papers investigating approaches to using limited supervision and filtering text to support information seeking were published by this author, and form the foundations for the work done in this thesis. A per-chapter listing of relevant publications is as follows:

In Chapter 3, we describe contributions included in:

- Avaré Stewart and Ernesto Diaz-Aviles. Epidemic intelligence: For the crowd, by the crowd. In *ICWE*, pages 504–505, 2012. [SDA12]
- Ernesto Diaz-Aviles and Avaré Stewart. Tracking twitter for epidemic intelligence: case study: Ehec/hus outbreak in germany, 2011. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, pages 82–85, New York, NY, USA, 2012. ACM. [DAS12]
- Avaré Stewart, Matthew Smith, and Wolfgang Nejdl. A transfer approach to detecting disease reporting events in blog social media. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 271–280, New York, NY, USA, 2011. ACM. [SSN11]
- Avaré Stewart and Kerstin Denecke. Can promed-mail bootstrap blogs? automatic labeling of victim-reporting sentences. In *Proc. of 1st International Workshop on Web Science and Information Exchange in the Medical Web, MedEx 2010, Raleigh, NC, USA, April 26, 2010*, 2010. [SD10a]
- Avaré Stewart and Wolfgang Nejdl. Self-supervised learning for medical web disease reporting events detection. In *Proc. of ACM WebSci'11, June 14-17 2011, Koblenz, Germany*, 2011. [SN11b]
- Kerstin Denecke, Peter Dolog, Pavel Smrz, Jens Linge, Wolfgang Nejdl, and Avaré Stewart. Using web data in the medical domain. In *Proc. of 1st International Workshop on Web Science and Information Exchange in the Medical Web, MedEx 2010, Raleigh, NC, USA, April 26, 2010*, 2010. [DDS+10]
- Avaré Stewart and Kerstin Denecke. Using promed mail and medworm blogs for cross-domain pattern analysis in epidemic intelligence. In *Proc. of 13th World*

Congress on Medical and Health Informatics Medinfo 2010, 12-15th September 2010, Cape Town, South Africa, 2010. [SD10b]

- Avaré Stewart, Kerstin Denecke, and Wolfgang Nejdl. Cross-corpus textual entailment for sublanguage analysis in epidemic intelligence. In *LREC*, 2010. [SDN10]

- Lda for on-the-fly auto tagging. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 309–312, New York, NY, USA, 2010. ACM. [DAGSN10]

- Avaré Stewart, Ernesto Diaz-Aviles, Wolfgang Nejdl, Leandro Balby Marinho, Alexandros Nanopoulos, and Lars Schmidt-Thieme. Cross-tagging for personalized open social networking. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09, pages 271–278, New York, NY, USA, 2009. ACM. [SDAN+09]

- Avaré Stewart, Ernesto Diaz-Aviles, and Wolfgang Nejdl. Mining user profiles to support structure and explanation in open social networking. *CoRR*, abs/0812.4461, 2008. [SDAN08]

In Chapter 4, we describe contributions included in:

- Mustafa Sofean, Kerstin Denecke, Avaré Stewart, and Matthew Smith. Medical case-driven classification of microblogs: characteristics and annotation. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, pages 513–522, New York, NY, USA, 2012. ACM. [SDSS12]

- Denecke K, Stewart A, Tim Eckmanns, Daniel Faensen, PeterDolog, Pavel Smrz. The Medical Ecosystem - Personalised Event-based Surveilllance. In *World Congress on Medical and Health Informatics*, Medinfo, 2010

- Nattiya Kanhabua, Sara Romano, and Avaré Stewart. Identifying relevant temporal expressions for real-world events. In and, editor, *SIGIR 2012 Workshop on Time-aware Information Access (TAIA'2012)*, TAIA'2012, 2012. [KRS12]

- Nattiya Kanhabua, Sara Romano, Avaré Stewart, and Wolfgang Nejdl. Supporting temporal analytics for health-related events in microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2686–2688, New York, NY, USA, 2012. ACM. [KRSN12]

- Avaré Stewart, Nattiya Kanhabua, Sara Romano, Ernesto Diaz-Aviles, and Wolfgang Nejdl. Leveraging social media for epidemic intelligence: Challenges and opportunies (under submission). In *ACM SIGIR Workshop on Health Search and Discovery: Helping Users and Advancing Medicine*. [SKR+]

In Chapter 5, we describe contributions included in:

- Marco Fisichella, Avaré Stewart, and Wolfgang Nejdl. Unified approach to retrospective event detection for event based epidemic intelligence (under review). *IEEE Trans. Knowl. Data Eng.* [FSN]
- A. Stewart and M. Smith. User centric public health event detection within social medical ecosystems. In *Proceedings of the 5th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2011)*, 2011. [SS11]
- Marco Fisichella, Avaré Stewart, Alfredo Cuzzocrea, and Kerstin Denecke. Detecting health events on the social web to enable epidemic intelligence. In *SPIRE*, pages 87–103, 2011. [FSCD11]
- Kerstin Denecke, Ernesto Diaz-Aviles, Peter Dolog, Tim Eckmanns, Marco Fisichella, Ricardo Gomez-Lage, Jens Linge, Pavel Smrz, and Avaré Stewart. The medical ecosystem [m-eco] project: Personalized event-based surveillance. In *Proc. of International Meeting on Emerging Diseases and Surveillance (IMED 2011), Vienna, Austria, February 4-7, 2011*, 2011. [DDAD+11]
- Marco Fisichella, Avaré Stewart, Kerstin Denecke, and Wolfgang Nejdl. Unsupervised public health event detection for epidemic intelligence. In *CIKM 2010: 19th ACM Conference on Information and Knowledge Management*, New York, NY, USA, 2010. ACM. [FSDN10]
- Avaré Stewart, Marco Fisichella, and Kerstin Denecke. Detecting public health indicators from the web for epidemic intelligence. In *3rd International ICST Conference on Electronic Healthcare for the 21st century (eHealth 2010)*. [SFD]

Background: Types of Limited Supervision Learning

In this section, we provide background: discussing the approaches to learning that use limited supervision (i.e., restricted amounts of manually labeled data) for building an automatic text filter; and highlighting areas in the research where our work is positioned.

2.1 Semi-Supervised Learning

Our task of detecting disease reporting mentions can be viewed as a binary classification problem. A binary classifier is defined as follows:

Definition 1 (Binary Classifier) *A binary classifier is a function* $F : \mathbb{R}^d \rightarrow \{-1, +1\}$ *that maps a d-dimensional feature vector* $x \in \mathbb{R}^d$ *to a label* $y \in \{-1, +1\}$.

The advantage of a supervised learner is that they do well at the harder task of predicting the true label for unseen, test data. The disadvantage, is that they do well if given enough labeled training data. For most learning tasks of interest, it is easy to obtain samples of unlabeled data; the World Wide Web, being a good example of a large collection of unlabeled data. In most cases, the only practical way to obtain labeled data is to have subject-matter experts manually annotate the data, an expensive and time-consuming process.

In this thesis, we seek to find a middle ground between building a good classifier, without relying heavily on the human annotation of a large quantity of training examples. In Chapter 3, we begin by using semi-supervised learning (SSL). A traditional semi-supervised (passive) learner in contrast to a supervised learner, does not require as much human effort since the system is bootstrapped with only a few manually labeled examples. In traditional, semi-supervised learning [Zhu05] a query component selects the most *reliable* instances at each iteration. More specifically, the semi-supervised learner takes as input, unlabeled data and a limited amount of labeled data; and, if successful, achieves a performance comparable to that of the supervised learner, but at a significantly reduced cost in the manual production of

training data. In our work with SSL, we focus on the following question: **Though we only require a small amount of labeled data for an SSL, from where do we obtain even these small labeled data?** Possibilities for obtaining SSL labels are:

1. manual labels
2. heuristic/intrinsic labeling scheme
3. a seed classifier
4. a budget for labeling instances that have been selected by the learner

In this thesis, we explore Options 1,2 and 3 in Chapter 3 and in Chapter 4 we explore Option 4.

2.1.1 Limited Supervision in Relation Extraction

The MUC and Automatic Context Extraction (ACE) initiatives played a significant role in driving the research challenges for addressing the label bottleneck problem with limited supervision systems, for the subtask of binary relation extraction. Relation extraction is defined as a mapping, $\varphi : \mathcal{D} \xrightarrow{R} \mathcal{F}$ of a set of documents, D, to a set of tuples $R \times E_1 \times E_2 \in \mathcal{F}$, where $e_i \in E_1$ and $e_j \in E_2$ are entities that form a tuple, denoted by $R(e_i, e_j)$, based on a semantic relation R. A binary semantic relation, $R(e_i, e_j)$, is a predication about a pair of (typed) entities. Early limited supervision systems under ACE were all built on the semi-supervised learning paradigm.

2.1.2 Distant Supervision

Distant Supervision (DS) [MBSJ09] is a more recent form of limited supervision, which attempts to acquire seed labels from an external source based on two assumptions. The first assumption is that if two entity pairs, within a *reliable* fact base, participate in a relation, then any proximate sentence (either same page or a hyperlink connection) that contains those two entity pairs might be an instance of the relevant relation and the relation can be extracted from the source text. Typical fact bases used in DS are Wikipedia info-boxes and YAGO [SKW07] or DBPedia [ABK+07]. The second assumption in DS is that of data volume and redundancy. Specifically, the same semantic relation will appear numerous times in a large volume of text, in different contexts. Although the entity pairs in the fact base are assumed to be a relevant (positive) instances of a known relation, the linguistic binding that describes the semantics, or context, for *how* the entities are related to one another, is unknown from the fact base, but can be discovered, presumably from proximate text. Given the amount of source text, and the number of relations in the fact base, it is infeasible

to search for all entity pairs within each sentence in the proximate text. Thus a common strategy in DS is to extract the context between entity pairs for a subset of the proximate text, and to then use these as a features for building a trainable classifier to detect more relevant instances from the source text. WOE [WHW09, WW10] for example, incorporates Wikipedia articles as training data to learn the extractor.

Similar to our work, this basic approach of semi-supervised learning is taken up in this thesis. We also focus on a predefined set of entity types that are relevant for the domain of EI. The main difference is that we take an implicit approach. In doing so, we relax the constraint of determining the predication, i.e., the type of relation that exists between entity pairs, as done in relation extraction. An implicit approach can be seen as a preprocessing step (identifying trigger sentences) for explicit forms of detection [NSC10b].

2.2 Unsupervised Learning for Clustering and Event Detection

Another counterpart to pure supervised learning is unsupervised learning with clustering. Unsupervised learning can be considered as a means of tackling the label bottleneck problem with respect to text filtering in that, it is also concerned with assigning instances to classes, but the clustering algorithm is only given instances and none of the labels for the classes. That is, in unsupervised learning, one seeks to find salient patterns in the data, which are above and beyond what would be considered pure unstructured noise [Gha04]. In particular, in Chapter 5 we will focus on generative models (mixture models), which have almost become synonymous with clustering.

Generative models have been widely used for the task of Retrospective Event Detection (RED). In RED, a document is assumed to contain the textual mention of one or more real-world, temporal events. A generative model is used to infer an event, where an event is considered to be a latent variable. Latent variables (as opposed to observable variables), are not directly observed, but are rather inferred by the model from some representation of the article's content that is observable, and directly measured (such as the distribution of its feature). When no new events are assumed to evolve over time, the problem can be cast as a classical document clustering problem [Gha04]. In Chapter 5, we tackle the limited supervision using a generative model for detecting disease reporting clusters.

2.3 Active Learning with Budgeted Labeling

Active learning (AL) is a methodology for building a trainable classifier, that attempts to reduce the cost, or burden of manually labeling training data. AL shares elements

of both supervised and unsupervised learning. Similar to a supervised learner, the goal of AL is to create an optimal classifier. Similar to unsupervised learning, the data come unlabeled. More precisely, the labels are hidden, and each of them can be revealed only at a cost. The key difference, however in active learning is to allow the learner to pro-actively select the "best" (informative) training instances, without having to label and supply the learner with more data than necessary. The label bottleneck is overcome by only asking the oracle for advice when the utility of doing so is high. The assumption of active learning is that if the learner is allowed to take part is selecting the more informative instances, it will ultimately lead to a learner that is supplied with as little training data as possible, for a desired optimal performance.

AL can be characterized by the manner in which oracles are queried. The popular pool-based learner [LG94] assumes a large data set with the majority of the data unlabeled. An item is chosen, by inspection, from the unlabeled pool. In an agreement method [LT97], a committee of learners is used to reduce the number of training examples required for learning queries; and selective sampling [Set09], where examples arrive successively and for each example, one has to decide independently whether it is informative or not. Independent of the query selection strategy employed, the central problem faced in all active learning is one of measuring the information content of the unlabeled data point.

Similar to previous works, we use a pool-based learner. We also take an approach to measuring the informativeness of a data point based on its distance from the separating hyperplane. This simple heuristic is a standard approach that has been shown to be efficient using a support vector machine for text classification [TK02]. Unlike previous works, however, we extend the traditional selection strategy with a ***semisupervised clustering algorithm*** that is not only capable of handling non-separable data in a principled way; but also allows us to reduce the number of data points that would be presented to an oracle when compared with traditional clustering approach.

Semi-Supervised Learning with Weak Labels

[1]

In this chapter, we use limited supervision to filter short text consisting of sentences. Similar to work done in Distant Supervision, as presented in Section 2.1.2, we tackle the label bottleneck problem for the task of detecting disease reporting mentions by using a *reliable* fact base. First, we use semi-supervised learning to weakly labeling the sentences within EI knowledge bases. Weak labels, as opposed to gold labels (those acquired from a human), are automatically obtained by exploiting properties of the knowledge base. In doing so, we acquire a large number of patterns for relevant and non-relevant instances of disease reporting mentions. Second, we apply these patterns to our desired copora to bootstrap the labeling of sentences therein.

[1] Image under License from Fotalia http://http://de.fotolia.com/

We refer to this approach as cross-corpora bootstrapping, or *xLabel*. Our approach to handling the label bottleneck using semi-supervised learning with weak labels that have been acquired from EI-Knowledge Bases, to bootstrapping the short text of blogs and news in a cross-copora setting is discussed in Section 3.4. However, before delving into the details of our *xLabel* approach, we first provide the reader with a deeper insight into what constitutes a disease reporting mention, in Section 3.1. We first present examples of relevant and non-relevant disease reporting mentions; secondly present guidelines for defining the relevance criteria; and finally show the set of features that we used for capturing these criteria to build a trainable classifier. In Section 3.2, we provide related work; in Section 3.3, terminology and a more formal statement of the *xLabel* problem is given. Experimental results evaluating the effectiveness of the *xLabel* approach is presented in Section 3.5. This chapter concludes in Section 3.6, summarizing the major results and providing an outlook for the future.

3.1 Short Text Characterizations of Disease Reporting Mentions

The textual mention of a real-world disease reporting mention is one which provides information about *Who-What-Where* with respect to a medical condition. It involves a persons suffering from an infectious disease; or its death related outcome that is above an expected level, for a particular time and place. It creates a need for action on the part of public health officials. For instance, an outbreak of cholera, or one case of a very rare, and highly contagious infectious disease, such as Ebola. Examples of relevant and irrelevant mentions are shown in Tables 3.1 and 3.2, respectively.

One can glean from these few examples how disease reporting mentions in short text can be characterized. Note also the importance of EI-specific entity types: Location, Disease, Temporal, and Victim, for the task. One of the main challenges however is that the presence of these EI-entities are a necessary, but not sufficient, criteria for detecting disease reporting mentions.

3.1.1 Relevance Criteria for Disease Reporting Mentions

We seek to establish a set of criteria for determining relevant and non-relevant. As a starting point, we examined the work that has been done in BioCaster [CKCC09]. BioCaster outlines a set of boolean and non-boolean criteria that can be used for annotating text, for a variety of events that potentially threaten public health, such as infectious disease outbreaks and chemical contamination. Their work is not intended to be exhaustive, and notably no criteria is explicitly given for when a disease reporting mention is not relevant. Moreover, there work was intended to be used for full documents, and not short text. *Most of the criteria described in the BioCaster guidelines are difficult to uniformly apply to short or sparse text given the limited*

Table 3.1 Examples of relevant disease reporting mentions in short text. Named entities offset with square brackets represent: ORG =victim of disease; DIS = disease; SYM= symptom; LOC=location; TEM= temporal mention.

Pattern and Example Short Text
Text: The Ministry of Health (MoH) of the [Kingdom of Cambodia]/LOC has announced a confirmed case of a [human]/ORG with the [avian influenza A (H5N1)]/DIS virus.
Text: About [142 passengers]/VIC were ill with [Norovirus]/DIS recently on an [Alaskan]/LOC cruise ship.
Text: While we are happy to have the negative tests for avian influenza in Bulgaria, confirmed outbreaks of [H5N1]/DIS in [Romania]/LOC and [Turkey]/LOC continue.
Text: The three [patients]/ORG tested positive for [Swine Flu]/DIS.
Text: This is the third case of [Ebola]/DIS observed within the past week.
Text: About 75 [H1N1]/DIS cases have been reported reported in [Salt Lake]/LOC.
Text: [China]/LOC confirmed its second case of [swine flu]/DIS on the mainland, in a [man]/ORG who had recently returned from [Mexico]/LOC.

Table 3.2 Examples of Non-Relevant disease reporting mentions in short text. Named entities offset with square brackets represent: ORG =victim of disease; DIS = disease; SYM= symptom; LOC=location; TEM= temporal mention.

Reason	**Example**
1. Off Topic	The first global conference on [SARS]/DIS will open tomorrow in [Kuala Lumpur, Malaysia]/LOC.
2. Outbreak Procedure	[Brussels]/LOC would take charge of future [foot and mouth]/DIS epidemics under a new [European]/LOC directive.
3. Vaccination Campaign	Of the health districts in [Burkina Faso]/LOC, 37 will benefit from a [yellow fever]/DIS preventive mass vaccination campaign on [13 Nov 2008]/TEM.
4. General Information	Challenges also exist in [China]/LOC and [Japan]/LOC, which together accounted for 82 percent of the region's population and more than 97 percent of its reported [measles cases] in [2008]/TEM.
5. Historically Outdated	The [Spanish Flu]/DIS of [1918]/TEM devastated the population/[VIC].

amounts of information contained within a single, short or sparse text message. Also, the boolean criteria is meant for a human assessment, so it is not straightforward to automatically extract the value of these boolean attributes from text, for the purpose of constructing features for a trainable classifier.

The boolean attributes of BioCaster are listed below.

- Was the victims of the disease involved in international travel potentially bringing the disease to new countries?
- Was the disease outbreak due to an accidental release?
- Was the disease reported to have crossed the species barrier between animals or from animals to humans?
- Was it reported that any victims of the disease failed to respond to regular drug treatment due to drug resistance?
- Did the victims of the disease catch the disease through contaminated food or water?
- Were any of the victims of the disease a hospital worker?
- Were any of the victims of the disease a farm worker?
- Did any of the victims of the disease catch the disease through contaminated or badly produced vaccines or blood products?

The non-boolean attributes of BioCaster are listed below.

- The country where the outbreak occurs
- The province in the country where the outbreak occurs
- The agent (pathogen) of the disease
- The species that was affected by the disease (either animal or human)
- The relative time when the outbreak occurred (hypothetical,present,recent past,historical)

We use a subset of the BioCaster criteria in our work. Specifically, we build upon the non-boolean attributes, which we extract via named entity detection. In the section that follows, we describe in more detail the named entities features, their extract, and the additional features we used to capture patterns of the type shown in Tables 3.1 and 3.2.

3.1.2 Features for Disease Reporting Mentions in Short Text

The meaning of relevance for our task is determined by the **context**, or the text surrounding the EI named entities. We use two types of features representations in *xLabel* to capture this context: non-structural and structural. Non-structural features ignore the relationship between tokens in the text, whereas structural feature take them into account.

Non-Structural Features

One of the most common method of representing a text document is in terms of a feature vector, that decomposes text it into its words; known as bag-of-words and has been found effective for text classification tasks. Bag-of-words ignore the order of tokens in the text, and the frequency of each token, along with its implicit co-occurrence with other tokens (i.e., context), is used as a feature. A weight can also be a boolean value for determining whether a given property holds within the text as in: "is the temporal mention within 3 months of today".

As illustrated by the examples in Tables 3.1 and 3.2, named entities play an important role in determining whether the short text is relevant. Thus, in addition to bag-of-words, we also rely upon bag-of-concepts; represented by the frequency of a set of predefined types of named entities that are deemed useful for the EI task. The EI-entity types we consider are: Temporal, Location, Medical Condition (symptoms, pathogens or diseases), and Organism. Each entity type is discussed below.

Temporal Entity: Extensive work has already been done in detecting temporal entities [SG12, KRSN12, CC10], and we exploit an existing approach namely, HeidelTime [SG12], for this entity type. The HeidelTime tagger is capable of resolving: preposition words (such as last Friday) or adjective and adverbs ("*5 months ago*"); absolute date (September 1, 1973); and a relative date, such "*yesterday*", which can only be determined from context. For example, given a date such as January 2, 2013, HeidelTime is capable of resolving the temporal mention "*yesterday*" to the date January 1, 2013. We found this adequate for our needs in filtering out mentions such as *Spanish Flu of 1918*, which took place over 90 years ago and is not considered a public health threat.

Location Entity: Extensive work has also already been done in detecting location entities [FGM05]. We experimented with Open Calais (http://www.opencalais.com/) and Stanford Named Entity Recognizer taggers (http://nlp.stanford.edu/software/CRF-NER.shtml) in our work. We found both taggers to be robust enough with respect to capturing: location granularity (city, state, providence); location mentions as adjective, (*Alaskan*); location disambiguations (Paris, Texas versus Paris Hilton); and locations used in a metonymy (e.g, *The Kingdom of Cambodia announced...*),

in which an inanimate object is used to express actions that would be taken by a sentient being.

Medical Condition Entity: For EI, in addition to the location and time entities we need: medical condition and affected organism. Although the extensive, domain specific annotator, Unified Medical Language System (UMLS: http://www.nlm.nih.gov/research/umls/), is capable of medical condition entity detection, we found that for our purposes it proved less effective. The main reason for this is that our domain experts were not interested in all the possible medical conditions tagged by UMLS, but only those related to infectious disease.

We were also interested in detecting aspects of a contagious medical condition, such as: symptoms, pathogens, virus as well as disease. In other exiting work, done by Dredze et al., [PD11a], steps are made towards determining aspects of a medical condition, yet their work is unsupervised and does not explicitly assign label to the aspects that are detected. The approach we take to medical condition entity detection is dictionary based. We used 723 English terms consisting of infectious diseases, their synonyms, pathogens and symptoms, which was manually built by our domain experts.

Organism Entity: As one can see from the examples in this section, the affected organism entity type is a fundamental characteristic in defining a relevant disease reporting mention. To the best of our knowledge, no other system has specifically dealt with an organism tagger, and we take up this issue in our work. Concretely, we defined an affected organism to be the semantic roles of a animals, including any concepts consisting of the following four types: (i) Persons-by-Population; (ii) Persons-by-Occupation; (iii) Persons-by-Geography; and (iv) Non-Human Organisms.

Persons-by-Population refers to the textual mention of a human by a family relation (e.g., brother, father), or a general population group to which a human belongs (e.g., elderly, group of children). Persons-by-Occupation refers to the textual mention of a human by their occupation (e.g., pilgrims, mine workers, nurse). Persons-by-Geography refers to the textual mention of a human by a geographical description (e.g., Moroccans, Brazilians)[2]. Non-Human Organisms refers to the textual mention of a non-human animal (e.g., swine, horse).

Each of aforementioned types of organisms entities were extracted with a simple dictionary based approach using LingPipe http://ir.exp.sis.pitt.edu/ne/lingpipe-2.4.0/. The complete list of terms used to construct each dictionary is provided in Appendix A and also available for download from the following web address: http://pharos.l3s.uni-hannover.de/~stewart/. One of the advantages of a Ling Pipe

[2]Cases for which person-by-geographical were tagged by both the Location and Organism NER tagger were safely ignored, since for the purposes of constructing features this overlap did not harm classifier performance

dictionary approach is its speed. LingPipe provides an implementation of the Aho-Corasick algorithm, which finds all matches against a dictionary in linear time, independent of the number of matches or size of the dictionary [AC75]. On the other hand, the limitations of a dictionary approach are that: morphological variations of entities must be explicitly enumerated (typically by using regular expressions), and entities names are not easily distinguished by their context if they are the same as common words. In spite of these limitations, we found the use entities as a feature consistently led to a improved performance for our classifiers, when compared with classifiers in which the NEs are ignored (e.g., see Figures 3.6 and 3.7).

Structural Features

In this section we discuss the structural features, in which the relationship between entities in the text are preserved. Common linguistic structures used for capturing word ordering within a sentence are syntactic parse [Zha08] and dependency parse [SG09, GS07] trees.

Syntactic Parse Syntactic parsing is the process of analyzing a sentence in order to determine its grammatical structure with respect to a given formal grammar (or set of rules). A syntactic parse tree represents the results of the parsing is a hierarchical tree structure in which each non-terminal node represents: either the grammatical part-of-speech (POS) or its grouping. The terminal nodes correspond to the tokens of the sentences and the edges in the tree denote the *is-a* relation between nodes. An example syntactic parse of is shown in Figure 3.1.

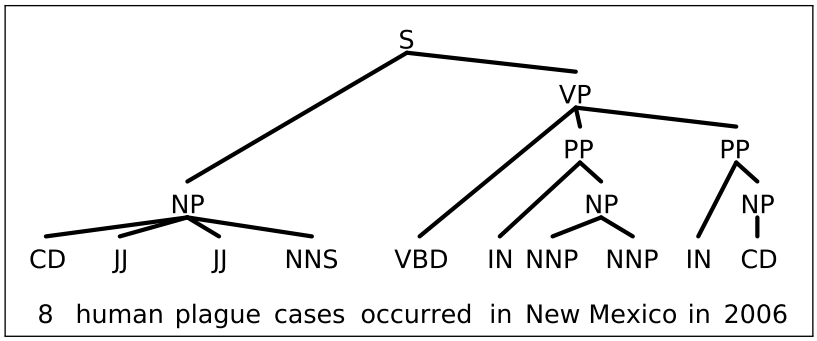


Figure 3.1 Example syntactic parse (POS) tree for the sentence: *8 human plague cases occurred in New Mexico in 2006.*

Dependency Parse Instead of capturing the relation between POS tags in a sentence as done by syntactic parse, dependency parsing capture the relationship between the words in a sentence; and is computable directly from its syntactic parse

[dMM08, NHN06, KMN09, SG09]. The dependency parse tree of a sentence, w, is a directed ordered tree, $T(w) = (V, E)$, with nodes, V and edges, E. Each node in the tree, v_i, represents a word, and each edge is labeled with a type as provided by the particular parser. The edges can either be typed or untyped. An example of a typed dependency parse tree is shown in Figure 3.2.

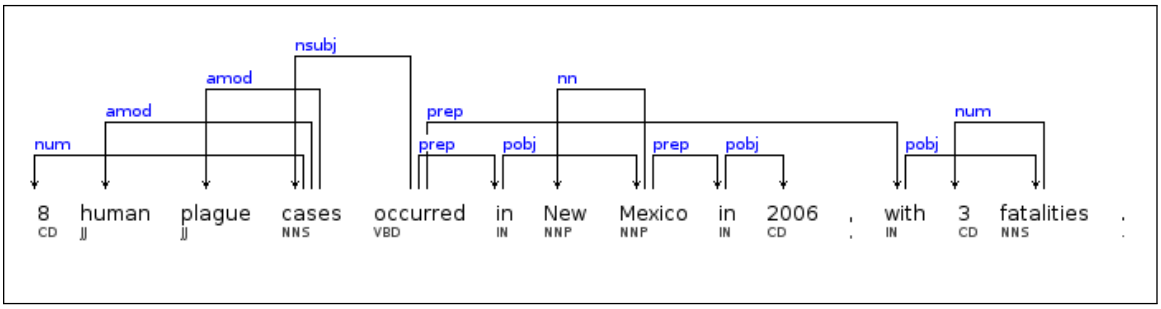


Figure 3.2 Example dependency parse tree for the sentence: *8 human plague cases occurred in New Mexico in 2006, with 3 fatalities.*

Substructures and Their Generalization Instead of using the entire tree, a portion of the structure may be used. Substructures allow a more narrowly defined context around the relevant entities, by pruning away, as much of the structure as possible, while preserving the relevant portions. This often leads to better features and classifier performance. Stevenson [SG09] and Zhang present an overview of several different types of structures. Generalization of structures is possible, by replacing the name of the entity with its type. In our experiments (Section 3.5), we evaluate the use of full structures, generalized substructures and the non-structural features that have been mentioned in this section.

3.2 Related Work

3.2.1 Distant Supervision

A number of systems exist that take steps towards addressing the label bottleneck problem for extracting general types of relations from large data collections. These systems are generally referred to as Open information extraction (OIE) [BCS+07, BE08, Ban00]. They are intended to address the label bottleneck problem; but also tackle the need for fast processing time and extraction of arbitrary tuples on a Web scale, where neither the relations (nor the participating entities) are known in advance [Bri99, AG00, ECD+04, BCS+07, BE08, Ban00, ZNL+09, EFC+11].

These system seed themselves with a set of heuristics [Ban00]; shallow NLP techniques [AA91]; or in the case of Know-ItAll [ECD+04] KNOWITALL, an extensible ontology is additionally used. An underlying assumption behind the use of labeling heuristics, is that most of the patterns for extracting the relevant relations can be grouped into a few categories. Automatic labeling heuristics are intended to capture the dependencies that would typically be obtained via syntactic parsing and semantic role labeling; without actually having to perform the syntactic parsing or semantic analysis. Differently from these works, we instead rely upon full sentential parsing.

TextRunner uses a set of heuristics based: on the length of the dependency parsing chain not being longer than a certain value; sentence structure (the path from $entity_i$ to $entity_j$ along the syntax parse tree, which does not cross a sentence boundary); and a parts of speech requirements (the entities of the relation do not consist solely of pronouns). Similarly, we as also rely upon heuristic to automatically label instances. However, one of the limitation of TextRunner is that by using light-weight techniques, the result contains relations that often have no meaningful interpretations. Unlike their work, we take steps towards interpreting of our results with the help of domain experts.

3.2.2 Transfer learning

Transfer learning is also similar to our work. The goal in transfer learning is to improve the learning of a target predictive function, using knowledge from a comparable, but different, domain. The work done in [ANC07, PY10] is closely related to ours, since no labeled data from the target domain is available a priori (i.e., transductive transfer learning). They also take into account additional domain independent properties of the training data (namely, the proportion of positive examples in the test data) to improve the classification performance. Numerous other works also exist where information gained from a learning task in one domain is transferred to improve the classification performance in another, related one [CJ09, ANC07]. For example, work has been carried out in a bootstrap setting [CJ09], where an event extraction system in one language is used to bootstrap another language. Similar to our approach, they rely largely upon unlabeled data and seek to apply a model across domains.

To date, none of these distant supervision based approaches to the label bottleneck problem consider the task of Epidemic Intelligence; use outbreak reports; nor syntactic parse based on Support Vector Machines kernel methods.

3.3 Terminology and Problem Statement

3.3.1 Terminology

In our approach, two distinct domains are considered, and the knowledge from a comparable, but different domain, is exploited to solve the same task in another domain. We constrain the auxiliary domain to contain outbreak reports and the target domain to contain blog and news. A **domain**, $\mathcal{D} = \{\Sigma, P(X)\}$, is a pair consisting of: a set of feature spaces, $\Sigma := \{(\sigma_1), \cdots, (\sigma_n)\}$, and a marginal probability distribution,P(X), where $X \in \Sigma$, is the set of surrogates or data instances. Each feature space, σ_i, determines how the raw data will be modeled. We categorize the feature spaces into two types: token based (i.e., bag-of-words, or bag-of-concepts) or structure-based (i.e., syntactic parse or dependency tree). As mentioned, structure-based feature spaces take into account the relationship between tokens. For example, if the goal is to classify sentences, and σ_i models a sentence as a binary, bag-of-words, then each sentence surrogate, $x_i \in X$ is a vector representing the presence or absence of a term in the sentence. Also, combinations of the features types may be used. Two domains are considered to be different, if they have different feature spaces, or different marginal probability distributions. We refer to the auxiliary and target domains as D_A and D_T, respectively.

A **task**, denoted by $\mathcal{T} = \{\mathcal{L}, \Psi\}$, is a pair consisting of a label space,$\mathcal{L}$, and a predictive function,$\Psi = P(L|X)$. The predictive function is not observed, but instead learned from a training set for a particular domain. For example, a training set for the auxiliary domain is represented as: $D_A = \{(x_{A1}, l_{A1}), \cdots (x_{An}, l_{An})\}$, where $x_{Ai} \in X_A$, is a training instance and $l_{Ai} \in \mathcal{L} = \{\textbf{true}, \textbf{false}\}$ is the set of all labels for a binary classification task. The function, Ψ, is used to predict a corresponding true or false label for a new instance that is in the target domain. Given the domains D_A and D_T, the learning tasks Θ_S and Θ_T are considered to be different, when either: ***i)*** the label spaces between the domains are different ($L_A \neq L_T$); or ***ii)*** the conditional probability distributions between the domains are different; i.e., $P_A(L|X) \neq P_T(L|X)$. In our problem setting, we consider the tasks to be the same.

3.3.2 Problem Statement

The problem faced in our setting is that, in general, neither the auxiliary outbreak reports, nor the new instances of the target blog data is labeled (L_A and L_T do not exist). Given a auxiliary domain, D_A; a learning task $\Theta_A = \Theta_T$; a target domain D_T; a feature space, Σ, that is common to both D_A and D_T; xLabel aims to: ***i)*** weakly label the data in the auxiliary domain; ***i)*** use a *subset* of the weakly labeled data as a training set, to learn a predictive function Ψ_A; and ***ii)*** use the knowledge from D_A and Θ_A, to improve the learning of a target predictive function,Ψ_T, where $D_A \neq D_T$, and $\Theta_S = \Theta_T$.

3.4 Cross-Corpora Bootstrapping of Disease Reporting Mentions

Laboriously annotating training data for building supervised learners is a commonly faced problem. But, what if enough reliable labels for our task could be easily gleaned from a comparable, auxiliary information source? What if this auxiliary is more amenable for acquiring labels due to: its nature, structure, topic matter, quantity, redundancy, reliability etc? Then perhaps, the knowledge of relevance and (not relevant) could be exploited to solve the same task in our unlabeled, target domain.

In this section we explore answers to the aforementioned questions and propose an approach `xLabel` to tackling the burden of labeling short text. An overview of our `xLabel` approach is depicted in Figure 3.3 and the algorithm is given in Algorithm 1. It consists of three phases: 1) *Auxiliary Domain* Semi-Supervised Learning; 2) *Cross-Corpora: bootstrapping*; and 3) *Target Domain:Semi-Supervised Learning.*

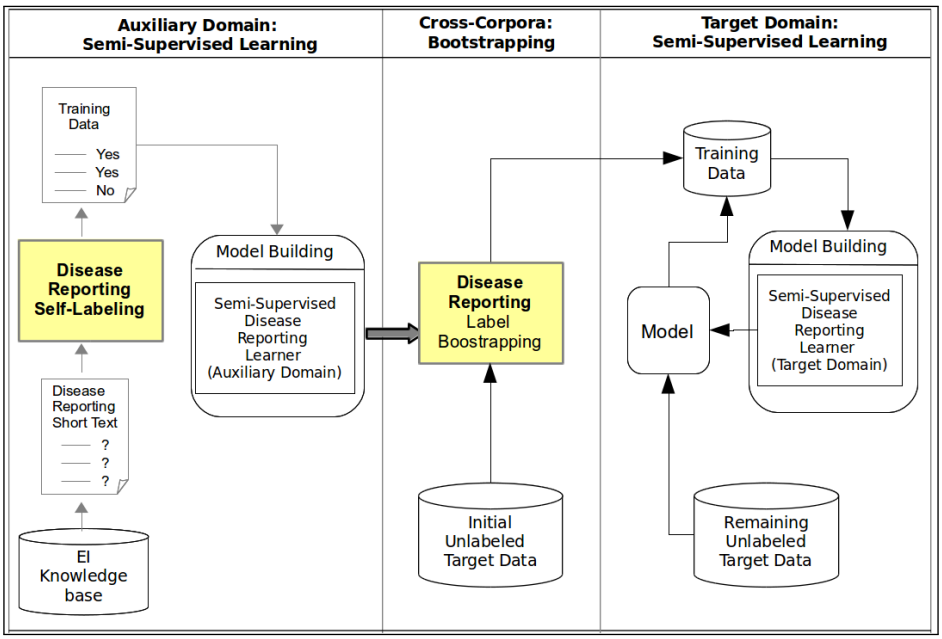


Figure 3.3 Overview of Limited Supervision Learning with xLabel: Cross-Corpora Bootstrapping. *xLabel* consists of three phases: 1) *Auxiliary Domain* Semi-Supervised Learning; 2) *Cross-Corpora: bootstrapping*; and 3) *Target Domain:Semi-Supervised Learning.*

Within the `Auxiliary Domain`, we rely upon a weak labeling of the outbreak report sentences(EI-Knowledge base), to build a classifier model. In the `Cross-Corpora Bootstrapping` phase, the model is used to label an initial set of short text from the desired target domain. Within the *Target Domain*, a target-specific model is constructed using the remaining, unlabeled target data. The underlying intuition behind

our approach is that the outbreak reports, acts as a type of "interlingua", which constrains the pattern a disease reporting text can have within another domain. In this section, the *xLabel* algorithm is presented and each phase is described, in turn, in the discussion that follows.

3.4.1 Auxiliary Domain Learning

The subtask of weakly labeling training data, first requires applying a score to each sentence. Various studies have been conducted for measuring the information bearing content of a sentence with respect to its document [LAJ01]. This includes: ***i)*** sentence position within the document; ***ii)*** the presence or absence of certain words or phrases in the sentence; ***iii)*** the title of an article. We incorporate these established results, by using several sentence weighting schemes.

3.4.2 Weighting Scheme

Sentence Position: For each document, in the corpus is represented as an ordered sequence of sentences. For the sentence position weighting scheme, we score the TopN sentences in a document with as '+1', to represent positive examples, for a threshold value of N. Further, we score sentences appearing towards the end of the sequence with '-1', so that the BottomN sentences are taken to be negative examples. All other sentences receive a score of '0'.

Sentence Semantics: We are interested in identifying disease reporting sentences. We say that a sentence is a disease-reporting one, if it contains a medical condition in conjunction with a case, time, or location, where the status of a case may be inferred from the context. The semantic information is thus represented by the presence of named entities (NEs) in the sentence. For the semantic weighting scheme, we modify the position weighting scheme. A value of '+1' is assigned to a TopN sentence if it contains the aforementioned entity types. A value and '-1' is assigned to the BottomN sentences if it does **not** contain these entity types. All other sentences receive a score of '0'.

Sentence Length: The sentences in the auxiliary corpus, vary greatly in length, due to conjunctions and phrases. Previous work using tree representations for sentences, has shown that longer sentences may contain too many irrelevant features, and over-fitting may occur, thereby decreasing the classification accuracy [CR03]. In this light, we propose that sentence length is also an important aspect for weakly labeling and investigate its impact in our experiments.

Algorithm 1: xLabel Algorithm: Semi-Supervised learning from Auxiliary domain and Cross-Corpora Bootstrapping for detecting disease reporting mentions in Target Domain

Input: Auxiliary Domain Sentences : $\mathcal{X}_{\mathcal{A}} := \{x_1, \cdots, x_m\}$
Auxiliary Text Sentence Weighting Scheme: Γ
$\mathcal{X}_{\mathcal{T}}$: target domain instances:
U_A: set of unlabeled instances for a auxiliary domain;
U_T: set of unlabeled instances for a target domain;
Φ: supervised learner
Output: $\Theta_T = \{(x_{T1}, l_{T1}), \cdots, (x_{Tn}, l_{Tn})\}$; target domain labeled instances
begin
1. Auxiliary domain weak labeling :
for *each* $\vec{a_i} \in \mathcal{A}_{\mathcal{S}}$ **do**
$\Theta_A := \Gamma(X_A) = (x_{Ai}, l_{Ai}, \cdots, x_{Am}, l_{Am})$ assign weak labels to instances based on weighting scheme,Γ
$P_A \subseteq \Theta_A$ initialize pool of weak labels for target
2. Auxiliary domain learning:
$\Psi_A = SSL(P_A, U_A, \Phi)$ (Algorithm 2)
3. Target bootstrapping :
$P_T \subseteq \Psi_A(U_T))$
4. Target learning:
$\Theta_T = SSL(P_T, U_T, \Phi)$ (Algorithm 2)
end

Algorithm 2: SSL: Semi-Supervised learning algorithm

Input: P: an initial seed of labeled instances; U: set of unlabeled instances;
Φ: a binary class learner
Output: $\Theta = \{(x_1, l_1), \cdots, (x_n, l_n)\}$: set labeled instances
begin
$L := P$
for *until a stopping condition is satisfied (e.g.,* $U = \emptyset$*, or threshold number of iterations is reached)* **do**
1. Train classifier: $\Psi := \Phi(L)$
$L_{cand} := \Psi(U)$, label examples
2. Select most confident instances:
$\tilde{L}_{pos} := SELECT(L_{candpos})$;
$\tilde{L}_{neg} := SELECT(L_{candneg})$; 3. Update: $L := L \cup \tilde{L}_{pos} \cup \tilde{L}_{neg}$
$U := U - L$
end

3.4.3 Cross-Corpora Bootstrapping

For the second phase, cross-corpora bootstrapping, once the optimal predictive function for the auxiliary domain has been derived, it is used to label text in the target domain. Kernel-based methods allow linguistic structures to maintains their (discreet) structural properties during classifier training. The alternative to kernel based methods, is non-structural bag-of-words features, in which the inherent structural properties of the text are not maintained as a feature during training. Kernel based approaches are popular method, when defining features for semantic tasks is not easily formulate-able. We point out that kernel based methods have not yet made their way into the domain of public health.

In our semi-supervised learner, we use a Support Vector Machine (SVM) with a tree kernel function as the base classifier. A kernel function is a similarity function satisfying two properties, that of being: symmetric and positive semidefinite [CST00]. The similarity function computes the inner product of two structured objects, such as a syntactic parse trees, first representing every sentence in the training data as a sentential parse tree. Then the feature space for the classifier is built by the (implicit) enumeration of all tree fragments in the training data [CD01, ZAR02]. One of the main statistical properties of the maximal margin solution is that its performance does not depend on the dimensionality of the space where the separation takes place. Thus, it is possible to map the data points into a very high dimensional spaces, such as those induced by using linguistic structures as features, without over-fitting.

When propagating labels from the auxiliary domain to the target the kernel function is used again to compute the proximity (or similarity) of an unseen example in the target domain, to the closest training examples of the auxiliary domain. The class of an unseen target example is determined by the side of the hyperplane on which it lies with respect to its proximate training example in the auxiliary domain.

3.4.4 Tree Kernels

The main goal of a tree kernel is to compute the number of common substructures between two trees T_1 and T_2 without explicitly considering the whole fragment space. Given the set of fragments $\{f_1, f_2, \cdots\} = F$, and an indicator function I_i(n) which is equal to 1 if the target f_i is rooted at node n, and 0 otherwise. A kernel, K, can then be defined as:

$K(T1, T2) = \sum_{n_1 \in N_{T_1}} \sum_{n_1 \in N_{T_2}} \Delta(n_1, n_2)$, where N_T1 and N_T2 are the sets of the T_1s and T_2s nodes, respectively and $\Delta(n_1, n_2) = \sum_{i=1}^{|F|} I_i(n_1) I(n_2)$. $\Delta(n_1, n_2)$ is equal to the number of common fragments rooted in the n_1 and n_2 nodes and is computed as follows:

1. if the productions at n_1 and n_2 are different then $\delta(n_1, n_2) = 0$;

2. if the productions at n_1 and n_2 are the same, and n_1 and n_2 have only leaf children (i.e. they are pre-terminals symbols) then: $\delta(n_1, n_2) = 0$;

3. if the productions at n_1 and n_2 are the same, and n_1 and n_2 are not pre-terminals then: $\Delta(n_1, n_2) = \prod_{j=1}^{nc(n_1)}(\sigma + \Delta(c^j_{n1}, c^j_{n2}))$,

where $\sigma \in \{0, 1\}$, $nc(n_1)$ is the number of the children of n_1 and c_j n is the j-th child of the node n. When $sigma = 0$, $\Delta(n_1, n_2)$ is equal 1 only if $\forall\Delta(c^j_{n1}, c^j_{n2})) = 1$ i.e. all the productions associated with the children are identical. By recursively applying this property, it follows that the subtrees in n_1 and n_2 are identical. The computational complexity of K is $O(\|N_{T1}\| \text{ x } \|N_{T2}\|)$. We refer the reader to the Moschitti [Mos04] for an efficient implementation that runs in linear time, on average.

3.5 Experiments

3.5.1 Experimental Goals

The objectives of our experiments are threefold. For the first objective, Auxiliary Domain Classification (Section 3.5.4), we begin by assessing the quality of a set of weakly labeled sentences by comparing them against human judgments. Weakly labeled seeds are those that have been labeled according to the heuristic properties of sentence length, position and semantics, as outlined in Section 3.4.1. Then we proceed by using these weak labels, we determine how well can we address the label bottleneck problem by automatically labeling the short text in ProMED-mail and WHO EI-knowledge bases using a **weak semi-supervised learner** (weak SSL). A weak learner is a semi-supervised learner that has been bootstrapped with weakly labeled seeds sentences. We compare the performance of a weak SSL against one that has been trained with manual labels.

Ultimately, we are interested in providing labels for the short text in the target domain of blogs and news according to our EI Scenario that was introduced in Chapter 1. Therefore, in the second objective, we measure how well we can address the label bottleneck problem within a target domain, using *xLabel*, our Cross-Corpora bootstrapping strategy using classifiers that has been trained with weak labels from the auxiliary domains of ProMED-mail and WHO. We consider two types of filtering strategies, which we refer to as: Precision Boosting (Section 3.5.5) and Recall Boosting (Section 3.5.6).

In the Precision Boosting Strategy, we use sentence length, semantics and position heuristics for weak labeling and ensure as large of a divergence as possible between the positive and negative examples by using the TopN sentences containing the named entities and the BottomN sentences *not* containing medical conditions named entities. In the Recall Boosting Strategy, we use only entity bearing sentences, regardless of position. The Recall Boosting Strategy is to a more challenging one for an EI

classifier, since the context surrounding the entities must be discriminating to assign an appropriate label. The purpose of these strategies is to examine how *xLabel* can be employed to build a classifier that simultaneously does not overfit the target domain (good accuracy) nor lead to recall gating (good recall).

In the third objective, Section 3.5.9, we the illicit feedback from domain experts with the intent of determining how we can adapt our xLabel strategy to come as close as possible to relevance judgments that would be given by an expert.

3.5.2 Data Sets and Summary

In realizing the aforementioned goals, two types of data sets were used: one for the auxiliary domain (outbreak reports) and the other for the target domain, consisting of a combination of blogs and news.

Target Domain: Blogs and News Blog data was collected by augmenting two different blog collections: MedWorm, [3] a medical blog aggregator; and AvianFluDiary[4] (Avian). For both MedWorm and AvianFluDiary, hypertext documents were collected for a one year period: January 1 - December 31, 2009. The new articles, we used was collected from the *url* column of the PULS online fact base [SFvdG+08b], a state-of-the-art event-based system for Epidemic Intelligence which provides public health event summarization and search capabilities. The hypertext documents were collected for a four month period, from September 1 - December 31, 2009, by crawling the website. The raw text was obtained by stripping all boilerplate and markup code using the method introduced by Kohlschütter et al. [KFN10].

Auxiliary Domain: Outbreak Reports For the auxiliary data, we used ProMED-mail [5] and WHO [6] outbreak reports. Both are global electronic reporting system, listing outbreaks of infectious diseases. These reports contain information about outbreaks and public health treats, which were moderated by medical professionals worldwide. The raw text documents were collected over a period of eight years: January 1, 2002 - December 31, 2009 from the outbreak report databases, freely available online.

Data Processing The raw text documents for both the source and target domains were processed using a series of natural language processing steps in order to extract the classification features outlined in Section 3.1.2. Each document was processed by applying the Standford Parser for sentence splitting; tokenization, and part-of-speech

[3] http://www.medworm.com/
[4] http://afludiary.blogspot.com/
[5] http://www.promedmail.org
[6] http://www.who.int/csr/don/en/

tagging. The Malt Parser [7] was applied to each sentence to construct both the parse tree and dependency tree, structural features (Section 3.1.2).

In addition, several named entity recognition tools were applied to each sentence to obtain the named entity features. Temporal entities were extracted using HeidelTime [SG12]. Location entities were extracted using the Stanford Named Entity Recognizer taggers [8]. Medical condition and organism entities were extracted using LingPipe [9], a simple dictionary based extractor supporting regular expression lookups. Table 3.3 shows a summary of the number of document collected and the corresponding number of sentences resulting after splitting and tagging. Table 3.4 shows the entity types and number of sentences tagged within the collection.

Table 3.3 The total number of document, sentences; and sentences containing named entities, for each type of data.

	No. Documents	***No. Sentences***	***No. Sentences(NER)***
ProMED-mail	14,665	347,822	84,423
WHO	1,541	16,213	3,469
Blogs	8,082	227,459	22,001
News	1,431	22,331	4,187

Table 3.4 shows a summary of the amount of sentences used for experimentation based on the named entity extracted.

Table 3.4 The number of sentences per data set for each of the three entity types medical condition (MED); location (LOC); affected organism (ORG); and temporal (TEM)

Entity Type	***MED***	***LOC***	***ORG***	***TEM***
ProMED-mail	195,077	134,227	47,142	26,458
WHO	9,239	8,969	4,462	2,928
Blogs	19,115	9,942	409	315
News	3,245	3,468	1,662	876
Total	402,904	53,675	53,675	30,577

To prune noisy and non-informative features, some sentences were eliminated from the experiments. Figure 3.4, shows a distribution over the sentence lengths (in characters) for the ProMED-mail and WHO. These distribution were used to determine the upper and lower bounds for the sentence lengths. Based on these distributions, sentences having a length below 12 and above 500 characters were excluded from the experiments.

[7] http://www.maltparser.org/
[8] http://nlp.stanford.edu/software/CRF-NER.shtml
[9] http://ir.exp.sis.pitt.edu/ne/lingpipe-2.4.0/

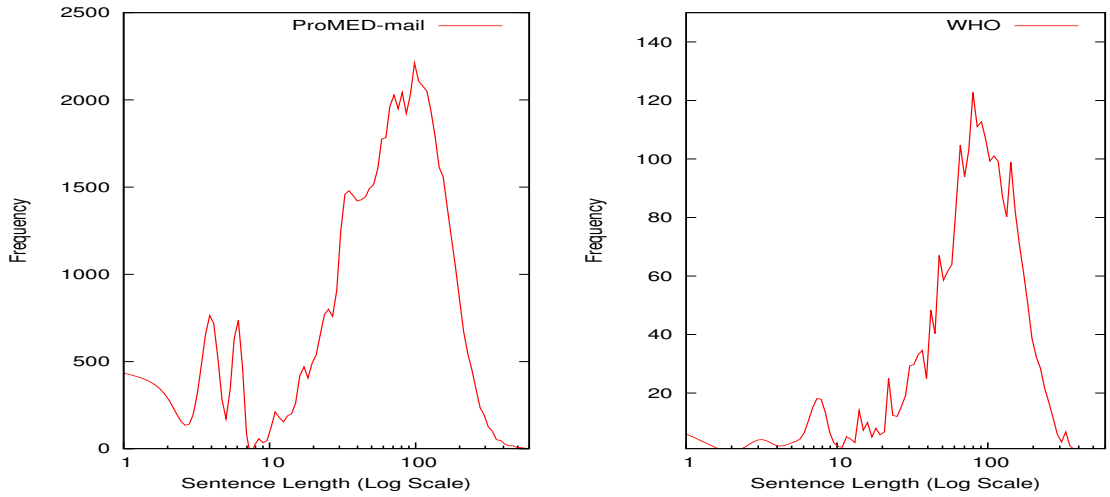


Figure 3.4 Average distribution of sentence lengths for ProMED-mail and WHO. Based on these distributions, sentences having a length below 12 and above 500 characters were excluded from the experiments.

3.5.3 Experimental Setting

Sentence-Level SVM Classifier and Features

The classifier used in our work was based on the implementation of SVM-TK by Moschitti [Mos06]. Experiments with various kernels and settings: linear, polynomial (degrees 2 through 5), RBF (gamma = 0.5, 2 .0 and 12.5) and sigmoid kernels, did not show a significance improvement over the default kernel settings, which were a polynomial of degree 3.

The set of features used as input to the classifier consisted of both structural features and non-structural features. We used Penn-Tree Bank, parts-of-speech parsing (POS), dependency tree (DEP); the term vector (VEC), and their combination (POS+DEP+VEC). Also used, were the presence/absence of negative terms (e.g., no, not, didn't, don't, isn't, hasn't); presence/absence of a modal terms (e.g., may, might, shall, should, must, will); and the number of a location (LOC), medicalCondition (MC); affected organism (ORG) entities.

For the purpose of these experiments we assume that the temporal filter is applied as a post-filtering stage, to exclude documents containing temporal mentions in a non-relevant time period. We opted for this approach under the assumption that sentence containing temporal mentions still contain relevant (non-relevant) patterns and we did not want to eliminate such instances when building our classifier. An example feature used in training the classifier is shown in Table 3.5.

As mentioned in Section 3.4.4, one of the main statistical properties of the maximal margin solution to classification is that its performance does not depend on the dimensionality of the space where the separation takes place. Thus, it is possible

Table 3.5 Example classifier feature for the non-relevant sentence: *Bird flu has killed at least 208 people worldwide out or 339 cases*; where $|BT|$ ($|ET|$) correspond to the begining (and ending) of a tree structured features and $|BV|$ ($|EV|$) correspond to the begining (ending) of the vector features, respectively.

0 $|BT|$ (S (NP (LEFT-M (NN bird) (NN flu))) (VP (VBZ has) (VP (VBN killed) (NP (NP (QP (IN at) (JJS least) (CD 208)) (NNS people)) (VP (VBN (MIDDLE-L worldwide)) (PRT (RP out)) (PP (IN of) (NP (NP (RIGHT-O (CD 339) (NNS cases)))))))))) $|ET|$
$|BT|$ (VBN (NN (NN (M *))) (RB (L *)) (CD (O *))) $|ET|$
$|BV|$ 7:1.0 13:1.0 156:1.0 240:1.0 939:1.0 1401:1.0 2255:1.0 10210:1.0 11027:1.0 20735:1.0
$|BV|$ 3:1.0 4:1.0 5:1.0 21:1.0 49:1.0 62:1.0
$|BV|$ 1:1.0 2:1.0 3:1.0 4:0.0 5:0.0 $|EV|$

to map the data points into a very high dimensional spaces, such as those induced by using linguistic structures as features, without over-fitting. Further, a tree kernel is capable of computing the number of the common substructures between two trees without explicitly considering the whole fragment space. This property is useful when the contexts defined by the text is very sparse and high dimensional. For example when using the tree structure as features, for both the dependency and parse tree, we have as many and 113,029 features for a training set consisting of 2,000 instances; and 465 support vectors.

Benchmark and Metrics Used

As a benchmark, we compared the performance of our xLabel approach with the: 1) Traditional (manual); 2) Random; and 3) state-of-the-art classification methods in EI. We used precision (P), recall (R), F_1 measure (F), and Accuracy (A), as metrics. The performance measures are computed as follows:

$$Precision = \frac{TP}{(TP + FP)} \tag{3.1}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3.2}$$

$$F_1 measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{3.3}$$

$$Accuracy = \frac{TP + TN}{(TN + TP + FP + FN)} \tag{3.4}$$

where TP = true positive; TN = true negative; FP= false positive; FN = false negative [MRS08]. The manual classification approach, where both the training and

testing is done using a hold-out on the manually labeled sentences for the target domain, represents a best-case scenario.

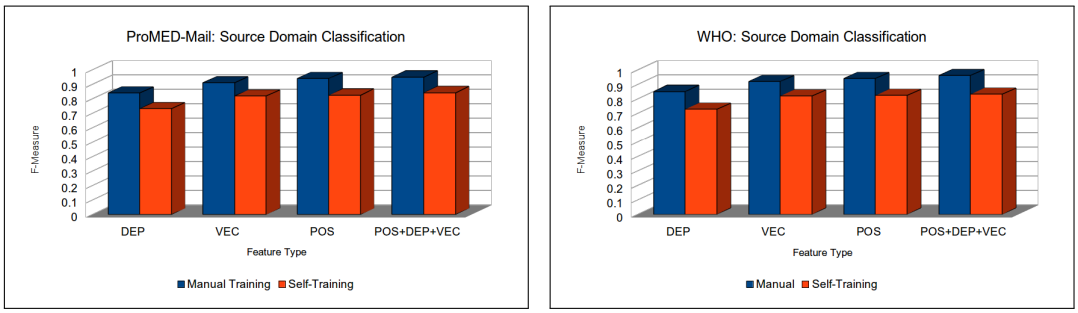
Gold Labels For the target domain 6,328 sentences were manually labeled, 835 were positive cases; clearly revealing the *needle-in-the-haystack* nature of the problem given the relatively low percentage of information-bearing sentences within the blog. The manual classifier was trained with equal amounts of positive and negative cases using a 10-fold cross-validation. Since the SVM classifier we used performs poorly on examples where the class imbalance is very high, all of the labeled data was not used since only about 10-12 percent of the sentences are positive examples. In contrast to the Traditional Classifier, the random classifier represents the worse case scenario in which the choice for the positive and negative training examples is taken at random from the auxiliary domains for weakly labeling sentences. For the auxiliary domain, one judge with guidance for domain expert labeled 2,000 positive and negative sentences in the both ProMED-mail and WHO. The labeled auxiliary sentences were used for the purpose of examining the performance of a manual classifier versus a semi-supervised one, before it was applied in a xLabel setting.

3.5.4 Results I: Auxiliary Domain Classification

We first verified the validity of weakly labeling, using three teams of judges that each labeled 100 sentences as disease reporting or not (positive or negative) from the topN and BottomN sentences of ProMED-mail dataset. We assessed the reliability of the human annotation using inter-annotator percent agreement. There was an overall inter-annotator agreement of 86%, showing that the assumption of weak labeling has promise in building a semi-supervised classifier. Next, we proceeded to compare a semi-supervised classifier with manually trained classifier i.e., one in which the labels were assigned by a human.

We trained several weak classifiers by varying the initial seed set, and averaged their performance. The results, as shown in Figure 3.5, show the superior performance of the manually trained classifiers over the semi-supervised one. The best performance for both ProMED-mail and WHO for the combined features of POS+DEP+VEC have an average F1-Measure of 92% for the manual classifier; whereas we acheive only 80% F1-Measure for the corresponding semi-supervised classifiers.

We notice that when using ProMED-mail there is 10% reduction in performance for the POS+DEP+VEC feature using semi-supervised learning, and for WHO the reduction in performance is larger by roughly 20%. We believe the different between the results for the auxiliary domains is attributed to the difference in their data sizes, and the manner in which disease reporting mentions are expressed in the collections. For WHO disease reports are made less frequently, hence the collection in general has fewer types of disease reporting mentions than ProMED-mail. Also overall WHO text contains shorter sentence lengths, as can be noticed by the distribution over their



(a) Promed-mail **(b)** WHO

Figure 3.5 Average F1-Measure for manual versus semi-supervised classifier on auxiliary domains of ProMED-mail (3.5a)and WHO (3.5b) using various feature types.

sentence lengths in Table 3.4. In addition, ProMED-mail tends to cross-reference WHO. An example positive and negative sentences using weak labels for ProMED-mail is:

1. Ex. Relevant: `Typhoid epidemic continues to spread in Kyrgizia.`
2. Ex. Irrelevant: `Fox predicts hunters will have a tougher time filling permits and tags this year because...`

See Appendix B for more examples of relevant and non-relevant sentences from ProMED-mail and WHO. Finally we note that the DEP feature alone under-performs all other features, this suggests that it might be a better feature to use in combination with the other structure features. As a whole, even though the results of the manually trained classifier outperforms the semi-supervised one, we propose that *if, very little or no manual annotation is possible for the task of building an auxiliary classifier from an EI knowledge base, then seeds obtained through weak labeling (or a mixture with manually obtained seeds) from the auxiliary domains of ProMED-mail and WHO is a viable option.*

3.5.5 Results II: Precision Boosting Strategy

In this section we consider a precision boosting strategy and in Section 3.5.6 we focus on a recall boosting one. Precision and recall boosting are important depending upon the task of the investigator. If for example, the investigator wants to detect a disease reporting mention as soon as possible with a high level of confidence, then a precision boosting detection strategy is used. On the hand, if the task of the investigator is to monitor the prevalence of an existing set of medical conditions, then it is better

to detect as many instances of a disease reporting mention as possible; so a recall boosting detection strategy would be put in place.

For the precision boosting strategy experiments, we used the classifier that was trained with weak labels from the auxiliary domains of ProMED-mail and WHO to seed labels for the unlabeled target data, which contains a mixture of blogs and news. The TopN entity bearing auxiliary sentences are taken as positive training examples; and the BottomN sentences that do not containing medical conditions are taken as non-relevant training examples. We manually labeled a set of sentences in the target domains and measured how well the auxiliary classifier was capable of predicting the true labels of the target with respect to Sentence Features; Sentence Position; Sentence Length; and Sentence Semantics.

Sentence Features

The results showing which feature yields the best results for the semi-supervised classifier is given in Tables 3.6, 3.7, and 3.8. We notice that in one-to-one comparison of the TopN sentences and training sizes, the POSVEC feature consistently outperforms the other features in terms of precision. This shows that using the combined features of POS and VEC, we are able to achieve a better performance than using each feature alone.

Sentence Position

A summary of the performance for our classifier showing how the position of the sentence within the document affects the weak labeling approach is shown in Table 3.6. We compared the semi-supervised labeling against a random selection of self-labels; and a Traditional Classifier, for which the labels were manually provided. We used the features POSVEC, varying the training size (Size) from 1,000 (1K) to 3,000 (3K) sentences; and using a fixed range for the sentence length consisting of 12 to 500 characters. The bold font in each table shows the maximum values obtained for each measure. The results clearly show that in terms of precision, we obtain results of 81.85% with the weakly labeling classifier, when compared with the Traditional Classifier, which has a precision of 86.50%. Also, in terms of a precision, the Top1 sentence positions prove to be the best. Noticeably, the random classifier performs significantly poorer (42.85% at best) than one which accounts for sentence position. These results clearly suggests that the sentence position is useful for weakly labeling. Interestingly, neither increasing the amount of training data, nor the TopN, yields a classifier with improved performance. Since, the precision performance of the Traditional Classifier is significantly better than a classifier built using weakly labels, this suggests that there is room for improvement, so we consider additional weakly labeling properties in Sections 3.5.5 and 3.5.5.

Table 3.6 xLabel Performance measures for various training sizes using weak labeling and features consisting of both a syntactic parse tree and a term frequency vector (POSVEC). The results for the Random and Traditional Classifiers are also shown. Size = training data size; N = topN sentences; P = precision; R=recall; F= F_1; A= accuracy.

Size	N	P	R	F	A
1K	1	**81.85**	60.63	69.66	73.59
	2	81.17	72.15	76.39	**77.71**
	3	78.71	70.51	74.38	75.72
	4	77.27	75.99	76.62	76.82
	5	76.51	**79.97**	**78.20**	77.71
	Random	41.37	40.11	40.63	41.61
2K	1	81.22	62.28	70.50	73.94
	2	80.76	70.23	75.12	76.75
	3	78.34	73.94	76.08	76.75
	4	76.89	78.05	77.46	77.3
	5	76.86	76.54	76.69	76.75
	Random	42.85	43.62	43.11	42.88
3K	1	79.32	60.49	68.64	72.36
	2	78.35	66.53	71.96	74.07
	3	76.98	74.76	75.85	76.2
	4	77.05	74.62	75.82	76.2
	5	75.2	76.54	75.86	75.65
	Random	33.37	32.66	32.94	33.40
Traditional		86.50	90.42	88.32	88.06

Sentence Length

The results showing how length of the sentences impacts the performance of the classifier that we build from weak labels is shown in Figure 3.6. We partitioned the set of sentences into four parts. The division points were computed from the quartile computation (i.e., each partition contained approximately 25% of the data), for the sentences having a length of 12...500 characters. In order to use the same sentence length as a division point across each TopN, we averaged the quartile values over each TopN, and used the average to create the partitions for the experiments, these partitions corresponded to sentences lengths (in characters) in the following intervals: Partition1=$[12\cdots69]$, Partition2 = $[70\cdots119]$, Partition3 = $[120\cdots171]$, and Partition4 = $[172\cdots500]$.

Based on the results shown in Figures 3.6, we see that when using shorter sentence lengths in the range of [12...69] characters, we achieve a higher overall precision values, when compared to using sentences in the other partitions (Figures 3.6b, 3.6c and

Table 3.7 xLabel Performance Measures for various training sizes using Automatic Labeling and features consisting a syntactic Parse tree from the Penn-Tree Bank parts-of-speech(POS). Size = training data size; N = topN sentences; P = precision; R=recall; F= F_1 measure; A= accuracy.

Size	N	P	R	F	A
1K	1	76.61	43.04	55.10	64.95
	2	76.72	64.25	69.86	72.33
	3	76.24	68.34	72.07	73.51
	4	75.68	73.50	74.54	74.94
	5	74.44	72.98	73.67	73.92
2K	1	***77.78***	46.45	58.16	66.60
	2	76.96	65.40	70.66	72.88
	3	76.57	71.00	73.67	74.62
	4	74.40	73.53	73.93	74.10
	5	75.08	73.83	74.44	74.66
3K	1	77.71	46.20	57.93	66.48
	2	76.14	65.62	70.46	72.51
	3	75.70	71.82	73.70	74.38
	4	74.64	73.52	74.05	74.25
	5	76.27	***74.07***	***75.14***	***75.49***

3.6d). These results are in fact consistent with the results reported in previous work using tree representations for sentences classifier[CR03]. In our case, the shorter sentences, contain the title (or headline), which is less noise and yield better classifier performance. We exploit, the manner in which actual outbreak information is written, when including the title of the outbreak report as the first sentence in the scoring scheme (Section 3.4.1).

Sentence Semantics

Figure 3.7 shows the results of our experiments when we combine the sentence length with semantics to evaluate the performance of our xLabel approach. Using the POSVEC feature and various training sizes, we notice that using a classifier which incorporates NEs yields significantly higher precision results, when compared with classifiers in which the NEs are ignored. A dense extractor (Lingpipe) extracted roughly five times more than that of sparse extractor (OpenCalais). Experimentally can see the extent to which such a difference in the annotation density impacts the performance of our classifier. For the dense (Figure 3.7a) and sparse (Figure 3.7c) classifier results, we achieve a rounded precision of 92% and 90%, respectively. In comparison, when NEs are ignored, we achieve a maximum precision of only 81.85%, as shown in Table 3.6, for the Top1 using a training size of 1K. We also notice in

Table 3.8 xLabel Performance Measures for various training sizes using Automatic Labeling and features consisting of a term frequency vector (VEC). Size = training data size; N = topN sentences; P = precision; R=recall; F= F_1 Measure; A= accuracy.

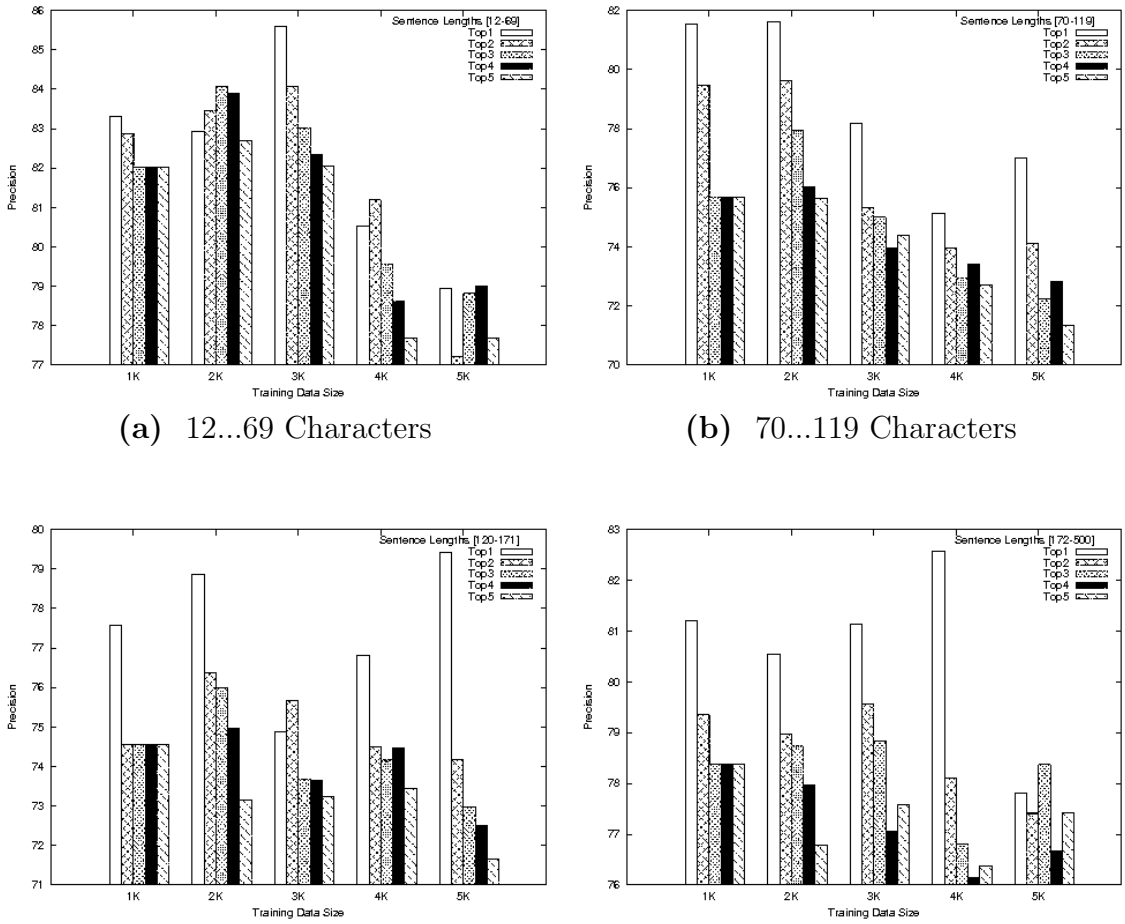
Size	N	P	R	F	A
1K	1	71.23	74.04	72.59	72.06
	2	70.68	78.63	74.42	72.95
	3	69.62	77.12	73.15	71.74
	4	68.75	79.81	73.85	71.73
	5	69.10	80.30	74.21	72.14
2K	1	71.76	74.46	73.05	72.54
	2	70.82	77.31	73.92	72.73
	3	70.87	80.71	75.45	***73.76***
	4	68.65	78.68	73.31	71.36
	5	68.06	80.55	73.76	71.35
3K	1	***72.91***	73.88	73.39	73.21
	2	71.08	76.43	73.64	72.64
	3	70.87	80.28	***75.27***	73.62
	4	69.63	79.97	74.43	72.54
	5	68.95	***81.65***	74.75	72.42

Figure 3.7, that the dense extractor achieves a higher overall precision than the sparse entity extraction. This is explained by the fact that the dense entity extractor has substantially more training data, even for the Top1 cases (15,756 examples of which 7,878 are positive and negative).

As we see from the results presented in the Figure 3.7a we are able to achieve precision as high as 92%. However, when we examine the recall for the classifiers that were built from the sparse and dense entity extractors using the POSVEC features, the results show the need for improvement. For the dense-classifier the maximum recall was 65.71%, which is just above recall values for the Top1 sentences, as shown in Table 3.6. In contrast the recall for the classifier built from the sparse entity extractor, had a maximum recall value of only 53.09%. In the section that follows, we consider what can be done to improve the the recall for our xLabel approach.

3.5.6 Results III: Recall Boosting Strategy

We now turn our attention to a recall boosting strategy determining - one in which we desired to detect as many disease relevant reporting mentions as possible. Differently from the precision boosting - where we used the entire structural feature, we instead use only a portion of the structure as outlined in Section 3.1.2. Substructures turn out to be an important type of linguistic feature, since they allow a more narrowly defined



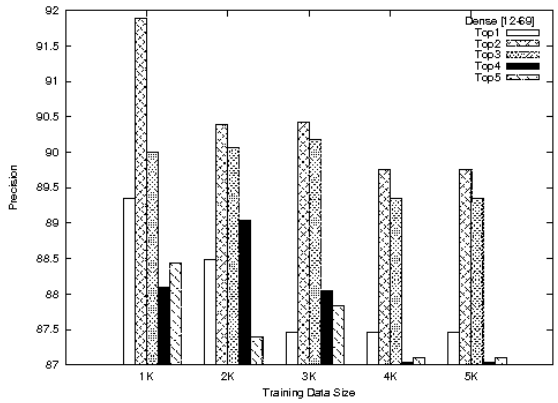
(a) 12...69 Characters **(b)** 70...119 Characters

(c) 120...171 Characters **(d)** 172...500 Characters

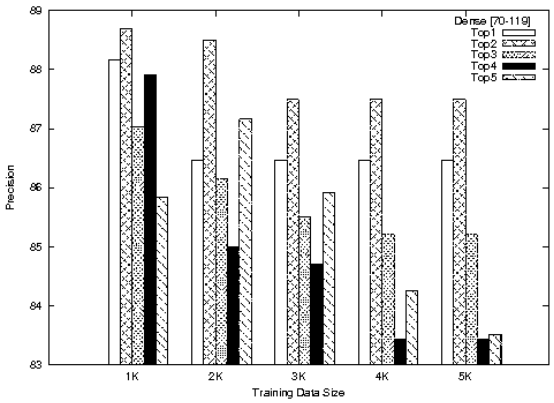
Figure 3.6 xLabel Precision based on a quartile partition of the sentence lengths into the intervals of: [12...69] characters(3.6a); [70...119] characters (3.6b); [120...171] characters (3.6c); and [172...500] characters (3.6d), for the POSVEC feature.

context around the relevant entities, by pruning away, as much of the structure as possible, while preserving the relevant portions. This leads to even less noise; and in our case a better recall performance, when compared with the recall obtained for the Precision Boosting Strategies. We use shortest tree path enclosing the first and last entity. In addition to substructure features, we also used generalization, in which the entity name is replaced with its type (as discussed in Section 3.1.2). Also, up to now, we only used sentences that did *not* contain medical conditions in the BottomN training set. In this phase of the experiments we use entity-bearing sentences only for assigning weak labels.

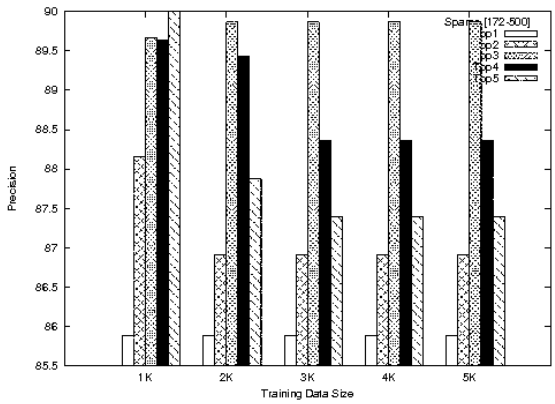
Table 3.9 shows the precision, recall and F1-Measure for the recall boosting strat-



(a) Dense:12...69 Characters



(b) Dense:70...119 Characters



(c) Sparse:172...500 Characters

Figure 3.7 xLabel Precision based on a partition of the sentence lengths for two dense entity extractors with sentence lengths: [12...69] characters (3.7a); and [70...119] characters (3.7b). The results using a sparse entity extractor with sentence lengths [172...500] characters is also shown (3.7c).

egy for different sentence lengths and positions. Overall, regardless of the sentence length and its position, the recall is significantly improved when compared with the recall for the best Precision Boosting strategy; whose maximum recall was 65.71%. In constrast, the best F1-Measure performance for the Recall Boosting strategy is obtained when using sentence lengths in the range of [12-500] characters; with a value of 89% in recall. Using a combinations of ProMED-mail and WHO showed no significant improvement, over using ProMED-mail alone: this is because ProMED-mail essential contains the text of WHO. For the same range of [12-500] characters, we also notice, that using Top2 sentences leads to a better recall boosting classifier, than using Top1 sentences.

These results suggest that the steps taken to generalize the auxiliary instances:

by ***i)*** replacing entities with their class; ***ii)*** using a broader range of sentences (all entity-bearing); and ***iii)*** incorporating the entire range of sentence lengths ([12-500] characters) - we are able to improve the auxiliary model's recall when classifying the target data - and avoid overfitting with respect to the target domain.

Table 3.9 xLabel Performance measures for a recall boosting strategy that uses substructures, structural generalization and entity bearing sentences exclusively for weak labeling with a training size of 2,000 instances. Aux. Domain= auxiliary training corpus; Length = sentence length in characters; P = precision; R=recall; F= F_1 Measure.

Aux. Domain	Length	P	R	F
ProMED:Top1	[12-500]	78.91	95.14	86.27
	[12-69]	71.78	97.32	82.62
	[70-119]	77.17	93.22	84.44
	[120-171]	77.88	90.96	83.91
	[172-500]	80.75	79.88	80.31
ProMED:Top2	[12-500]	81.25	98.29	88.96
	[12-69]	73.40	96.49	83.37
	[70-119]	77.52	92.88	84.51
	[120-171]	78.95	93.17	85.48
	[172-500]	80.96	89.43	84.98
WHO:Top1	[12-500]	84.92	88.22	86.54
	[12-69]	81.05	79.43	80.24
	[70-119]	84.77	92.31	88.38
	[120-171]	**85.71**	93.48	89.43
	[172-500]	84.55	87.87	86.18
ProMED+WHO:Top2	[12-500]	82.66	**98.74**	**89.99**
	[12-69]	73.45	95.76	83.14
	[70-119]	77.52	92.88	84.51
	[120-171]	81.24	90.37	85.57
	[172-500]	81.75	84.05	82.88

3.5.7 Discussion

The experiments presented above allow us to see that the properties of Sentence Position, Sentence Length and Sentence Semantics do, in fact, impact the ability of a classifier built from the weak labels to detect the relevant disease-reporting sentences and several points should be noted regarding of the properties.

Weak Labeling

Sentence Length: The results for sentence length have shown that using shorter sentence lengths in the range of [12..69] characters, yields the best performance for the Precision Boosting strategy. In constrast, for the Recall Boosting strategy, the best performance is obtained when a larger range of sentence lengths ([12-500]) was used. Given the fact that we used a syntactic parse tree as an underlying representation and a tree kernel method, we were able to generate a high number of syntactic features from syntactic fragments, from which the classifier could learn.

Sentence Position: It should be noted that the Sentence Position is not independent of Sentence Length. The shorter sentences that appear in the top first and second positions in the outbreak reports consist of titles, which suffice to summarize the reports; and offer a precision of up to 92% was obtainable. Using more TopN sentences in these cases, did not lead to an improved classifier performance. Moreover, in terms of the classifier learning rate, with as little as 1,000 examples, our classifier is able to reach this peak performance.

Sentence Semantics: Finally, we note that we achieve the best overall performance using a more dense named entity extraction tool that was built using a dictionary matching algorithm to extract entities. We find that the additional overhead required for the entity extraction, is worth the benefit, since we see that there is a substantial gain in term the classifier performance for both the Precision and Recall Boosting strategies.

Trade-offs of Tree Kernel

Feature Engineering: Tree Kernels allows linguistic structures to maintain their (discreet) structural properties during classifier training. Kernel based approaches are useful when defining features for semantic tasks is not easily formula-table. In practice, most kernel-based systems, have augmented non-structural features, to achieve a performance boost. Finally, we point out that kernel based methods have not yet made thier way into the domain of public health.

Kernel Computation: Kernels can be computationally expensive for large volumes of data and time sensitive tasks. For some complex kernels, such those in the state-of-art work by [RKP10, FRP09], using dependency parse tree kernel, it took up to twelve hours to train their best kernel. In time sensitive tasks, such as surveillance, this turn-around time can be prohibitive. In our case, we relied upon the Tree Kernel SVM of Moschitti [Mos06]; which operates with a complexity of O(m+n), on average.

Feature Construction: Linguistic structures (structural features) must be built for both the training set and the new, unseen examples on which the classifier is deployed. Depending upon the amount of data, a full sentential parse may also be prohibitive. An alternative is to use shallow parsing [ZAR02] such as: chunking to dividing sentences into noun or verb phrases [AA91]; or parts-of-speech tagging, which assigns a syntactic label such as NP to a chunk. Although we do obtain a performance boost with a composite kernel (structural and non-structural features), compared to vector-based features alone, when a time-critical deployment strategy is needed, it is better to use tokens features, if the full sentential parsing is too costly.

Parse Tree and Grammar: Parse trees semantics are not always clear. Parse trees may share similar structure, but have entirely different meaning. Conversely, parse trees may be structurally different, but actually mean the same thing. Finally, some long range similarities may be missed because no common subtree in the kernel computation covers them. Our goal was the assessment of structure features, but this was tied to the SVM implementation we used. MaxEnt classifiers have recently been reported to have a better performance than an SVM on certain tasks. It is an open question as to whether this is the case for short text used here for EI; we consider this in future work. Finally, we also note that we assumed our short text to be grammatically correct, hence, we were able to make a sentential parse of for the tree kernel. When is comes to grammatically incorrect text such as tweet our kernel approach is not as effective. and the pure vector is likely to be a more promising feature type.

EI Knowledge Bases: In order to overcome the expense associate with creating a sufficient size training set, distance supervision has been used. These systems overcome the burden of requiring hand labeled training data and has shown success with Wikipedia info-boxes and YAGO as auxiliary text. One common limitation, however, is in finding the so -called universal knowledge base that is suitable for the task.

3.5.8 Comparison with the State-of-the-Art

Short Text Classification

Finally, we compare the performance of the xLabel approach to reported results for the similar sentence-level classification tasks [Zha08, NSC10b]. Work has been done by Zhang [Zha08, yZhL09] for classifying disease reporting sentences. In their work, an F_1 measure value of 76% is reported. When considering sentence position alone, we obtain a comparable F_1 of 78.20% (see Table 3.6) instead, using weakly labeling. Work has also been done by Naughton et.al, [NSC10b]to detect events in sentences. The focus of their work was not on e-EI, but on the more generic Automatic Content

Extraction (ACE) event types. They achieved an F_1 measure of 90% for the event type *Die*, which contains incidences of death due to traffic accidents and natural disaster. These results were based on using 5,000 features, that were subsequently pruned using information gain theory. In contrast to this work, our F_1 measures are much lower for the precision boosting strategy and only slightly low for the recall boosting strategy.

One of the major drawbacks of short text classification at the sentence level, is the assumption that all the context is available within a single sentence (compared to a document). Patwardhan addresses the problem of limited token by *expanding the field of view* around key sentences to include their surrounding sentence [PR09]. Future work includes steps in this direction. Finally, we note a limitation with all semi-supervised strategies that use the most confidently predicted candidates to build the training set. A common problem is that if the classifier makes a confident, but incorrect prediction, the example is still fed into the next round to the the same classifier. Co-training is a common approach to handling such limitations [BM98] when the features are capable of being split into non-overlapping views. In our experiments with co-training we found it difficult to find appropriate non-overlap views, using the features presented here. Further work would be needed to devise such a set of dichotomous features for co-training.

Automatic Labeling: Work has been done in several areas to reduce the human labeling effort [TSZ07, FKG+09]; one such area is active learning [MS09, TSZ07, Set09]. Active learners are self-learning systems that: ***i)*** may construct their own (learning) examples; ***ii)*** request certain types of examples; or ***iii)*** determine a set of unsupervised examples that is most usefully to be labeled. The human effort in labeling a set of data is reduced, since the learner queries the human for labels of intelligently chosen examples. Much of the work done in this area has focused on techniques for selecting the unlabeled instances that are to be labeled. This differs from our approach since no human interaction is assumed in order to acquire labels.

Supervised Detection

Numerous supervised classifiers exist for detecting disease reporting events within unstructured text [CCD09, KBHT09, Zha08]. The work done in [DKCC08] incorporated the use of roles within epidemiological processes. Roles are central within ontological modeling because they encapsulate how entities are involved in processes and situations. In future work done by the same authors. [DKCC08] they additionally incorporated the use of three types of roles within epidemiological processes: a) case; b) transmission medium or vector (e.g., nonhuman, product or anatomy); and c) therapeutic agent (e.g., chemical substance). Roles are central within ontological modeling because they encapsulate how entities are involved in processes and situations. Using a combination of entity types and roles, they are able to acheive an F-Measure of 85.54% (Precision=83.74%; recall=87.43%) using the PERSON entity

and case role, with a Support Vector Machine (SVM).

In the work done by Conway,et.al [CCD09], semantic features include the use of hedges; *the means by which writers can present a proposition as an opinion, rather than a fact.* Some example hedges used in their work include: *reported*, *suspected*, *probable*, or *suspect*. A classifier built using hedges and unigram features achieved only a slight improvement compared with one that used unigram features alone. Interesting, using hedges they were able to associate a high, medium and low speculative category, to both relevant and non-relevant classified documents, thereby associating a form of confidence level to the classified documents. Collier et.al, [CDKC09], use WordNet style synonym sets in order to capture the distinctive semantic characteristics verbs and nouns within disease outbreak reports. KelBle09 [KBHT09] seeks to classify the documents from ProMED-mail outbreak reports using a Digramic Bayesian Classifier (DBACl) into event, nonevent and spam. Also at the sentence level, the work done by Zhang et.al, address the challenge of identifying the location.

In all cases, the authors incorporate the use of some type of semantics in order to capture relevant entity co-occurrences within a document. A limitation however is that they all also use manually labeled data to build their models. In our work, we seek to go beyond the human effort associated with building a training a supervised classifier by taking an limited supervision approach to detecting disease reporting mentions.

3.5.9 Results IV: Expert Interpretation and Assessment

We now turn our attention to a case study in which we seek to determine, which set of relevant sentences are best. The goals of our study was to: ***i)*** illicit feedback so we could determine how to adapt our xLabel strategy to come as close as possible to relevance judgments that would be given by an expert; ***ii)*** determine what percentages of the sentences that are labeled as relevant by the supervised classifier are considered relevant by domain experts? Ultimately we want to use the filtering sentences as input to detecting temporal anomaly (or signals); so we want to ensure that signals are being generated from sentences that the experts would consider to be relevant. If non-relevant sentences are being used to generate signals, this could lead to false positives and result in an overload for the domain experts.

We will refer to the sentences that have been labeled as relevant by the classifier as a "trigger sentence". We used the precision boosting strategy to build the target classifier and used the classifier to labeled a total of 2000 sentences from combined news and blogs. For the study we attempted to reconstruct a portion of the context surrounding the trigger sentence, in the form of an incident report,i.e., short snippets consisting of at most 3 sentences, including the trigger sentence and the two immediate sentences following the trigger sentence.

Experimental Setting

In an initial validation step (before user assessment) we constructed incident reports using only the top 100 most confidently classified sentences. As can be seen in Figure 3.8 the top 100 sentences all followed a similar structural pattern involving maladies in the form of an affected organism as a *case* of some disease.

Id	Report	Rating
101	**There is seems to be talking about the first recorded case ever in the world of human influenza infection among mink .** `` We do not know how the virus entered the farm , and we get it may never be solved . But it is probably a single factor , because there is a geographically limited area .	○ relevant ○ undecided ○ not relevant
102	**Four or five cases of human rabies in an area in the US would be an outbreak or even an epidemic.** Hundreds of colds or even serious pneumonias in an urban area is normal. It 's not an epidemic.	○ relevant ○ undecided ○ not relevant
103	**The WHO is reporting another modest increase in confirmed cases over yesterday , with 493 new infections .** Given that this is an update of Sunday 's numbers , and many labs were closed over the weekend , a small increase was to be expected . Additionally , many countries remain incapable of testing for the virus , while some other countries have the ability , but have either a limited capacity or limited desire to do testing .	○ relevant ○ undecided ○ not relevant
104	**The 140 confirmed cases in Canada are spread over eight provinces.** British Columbia has been hardest hit with 39, followed by Nova Scotia 38 , Ontario 31 and Alberta 24 . Québec 3 , New Brunswick 2 , Prince Edward Island 2 and Manitoba 1 also have reported cases.	○ relevant ○ undecided ○ not relevant
105	**Just as the 27,737 confirmed cases as reported in the WHO update # 46 represent some percentage of the number of actual H1N1 cases .** No one really knows the size of this elephant , since we are only privy to select pieces . It is unlikely that this virus is much more virulent than seasonal flu , else we 'd probably have seen some spikes in the 122 MRS 122 Cities Mortality Reporting System .	○ relevant ○ undecided ○ not relevant

Figure 3.8 Examples of the incident reports selected from the most confidently classified instances.

From this, we could clearly see the impact of the structural features in grouping case mentions as relevant instances. However, we wanted to avoid presenting triggers sentence with repetition to the user. Thus, we opted to build incident reports using a random subset of the relevant triggers only. Notably taking a random a subset of the relevant sentences also implies that the least confidently classified instances would also be included. Examples of the randomly selected incident reports are shown in Figure 3.9. We believe that a random selection from all the relevant classified to be more insightful with regard to the goal of determining how we could adjust the classifier performance to mimic the judgments that would be provided by and expert.

Data: Data for testing consisted of 100 trigger sentences (labeled relevant by the supervised classifier) partitioned into groups of 50 sentences and converted to incident reports. Experts were asked to label the incidents in one of three categories: relevant with respect surveillance at a national level; not relevant; or undecided. The assessment was done by nine domain experts; consisting of the Mekong Basin Disease Surveillance, World Health Organization; and European Centers for Disease Control

Id	Report	Rating
52	**Most of those sickened from the H1N1 virus have complained of mild, flu-like symptoms such as fever, cough, sore throat, aches and fatigue.** As of Sunday afternoon, health officials had reported five other deaths in the United States: three in Texas, one in Washington state and one in Arizona. People with underlying health issues seem most susceptible to the virus.	○ relevant ○ undecided ○ not relevant
53	**In the Northern Hemisphere, Japan continues to experience an early start to its annual flu season.** Countries in the equatorial and tropical regions of South America, such as Ecuador, Venezuela, Peru, and parts of Brazil, continue to experience regional or widespread influenza activity, with many reporting an increasing trend in the level of respiratory diseases. Widespread geographic activity is also reported in Central America and the Caribbean including Costa Rica, El Salvador, Guatemala, Honduras, Panama, and Cuba, but most of these countries are now reporting a declining trend, WHO said.	○ relevant ○ undecided ○ not relevant
54	**He pointed out that 500 Jordanian pilgrims were the first batch to take vaccines before their departure to Saudi Arabia on Monday.** He reported the discovery of 143 swine flue cases over the past four days that raised to 2,747 the accumulative number of people testing positive to the H1N1 virus so far. The minister said that 85 of those who had contracted the disease were still receiving medical treatment at hospitals.	○ relevant ○ undecided ○ not relevant

Figure 3.9 Examples of the incident reports selected at random from the classified instances.

(ECDC). As well, domain experts from the Joint Research Center in Italy, National Health Agencies in Germany (Robert-Koch Institute) and France (Health Institut de Veille Sanitaire); and a State Health Agency in Niedersachsen, Germany. Eight participants labeled 50 sentences, and one participant labeled all 100 sentences, in total 500 instances of sentences were labeled.

Users were instructed as follows:

Instructions: *In this evaluation you will be presented with a set of sentences in bold text, which may be potentially relevant as input for generating a signal. Your task will be to judge whether the sentence is actually potentially relevant or not, from a clinical or epidemiological point of view.*

1. Please read the sentences in bold text. If the meaning of bold text is not clear, then read the non-bold text immediately following the bold text.

2. After reading the text, put an X in one of the three circles: Relevant, Undecided or Not Relevant, according to the following:

Relevant: if the sentence has the potential to be clinically or epidemiologically relevant for generating a signal.

Undecided: if you are undecided about the potential of a sentence to be clinically or epidemiologically relevant for generating a signal.

Not Relevant: if you do not think a sentence is potentially clinically or epidemio-

logically relevant for generating a signal.

3. If you mark a sentence as "Undecided "or "Not Relevant" briefly comment why in the space provided.

Agreement Among Experts

Table 3.10 shows the overall percent agreement, PA, among the experts for the five teams. $PA(i)$ for a team,i, is computed according to:

$$PercentAgreement(i) = \frac{agreement}{total}$$

where *agreement* is the number of agreements among the judges within a team and *total* is the total number of units that where judged by the team. Percent agreement is bounded by 0 and 1.0: an agreement score of 0 represents no agreement, and a score of 1.0 represents perfect agreement [AP08].

Table 3.10 Per-team percent agreement for the relevance judgments of ten EI field practitioners and the overall average among the teams.

Team No.	Percent Aggreement
1	0.73
2	0.54
3	0.80
4	0.63
5	0.85
Average	0.71

Although the percent agreement is a simple agreement metric and does not take into account agreement to due to chance, we can observe the variations in the agreement across the teams and among the judges. We believe this variation to be due in part to the fact that the experts have a wide range of expertise and diverse responsibilities as EI investigators and we take this agreement measures into account when drawing final conclusions.

Expert and Classifier Agreement

Of the 500 instances that were labeled by the domain experts, the following results compared with the classifier were obtained:

- 275 instances of the incidents were considered to be relevant;
- 170 instances of the incidents were considered to be undecided;
- 55 instances of the incidents were considered to be not relevant.

Although 55 of the incidents that were classified as relevant by the classifier were not considered relevant by the domain experts. However, there is room for improvement since many of instances were undecided. The comments given by the users provide useful insights and in Table 3.11, we examine the user comments to understand these results in more detail.

Table 3.11 Summary of user comments obtained by manually grouping similar explanations given by the users during their assessment of automatic filtering classifier labels. The value in parenthesis denotes the comment's frequency.

Not Relevant Category	**Comments**
Off Topic *(29)*	advertisement, clinic trial, vaccination campaign
Historical*(11)*	outdated information
Non-Transmittable *(9)*	non-infectious disease
Personal Opinion *(7)*	expresses opinion of author
General Information *(45)*	not epidemiologically significant, literature review
No Outbreak *(37)*	no threat (action) required
Relevant Category	
Useful Knowledge *(17)*	new strain
Monitoring *(7)*	relevant for detection and monitoring
Undecided Category	
Not Enough Information *(11)*	confirmed case missing information symptoms or pathogen
Depends *(10)*	what illness; number cases; when, where take place

In some categories refinements to improve the classifier are easy. For example, with the category *Historical*, outdated information can be handled with an appropriate temporal filter. Also refining a list of medical conditions could help improve the fact that the classifier detected sentences containing *Non-transmittable* diseases as relevant. The category *Not Enough Information* suggests that even though we provided snippets of three sentences, there was still not enough information for the users to be certain about the relevance of the incident. We believe this is due, in part, to the fact that we only used sentences following the trigger and it would be

useful to build context from the sentences also **proceeding** the trigger sentences as well. This also has to do with the nature of short text classification in general and a broader scope may help to tackling this problem. For example, users commented that if no reason was explicitly linked to the medical condition, then it was hard for them to make a relevance judgment.

More challenging categories are: *Personal Opinion* and *General Information*. Text if often written in a factual way, but may not be useful epidemiologically. For example, whether an outbreak is relevant for an expert, depends of context of the user; their area of responsibility; or whether the cases have been imported into their country. Some users found messages that originated from a health department useful, while other users preferred the text that was identified strictly from the non-official sources.

Another important challenge in filtering, was the distinction between detection (sudden outbreak) versus monitoring (ongoing activity of a particular disease at a global level). For domain experts these are separate tasks, and they expect results to be filtered along these lines as well. For example, decreasing activity is not relevant for outbreak, but relevant for monitoring. Users suggested a widget whereby they could see information about general events, and not necessarily those related to outbreaks.

3.6 Chapter Summary and Outlook

On the onset, building a text filter for the task of EI seemed rather straightforward, but in practice, building a single automatic classifier to meet the needs of such a diverse set of epidemiologists is a challenge. In future work we will focus on ***an ensemble of staged classifiers, rather than a single, universal one***. An ensemble of classifiers would then allow us to tune the classifier to different types of information needs such as a filter for travel related infections; or a classifier for a particular concept, such as personal versus non-personal text. Also in future work, we will use clustering to group relevant instances and present incidents from clustering instead of a random selection as was done in this case study.

In our case study, we built our self-trained classifiers, by selecting the most confident instances at every iteration, to include as a new training instance for the following iteration. Using a most confident selection strategy automatically avoids including any instances for which the classifier is uncertain; but this does not imply that the uncertain instances are not good discriminators for the classification task at hand. The impact of a most confident selection strategy is that it is more difficult to build a classifier with a high accuracy (not recall gated). We also realize that when it comes to domain expert assessment forms of personalization and adaptation are needed to cope with the individual tasks.

One of the limitations of this case study, was that we did not get feedback from the users on the possible false negative predictions that were made by the classifier. Finally, as previously mentioned, one of the major drawbacks of short text at the

sentence level is the assumption that all the context is available within a single sentence. Existing work to address this problem expands the field of view around key sentences to include their surrounding sentence [PR09]. Ongoing work in this direction already shows promise with an F-Measures of 90% using named entities features for German text with a traditional classification task and a Support Vector Machine as a base classifier. Future work is needed to assess the performance level achievable for semi-supervised learning in the same setting.

In this chapter we have demonstrated that with our xLabel approach, it is possible to exploit the domain knowledge from disease outbreak reports to build a binary, syntactic parse tree-based, classifier that is capable of detecting disease reporting sentences in blogs and news. Our experiments show that with weakly labeled training data, we achieve a precision of 92% for the Precision Boosting strategy and a recall of 89% for the Recall Boosting strategy.

In the chapter that follows, we consider a selection strategy that focuses on using the *least* uncertain instance during the semi-supervised selection. Since the vast majority of recent work in EI is now devoted to disease reporting mentions in Twitter, we do so for the sparse text of Twitter messages.

Active Learning with Label Resolution

[1]

In this chapter, we focus on automatic text filtering using limited supervision with sparse text. The sparse text we use consists of tweets, or Twitter messages. In the medical domain, there has been a surge in detecting public health related tweets for Epidemic Intelligence (EI). Twitter is a popular micro-blog service that continues to surge as a means of sharing information in social networks. Given its real-time nature, coupled with the ease with which content can be created Twitter messages (or *tweets*) are now seen as a valuable auxiliary of relevant information for intelligence gathering tasks, such as natural disaster detection [SOM10, VHSP10] and tracking

[1] Image under License from Fotalia http://http://de.fotolia.com/

flu outbreaks [AMM11, CSN11, Cul10, PD11b, SKdQ10, SKL11, Dre12].

Unlike the short text presented in Chapter 3, sparse text is significantly different. Thus in Section 4.1, we begin by providing the reader with a deeper insight into what constitutes a disease reporting mention for tweets. We present examples of relevant and non-relevant disease reporting mentions and guidelines for defining the relevance criteria. We then motivate the active learning approach we take to detecting disease reporting mentions within sparse text. After providing an overview of related work, in Section 4.2, we then delve into the details of our *LaSAL*, approach (in Section 4.3), which uses semi-supervised clustering in an active learning setting to tackle the label bottleneck problem by mitigating the costs associated with a budgeted labeling strategy. In *LaSAL*, we attempt to offset the burden on HIT workers, by selecting from the large pool of unlabeled data points, a small, but productive number of instance that would still allow us to build a quality classifier, with as low a cost as possible. In Section 4.4, we present the experimental results validating our approach, including the results of a case study assessing the quality of HIT labels compared with experts. Finally in Section 4.5 we present a summary and outlook for the chapter that follows.

4.1 Sparse Text Characterization of Disease Reporting Mentions

4.1.1 Relevance Guidelines for Tweet in EI

Twitter is a micro-blogging service which serves as a means for sharing large volumes of real-world events ranging from a user's personal status to news reports. A tweet is a highly dynamic short, multilingual text, containing up to 140 characters. In addition to its sparsity, tweets have many peculiarities. They may consist of abbreviated lingo, URLs, and tweet tags or *hash-tags*. A tweet may refer to another tweet that was originally posted by the others (called a *re-tweet*); or a tweet may contain a mention to other tweeters (via "@" symbol). Despite these peculiarities, (in comparison to full and short text) Twitter messages are seen as a rich source of information.

Unfortunately, we found that related works in EI using Twitter do not explicate their relevance criteria. In an effort to address the needs of our EI system we undertook the task of first labeling tweets to better understand how they could be characterized. We summarize the work done in the context of the M-Eco project to develop labeling criteria for tweets. Then in conjunction with domain experts we converge on a set of criteria that we used in this work for determining the relevance of tweets for the task.

Unlike existing work in the domain of EI, the annotation guidelines set forth in by Collier et al. [CKCC09] for full text; as well as the criteria we outlined in determining a relevant disease mention in short text (Section 3.1); simply do not apply to sparse

text. The main reasons, similar to sentential level detection, is that diseases may be mentioned in a tweet in contexts, which do not imply relevance. This ambiguity problem is compounded by the sparsity and peculiarities of tweets. Within the M-Eco project, work done by Mustafa, et.al set forth the following criteria for determining a relevant tweet.

1. *Putative Case*, e.g, three Chinese are suspected to have swine flu.
2. *Probable Case*, which is a suspect case with some evidence like X-rays
3. *Confirmed Case*, which confirms that someone is directly infected by an outbreak.
4. *Self Reporting*, an individual mentions that they currently have symptoms or ailments associated with and infectious disease
5. *Third Person Reporting*, an individual mentions that someone they know currently has symptoms or ailments associated with and infectious disease

A tweet was considered *relevant* when an individual person is providing information about his own or someone else's health status and *irrelevant* does not fulfill this criteria. Any tweet which confirms that there is no case or which contains text that is unrelated to a case is labeled *irrelevant*. Examples of relevant and irrelevant instances of tweets are shown below.

Examples of relevant tweets:

Any tweet should be labeled as relevant regardless whether it is a confirmed, putative or probable case, if:

1. It confirms that the user is infected with a disease or symptom, e.g, *I am sick now. I got influenza and I need medicine*,
2. It confirms that another subject (e.g., animal,) has a disease or symptom,
3. A test result is mentioned which confirms an infection, e.g. *Tyler is influenza positive!!!!.*
4. A suspicion is mentioned, e.g., *my son is suspected to has swine flu.*
5. Another outbreak or danger is described.

Examples of irrelevant tweets: An irrelevant tweets is any tweet that:

1. is a question, e.g. *What is this Bieber Fever Thing?*
2. contains a condition, e.g. *If I have the flu again I will kill someone.*

3. offers advices like *#Kids health:you should prevent your child from getting #dengue fever*
4. negates an infection, e.g, *I don't have measles*,
5. contains a disease definition, statistics, or describes past outbreaks, jokes about diseases or outbreaks.
6. is outside of the disease outbreak domain.

Mustafa, et.al also take into account verb classes as a set of useful features; example trigger verbs are shown in Table 4.1.

Table 4.1 Examples for useful words for labeling a tweet as relevant

Word Category	**Example**
Infection Verbs	affect, infect, got, come down, suspect, down with, have, has
Detection Verbs	find, confirm, detect, discover
Medical Terms	death, fatality, case, hospital, patient, victim, clinic, pain, ill, sick, ache, doctor, outbreak, hurt, inflammation, negative test

4.1.2 Feedback from Domain Experts

After the initial information gathering stage, together with epidemiologists, we refined the above set of criteria so that even non-domain experts could apply the criteria consistently. In this way, we could tackle the label bottleneck with a budgeted strategy and obtain relevance judgments from hired non-expert, Human Intelligence Task (HIT) workers, within an active learning setting. Table 4.2 lists the criteria we developed with domain experts to decide if a tweet is relevant for EI or not.

Table 4.2 Criteria for labeling tweets according to their relevance for EI in conjunction with domain experts.

Relevance	**Description**
Relevant	Somebody reports himself or another person being ill
Irrelevant	No one is suffering from symptoms; i.e., mentions refers to opinion, advertising, jokes, music, books, films, artists, landmarks, sporting events, slang, etc.

The criteria are notably less strict then, short and full text criteria, and often times in practice a relevance judgment is not to make, given the limited information contained within a tweet, but experts found these criteria easier to apply.

Table 4.3 Examples of Irrelevant Tweets for the task of Epidemic Investigation.

Literature	"A two hour train journey, Love In the Time of Cholera ..."
Music	"Dengue Fever's "Uku," Mixed by Paul Dreux Smith Universal Audio..."
Marketing	"Exclusive distributor of high quality #HIV/AIDS Blood & Urine and #Hepatitis #Self -testers.
Vaccination Campaign	"Rotavirus vaccine greatly reduced gastroenteritis hospitalizations in children: ...
General	"Identification of genotype 4 Hepatitis E virus binding proteins on swine liver cells: Hepatitis E virus..."
Negative	"i dont have sniffles and no real coughing..well its coughing but not like an influenza cough."
Joke	"Thought I had Bieber Fever. Ends up I just had a combo of the mumps, mono, measles & the hershey squ..."
Off Topic	"Its raining like severe diarrhea"

4.1.3 Ambiguity and Limited Context of Tweets

The examples presented above represent rather clear cut cases. Given the lack of contextual information, we found is was even challenging for the experts to determine if a tweet is relevant or not. Given the volume of tweets, an automatic interpretation is necessary, yet challenging, due to the variance in content, ambiguity of language, and sparsity of the text. Consider the following tweets. ‘‘`This condition is making me sick?`’’ which could imply that somebody suffers from a medical condition, where as the tweet ‘‘`...Beraee is my buddy cough, cough future wife cough`’’ does not explicitly mention aliment, but uses injection to express feelings about some situation. Whereas tweets such as: ‘‘`face it me and ace is sick like malaria carriers`’’ mention infectious disease used as analogy or metaphor. Still further, the mention of symptoms such as `fever, cough, headache` or infectious diseases such `cholera, anthrax` in English natural language can be used to express many concepts from the excitement over an event: "Royal Wedding Fever"; "Spring Fever"; liking for a celebrity: "Justin Bieber Fever", "Austin Fever", "Canuck Fever"; preference for human characteristics: "Scarlet Fever", "Korean Fever"; "Yellow Fever"; performing artist: "Anthrax Band"; book titles "*Love in the Time of Cholera*", etc. These examples only represent the tip of the iceberg.

We see from these examples, that in addition to the limited context, we are also faced with the problem of discriminating the sense of a word. In the section that follows, we discuss semi-supervised learners that seek to tackle the problem of limited supervision while simultaneously handling the exclusion between ambiguous concepts that are being learned.

4.2 Related Work

4.2.1 Semi-Supervised Learning with Mutual Exclusion

One of the main challenges faced by many learners, is the ability to handle the mutual exclusion between relevant and irrelevant target concepts. Naive approaches operate under the one-sense-per-meaning principle, namely, that: (1) terms only have a single sense; and (2) context used to represent the concept only extract terms with a single sense. This assumption that context are mutually exclusive is, in general, for human language text far from true in practice, as illustrated by numerous examples in Section 4.1.3.

For semi-superivsed learning this problem is surfaces as semantic drift. Semantic Drift occurs when a term with multiple senses is selected and incorrectly incorporated in the training data. As a result, the context of candidate instances tend to become more similar to the recently added instances, drifting away from the context of the seeds and the high precision instances of the early iterations. In the semantic drift problem, the goal is to maintain a word's sense, and additionally its labels (relevant

or irrelevant) along with it.

Bootstrap aggregating (bagging) [MC09] is a general approach which seeks to exploit knowledge from multiple views over the data. It proposes to combine multiple supervised learning models trained by randomly resampling data from the same source training set. In contention methods, such as mutual exclusion bootstrapping [CMS07] Weighted MuxB [MC09], multiple semantic categories simultaneously compete with one another in an attempt to actively direct categories away from each other. In agreement based methods such as: co-training [BM98], instead of asking a human, the response from multiple classifiers is used, and resolution of a label is accomplished by specifying an agreement (disagreement) strategy on the labels provided by each automated prediction. Mutual exclusion in bootstrapped learning has also been ensured by manually craft negative categories [McI10] and knowing which tasks share semantic spaces, in coupled learning [CBW+10]. In general is difficult to predict if a category will share overlapping context and compete with other target. Moreover, the success of the approach also greatly depends on the negative categories.

In bootstrapped learning, clustering has also been used to ensure mutual exclusion in the presence of semantic drift [CMS07, MC09]. The underlying intuition for using clustering to support mutual exclusion is that we can disambiguate a concept by the words it co-occurs with [PP04]. We build upon this basic assumption and incorporate the use of clustering. Differently however, these works is the sampling method used. These passive passive learners selects the most confidence instances at each iteration. However, we employ an aggressive selection strategy - using the most uncertain instances at each iteration of the semi-supervise learning process. Moreover, we seek advice for labeling data from humans with a limited budget, thus take up an active learning approach.

4.2.2 Active Learning with Clustering

Active learning (AL) is a methodology for building a trainable classifier, that attempts to reduce the cost or burden of manually labeling training data. Clustering has also been successfully used in many active learning strategies to help reduce the number of queries needed [NS04, BMC11]. They start by clustering the entire pool of unlabeled data (global clustering) and then select for each cluster a set of instances to be labeled by the oracle. However, such approaches have been known to suffer when: (i) no obvious clustering exists; (ii) clusterings exist, but are at an unknown granularities; (iii) the classifier labels themselves are not be aligned with the active learner clusters (label-cluster alignment problem) [Das11].

In our approach to active learning, we also using clustering techniques to help reduce labeling costs, but in unlike existing active learning approaches, we not only tackle the problem of ensuring mutual exclusion; but also use semi-superivsed learning as an additional mechanism to tackling the problem of have a limited amount of training data to build the learner. Moreover, we perform a *locally* clustering (i.e.,

during each iteration of the semi-supervised learning process) as opposed to a global clustering strategy. Local knowledge captures the state of the current inferences and feature for the latest batch of data.

Finally we build upon generative model for clustering tweets. The fundamental difference is that a generative model describes how instances of each class are generated (commonly denoted by plate notation); For example the model associates each tweet with some (latent) topics and each topic with some significant words. A discriminative model on the the other hand, does not specify how to generate instances, but rather by specifying how to divide up the space into class regions. In our approach we also seed to simultaneously seek to minimize the side-effect of mislabeling by taking into account confidence levels to capture the conditions under which we can trust the clusters. In the section that follows, we introduce terminology to describe our problem more formally and then present the details of our approach.

4.3 LaSAL: Semisupervised Clustering with Active Learning

4.3.1 Motivation

In Chapter 3, we tackled the label bottleneck problem using forms of weak labeling, and semi-supervised learning (SSL). In this chapter we integrate the limited supervision approach of SSL with a budgeted labeling strategy, or active learning. During the process of SSL we pro-actively select instances, with the intension of building the best performing classifier, with as few requests as possible, for a fixed budget. Independent of the query selection strategy employed, the central problem faced in all active learners is one of measuring the information content of the unlabeled data point and the use of sampling the most uncertainty instances has consistently been showed to be a simply, yet effective query selection strategy, yielding a good classifier performance that converges with fewer requests for labels[TK02].

Ideally, we would like to select the most uncertain samples, (thereby having a fast converging classifier) but also resolve the labels without human intervention. Specifically, resolving as many of the labels for uncertain samples as possible without asking an oracle, would allow us to tackle the label bottleneck problem and reduce the cost associated with our budgeted labeling strategy. This is the aim of this work.

Existing approaches to active learning have used clustering to reduce the number of annotations that would be otherwise be deferred to a human, but suffer when the classifier labels themselves are not aligned with the active learner clusters (label-cluster alignment problem) [Das11]. In our work, we also use clustering to reduce the number of annotation needed, but facilitate the assessment of a true label for a dubious instance with a semisupervised clustering based on a Partially Labeled Dirichlet Allocation (PLDA) [RMD11]. The PLDA allows us to align the clustering

with the labels of the training instances. We refer to our approach as Label-Aligned Semi-superivsed Active Learning or *LaSAL*.

4.3.2 Terminology and Overview

An overview of our approach to Label-Aligned Clustering (*LAC*) using pool-based, semi-superivsed active learning is depicted in Figure 4.1. It is similar to the the SSL algorithm outlined in Algorithm 2. The main differences however, are: oracle intervention, which is needed, since we now select instances that have the most **uncertain** classifier predictions, and a component which attempts to automatically resolve the labels of these uncertain instances. This is in stark contrast to the approach taken in our previous chapter, in which the most certain classifier predictions were used, no human intervention was considered.

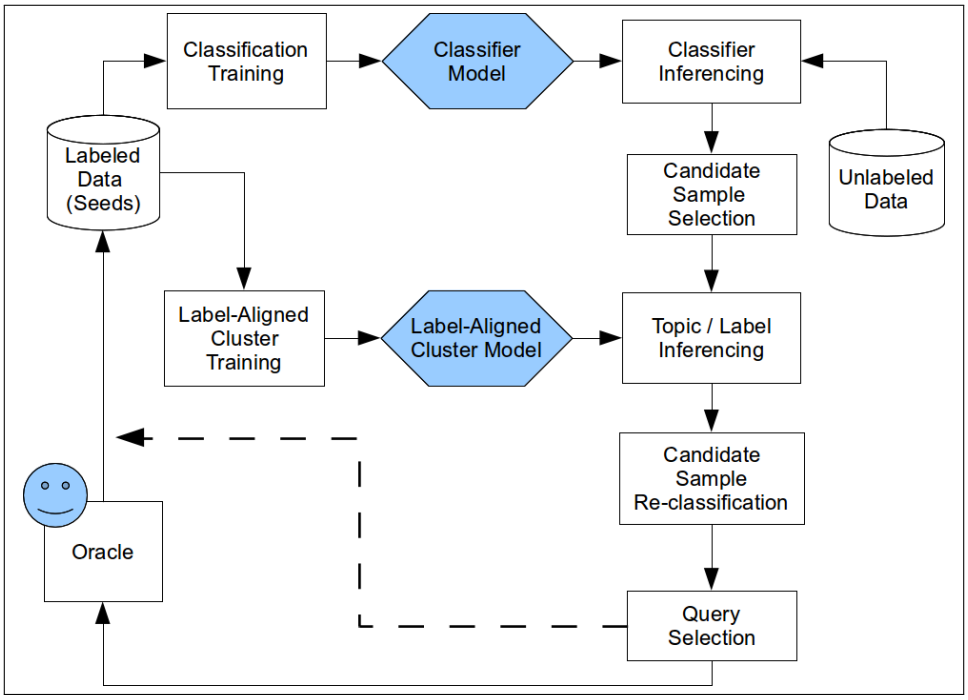


Figure 4.1 Overview of a *LaSAL*, a pool-based semisupervised active learner using label-aligned clustering for reducing the number of queries presented to an oracle.

We start with a small set of initial `Labeled Data` (or seeds), $\mathcal{S} = \{s_1, ds2...s_l\} \in \mathbb{R}^N$, representing instances of documents that have been manually labeled by a human as either relevant, +1, or irrelevant, -1 with respect to our task. Throughout the discussion, we ignore the distinction between *relevant* or *irrelevant* instance when doing so lends to no ambiguity.

Classification Training. We use a pool-based active learning. A pool-based active

learner is a quadruple, $\Gamma = (\mathcal{H}, \mathcal{Q}, \mathcal{S}, \mathcal{U})$, where $\mathcal{U} = \{u_1, u_2...u_l\} \in \mathbb{R}^N$ is a set of `Unlabeled Data` instances; $\mathcal{H} : U \rightarrow l \in L = \{+1, -1\}$, is a *binary classifier*, which have been trained on the current set of labeled instances in S.

Classifier Inferencing. Once training on the current set of seeds is complete for the current iteration,i, the `Classifier Model`, H is used to map $\mathcal{U}$ to L. We denote these as inferred instances and denote then as D.

Candidate Sampling. The query function, $Q(D) : \tilde{D} = \left\{\tilde{d_1}, \tilde{d_2}...\tilde{d_\beta}\right\} \mathcal{U}$, is used to select a candidate set of instances from the inferred instances (sampling), $\beta \geq 1$ is the size of the sample. We use the notion $\tilde{d} \in \tilde{D}$ to denote a instance whose label is inferred with uncertainty; and $\hat{d} \in \hat{D}$ to denote a instance whose label is inferred with confidence. We create a set of Candidate Samples,$\tilde{D}$, by selecting a subset of the most uncertain instances.

Topic/Label Inferencing. The query function, Q is also used to select a subset of confident samples,$\hat{D}$. We use: (1) $\hat{D}$; (2) along with a subset of seeds from any previous iterations ($\hat{S}_{0,\cdots,i-1}$); and (3) their corresponding labels, L, as input seeds to train a `Label-Aligned Cluster Model`. The cluster model is used to map an instance to a set of topics, $\Phi = 1, 2, ..\mathbb{N}$; and (2) infer (possibly new) labels for the uncertain Candidate Samples for the current iteration, $\tilde{D}_i$.

Candidate Sample Re-classification. Optionally, the set of topics, Φ, can be used as features to retrain a SSL classifier using the seeds for the current iteration($\hat{D}_i$); along with a subset of seeds from any previous iterations ($\hat{S}_{0,\cdots,i-1}$) to infer new labels for the uncertain samples of the current iteration.

Query Selection. Using either the labels inferred by the `Topic/Label Inferencing` stage or `Candidate Sample Re-classification` stage, Candidate Samples, $\tilde{D}$, are fed into the pool of `Labeled Data`. Those candidates $\tilde{D}$, which fall below a threshold level of confidence are ushered to the oracle so that their true labels can be obtained. All samples, those whose labels have been successfully resolved; either by the oracle; the Label-Aligned Clustering inferencing; or `Candidate Sample Re-classification`, are included in the existing set of labeled data, and the cycle repeats until a desired stopping condition is met.

4.3.3 Problem Statement

Let $\mathcal{S}$ be a set of instances and $L = \{+1, -1\}$ be a set of possible labels. Further, let H be their hypothesis, a set of functions mapping $\mathcal{S}$ to L. We assume there is a distribution, P, over the instances in $\mathcal{S}$, and that the instances are labeled by multiple, heterogeneous oracles, O. Let acc_O denote the minimum accuracy of any hypothesis in H with respect to the distribution induced by $\mathcal{S}$ and the labeling oracle, O; and η_O be the number of queries needed to achieve accuracy, acc_O. Similarly, let acc_R denote the maximum accuracy of any hypothesis in H with respect to the distribution induced by $\mathcal{S}$ and an automatic label resolution strategy, R; and η_R be the number of

queries needed to achieve accuracy acc_R. The goal is find a hypothesis $h \in H$ using R, that has an accuracy within a tolerance, ϵ of acc_O, such that $\eta_R \leq \eta_O$.

The intuition is that although the majority of uncertain samples might be misclassified, some of the uncertain samples can be more readily resolved automatically. We propose an active learning selection strategy to reduce the annotation effort by exploiting semi-supervised clustering to identify and handle: uncertainty that arises due to non-separable context between training seeds. In the discussion that follows, we describe the details of our *LaSAL* approach, and the machinery underlying the automatic label resolution strategy, R.

4.3.4 Label-Aligned Cluster Training

Partially Labeled Dirichlet Allocation (PLDA) probabilistic generative graphical model.It is built upon the Labeled Dirichlet Allocation and is used to express the coupling between words and *labeled* documents, by introducing an unobserved (hidden) variables to capture the notion of a topic [RMD11, SG07]. Given a collection of documents $s \in \mathcal{S}$, each containing a multi-set of words w_s from a vocabulary W and a set of labels, $l \in L$. The goal of PLDA is to recover, for each label, $l \in L$, an association that fits the observed distribution of words in the labeled documents. For a set of topics, Φ, the multinomial probability distribution, $P(\Phi|s))$, describes the probability that the document is devoted to a topic; where each topic is a distribution over words, $w_s \in W$, that tend to co-occur with each other. To train the PLDA, we use the labels from the confidently labels seeds, $\hat{D} \cup \hat{S}$ at each SSL iteration. Each $l \in L = \{+1, -1\}$ is assigned some number of topics that are unique to that label. This is a crucial aspect in our ability to uniquely assign a label to an unseen instance, since we have binary classification task, there are only use two labels; so the label and clustering are aligned (hence the term label-aligned clustering). We compute the number of topics assigned to the relevant labels, $\|\Phi^{\oplus}\|$ and the number of topics assigned irrelevant labels, $\|\Phi^{\ominus}\|$, for the current iteration,i, according to equations 4.1 and 4.2, respectively.

$$\|\Phi_i^{\oplus}\| = Ceiling(|S_i^{\oplus}| * \gamma) \tag{4.1}$$

$$\|\Phi_i^{\ominus}\| = Ceiling(|S_i^{\ominus}| * \gamma) \tag{4.2}$$

where $\gamma = .05$. The PLDA assumes that each topic, Φ takes part in exactly one label, so $\Phi_i = \Phi_i^{\oplus} \cap \Phi_i^{\ominus} = \oslash$. Additionally, let L^b represent a set of background labels, used to denote a *shared* latent, set of topic classes, Φ^b that is applied to all seeds, in the collection for the current iteration, where $\Phi_i = \Phi^{\odot} \cup \Phi_i^{\ominus} \cup \Phi_i^b$. We compute the number of background topics, for each iteration, Φ_i^b, according to the formula in Equation 4.3:

$$\|\Phi_i^b\| = Floor(\log(\|\Phi_i^{\oplus}\| + \|\Phi_i^{\ominus}\|)) \tag{4.3}$$

By incorporating the latent class of topics in addition to the label classes, the model is essentially driven by deciding whether each word to better modeled by a broad, latent topic, or a topic that applies specifically to one of its document's labels. The outputs of the PLDA are: a distribution, $P(\Phi|s)$, describing the probability that each seed document,$\hat{s}$ is devoted to a topic in Φ; and a distribution describing the probability that each word belongs to a topic.

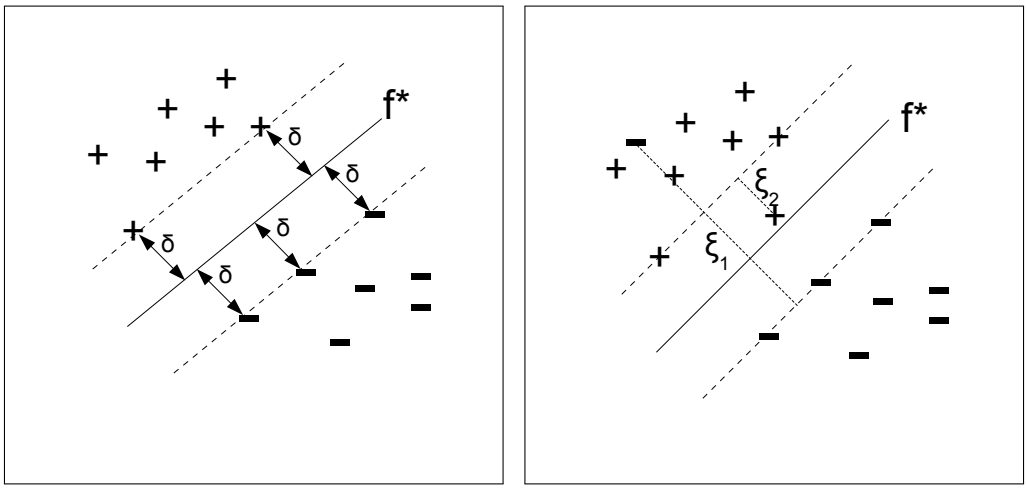
We emphasis the two crucial aspects of our appraoch using PLDA. First, since each $l \in L = \{+1, -1\}$ is assigned some number of topics that are unique to that label: so the label and clustering are aligned (hence the term label-aligned clustering, $\Phi_i = \Phi_i^{\oplus} \cap \Phi_i^{\ominus} = \oslash$). This alignment allows us to unambiguously assign a class label to an unseen instance. Second, the use of background topics allows us model any context that might be overlapping with a separate set of topics, which we can subsequently eliminate to ensure mutual exclusion between the seeds is maintained.

Global versus Local Clustering. Intuitively, we would expect that so long as (1) the context (i.e. topics) of the seeds are in separable classes; and (2) the uncertain samples fall into one of the mutual exclusive classes; then resolution of the uncertain samples can be accomplished using other types of clustering, besides PLDA, such LDA, globally. The advantages of a global strategy is that we could cluster once and apply the same clustering to each iteration. Differently from our PLDA strategy where a new clustering is performed at each iteration, albeit for a relatively small number of documents.

Is important to note however, that this global clustering does not ensure a mutual exclusion, nor does it align labels with clusters, since the labels for the entire pool of data is simply not known in advanced, but learned per iteration via SSL. Nonetheless, if we were to use a global clustering strategy, in a manner that is similar to PLDA, we could heuristically enforce mutual exclusion among the seeds using the LDA topics on at each iteration, to assign an unseen instance to a class in mutual exclusion-aware manner. We take up the notion of a mutual exclusion-aware strategy in the experimental section. Next we describe the selection of the most uncertain candidate instances,$\tilde{D}$.

4.3.5 Candidate Sample Selection

Bag-of-words is a common method of representing a text document in terms of a feature vector and is use in this work for tweets. Even though tweets are sparse, their bag-of-words representation still leads to a high-dimensional feature. This means that a classification algorithm must be able to handle high-dimensional training input well. Support Vector Machines (SVMs) have been shown to be a good choice for



(a) Linear Separable **(b)** Linear Inseparable

Figure 4.2 Fig (a) shows the hyperplane f^* that separates those examples with maximum margin δ. Fig (b) depicts a linearly inseparable classification task in 2-dimensional space. f^* is a solution for the soft-margin optimization problem with slack variables ξ_1 and ξ_2.

high-dimensional text classification tasks [VLC94, TK02]. An SVM learns a linear function, given in Eq. 4.4, to find the largest margin between the closest negative (irrelevant) examples and the closest positive (relevant) examples. This margin δ is shown in Fig. 4.2a. The examples closest to the hyperplane, with distance exactly δ, are the support vectors.

$$f(x) = sgn\,(w \cdot x + b) = \begin{cases} +1 & \text{if } w \cdot x + b > 0 \\ -1 & \text{otherwise} \end{cases} \tag{4.4}$$

The sgn function returns -1 if the argument is negative and +1 otherwise. This means the predicted label of an example is given by the side of the hyperplane $w \cdot x + b$, with normal vector w and bias b, it lies on. Formally, the requirement that all training examples x_i with labels y_i lie on the correct side of the hyperplane is given in Eq. 4.5.

$$y_i(w \cdot x_i + b) > 0 \tag{4.5}$$

Most words in natural language have multiple possible meanings that can only be determined by considering the context in which they occur. Often a well defined context can not be easily defined. In practice, the training examples of most text classification tasks, are not linearly separable, and these constitute uncertain samples that are selected with our `uncertainty sampling` selection strategy. Figure 4.2b depicts depicts a linearly inseparable classification task in 2-dimensional space.

These uncertain samples are selected with our query function, Q as follows. First, the current classification model, H, is applied to the set of unlabeled instances, U, to infer their labels. Second, an information measure is applied to the relevant and irrelevant instances; and a partial ordering over the the instances are made using the measure. We use the magnitude of log-likelihood of the classifier predictions for as an information measure. Third, we define a small fixed number of instances, as the sample size, β; and build a set of uncertain samples by taking the smallest log-likelihood values(most uncertain samples) as the Candidate Samples, denoted by $\tilde{D}$.

4.3.6 Topic-Label Inferencing

PLDA also allows for approximate inferencing so that the assignments for both: per-document distributions over labels and topics; as well as the set of words associated with each label, can be made on a previously unseen document collection. Inferencing is performed on the most uncertain instances $\tilde{D}$, and we refer to this as **label resolution**. A key point of the PLDA is that we can eliminate all background topics from the seeds, $\mathcal{S}$ and candidate samples, $\tilde{D}$ in this way, mutual exclusion among $\mathcal{S}_i^{\oplus}$ and $\mathcal{S}_i^{\ominus}$ is guaranteed. Specifically, let $\Phi_i^{\oplus}$ be the set of topics induced by the PLDA for the relevant set of seeds, and $\Phi_i^{\ominus}$ be the set of topics induced from the irrelevant set of seeds, then after background topics are eliminated from each set of seeds, $\Phi_i^{\oplus} \cap \Phi_i^{\ominus} = \oslash$.

PLDA is feasible approach for handling overlapping context since any background topics uncovered by the PLDA model are eliminated prior to constructing the feature vector for a classifier. Moreover, the PLDA model essentially transforms the topic space, for both relevant and non-relevant seeds, into a set of sub-topics, which may not be revealed in a global clusterings. This, in effect, allows a once non-shared context between the seeds and uncertain samples, to share context. We make use of this exclusion and sub-clustering during the `Candidate Sample Re-classification` phase, described next.

4.3.7 Candidate Sample Re-classification

In this stage, the inferred labels and discovered topics as used as features for retraining a local (per iteration) classifier. In order to do so, we must first convert the topic distribution to features. The results of the PLDA, yields a distribution such that every document is associated with each topic,with some probability. We can, at one extreme, construct a topic context using all the topic is the PLDA model. The practical limitation in doing so however, is that not all topics assigned to a document are equally confident. Thus, we opt for limiting the number of topics applied to a single document, and define an additional parameter, α, used to represent the threshold probability a topic must have to be included in the resolution process.

4.3.8 Query Selection

We do not want to introduce more mislabeling of the uncertain examples than is already introduced by the base classifier's inference. It should be noted that PLDA is not always capable of making an inference prediction. There are three cases for which we do not choose to rely upon the predicted label of the PLDA: (1) given the sparsity of a tweet, the features sometimes degenerate so that counts for building the PLDA model are insufficient to make a prediction; (2) a prediction is not used when the instance only belongs to a background topic; (3) finally, when the prediction probability is below a threshold level, we also do not consider its label in our task. If one of these criteria fails, we then defer to the oracle to obtain a true label for an instance. Otherwise, the label of the uncertain document is considered to be resolved automatically by the PLDA inferencing. Ideally a very small subset of the original uncertain samples $\tilde{D} \subset \tilde{D}$, is presented to the `Oracle`: we assume $\left|\tilde{D}\right| \ll |U|$. When the label is resolved, either from inferencing, re-classification or the oracle, the document is augmented with the set of labeled data to be included as a new seeds for the next iteration.

4.4 Experiments

We have proposed, *LaSAL*, semi-supervised, pool-based active learner. *LaSAL*, uses label-aligned clustering, *LAC* in order to reduce the number of requests for labels that are presented to an oracle. In this section we evaluate the efficacy of our approach.

4.4.1 Experimental Goals

Recall from our previous discussion that that PLDA affords us a clustering, per class label (Section 4.3.4) ; as well as capabilities to perform inferring for these labels (Section 4.3.6). We are interesting in evaluating the extent to which both can be exploited in an active learning setting. More specifically, we consider two different approaches to the tackling the label bottleneck problem: **PLDA Topic Driven Resolution** and **PLDA Inference Driven Resolution**.

In PLDA Topic Driven Resolution, we use the topics afforded by the semi-supervised clustering as features to retrain a secondary binary classifier in which the background topics have been eliminated from the the training seeds. The label of the uncertain sample is obtained by classifier inferencing. In PLDA Inference Driven Resolution, we exploit the inferences capabilities to automatically resolve labels of uncertain samples and feed them, as seeds, back into the SSL process.

In Section 4.4.4, we first compare the accuracy of different active learning strategy (*ALU*, *ALR*, *ALC*) and passive learning strategies (*PLU*, *PLR*, *PLC* - similar to the ones used in Chapter 3). The strategies presented in Part I form the baselines

for the remaining experiments. For ALU, samples closest to the hyperplane are selected; ALR, a random subset close to the hyperplane is used and; for ALC, the most confident of the uncertain samples were selected. Likewise we PLU, PLR, PLC except that the selection is made from the instances **farthest** away from the hyperplane. Section 4.4.5 focuses on PLDA Topic and Inference Driven Resolution for tackling the label bottleneck. We compare the accuracy and costs of classifier that we build by automatically resolving the labels of uncertain samples using label-aligned clusters LAC for which background topics have been eliminated. We compare the LAC against a baseline strategy $MuxA$ in which global clusters are using and mutual exclusion is enforced. PLDA Inference Driven Resolution for tackling the label bottleneck problem. We use the baselines from Part I, but instead compare the accuracy and cost of a classifier that has been built purely from resolving the labels of uncertain samples with PLDA-inferenced labels. Finally, in Section 4.4.6, we measure the percent agreement between labels obtained from crowdsourcing labels again a baseline consisting of labels obtained from subject matter experts. When labels can not be automatically resolved with LAC, we must defer to the oracle for providing a judgment; we examine the extent to which such judgments obtained from for-hire workers can be used, as a proxy for judgments that would otherwise be obtained from subject matter experts.

4.4.2 Data Set and Summary

In realizing the aforementioned goals, we collected 14,725,788 Tweets gathered during a 3-month period from April 1, 2011 through June 30, 2011. The data was collected by using the Twitter stream collector for list of 1,258 terms consisting of infectious diseases, their synonyms, pathogens and symptoms, which are provided by the domain experts. A random subset of consisting of 36,067 Tweets were used and exact duplicates where removed, leaving a total of 22,514 for experimentation. All tweets were processed using a series of language processing tools, including Lucene English Analyzer [2] for tokenization, stemming and indexing; and Ark Tweet [GSO+11] for part-of-speech tagging; and entity taggers outlined in Section 3.1.2 for named entity recognition.

Stop words were also removed. The stop word list was created using the tags produced by the Ark Tweet parts-of-speech tagging. The stop list consisted of words whose class was tagged as: determinants; subordinating conjunction; WH-words, non-possessive pronouns; and urls. The tagger allowed us to eliminated lingo-specific variations that would not be detected with a traditional stop list of grammatically correct terms. Table 4.4 show example terms included in our stop list.

[2] http://lucene.apache.org/

Table 4.4 Example stop list terms obtained by filtering Twitter terms based on their word class (i.e., part-of-speech).

Word Class	Stop List Term	Grammatical Term
Determinants	da, dah, de, deh, teh, thee, theeeeee, tha	the
	dat, daaat	that
	alll, allllll, alllllll	all
Subordinating Conjunction	dwn	down
	b/c, bc, bc0z, Bcoz, bcus, bcuz, bcz, cus, cuz, cuuz	becuase
	b4, be4	before
Non-personal pronuoun/WH	yuu, Yooooouuuu, yoooouuuuu, YOOU, youh, youu, youuuu	you
	wen, wenn,	when
	wth, wit, wt	with
	w/, w/'da, w/2, w/a, w/his, w/my, w/o, w/the, w/this	with*
	whaaaap'n	what's happening
	wha, wutttt, wut, wht, Whud, Whut, wot	what
	ether, ethr	either
	whre, whr	where

4.4.3 Experimental Setting

The classifier performance was evaluated on a hold data set consisting of 2,000 examples, (1000 positive, and 1000 negative) using the precision, recall, F_1-Measure and accuracy, as computed according to 3.1, 3.2, 3.3, and 3.4, respectively.

We measure the oracle labeling cost, $Cost_O$, of an active learning strategy as some multiple of the number uncertain samples for which an oracle is asked to give the true labels. We assume the number of samples per iteration is a fixed number; and that there is a unit cost is associated with each request. The total oracle labeling cost is then computed based on the number iterations n, of the learner as given in Equation 4.6.

$$cost_O(n) = \sum_{i=1}^{n} (cost * nrSamples_i) \tag{4.6}$$

,where $cost$ is the cost per unit, and $nrSamples_i$ is the number of samples for iteration, i We measure the cost of a LAC strategy, $cost_{LAC}$, as the number uncertain samples that we were able to resolved without requesting input from the oracle. The total cost a LAC strategy is then computed based on the number iterations n as given in

Equation 4.7.

$$cost_{LAC}(n) = \sum_{i=1}^{n}(cost * (nrSamples_i - nrResolved)) \tag{4.7}$$

where $nrResolved$ is the number of resolved labels. The cost savings is thus computes according to Equation 4.8:

$$savings = cost_O(n) - cost_{LAC}(n) \tag{4.8}$$

PLDA Parameter Settings: For training the PLDA, we set the number of topics per class to be some fraction of the total number of documents per class, as given by Equation 4.2. We set the number background topics relative to the total number of topics, as in Equation 4.3. We set both the PLDA term and topic smoothing to be 0.01; and run the PLDA for a maximum of 1,000 iterations.

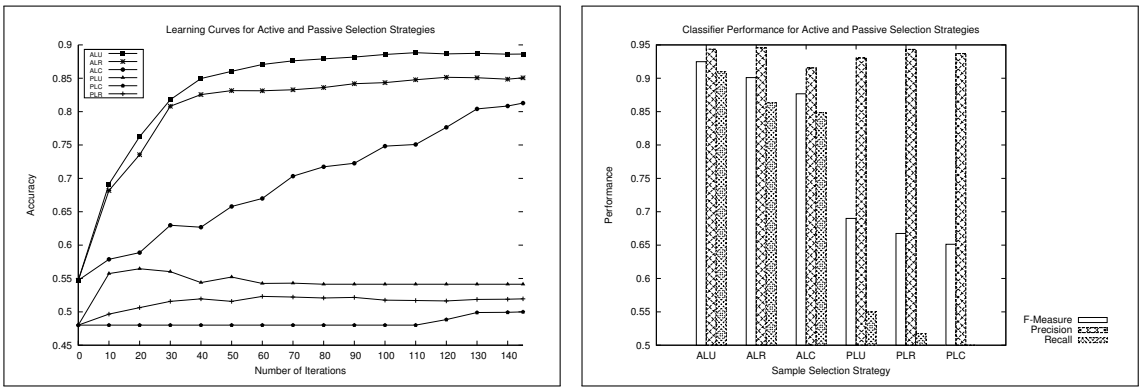
4.4.4 Results I. Selection Strategy and Ngram Features

The goal of this experiment is to determine whether there is a clear advantage to using and active learning strategy (ALU, ALR, ALC) in comparison to the a passive strategy (PLU, PLR, PLC similar to the ones used in the Chapter 3). We also seek to determine which active (and passive) learning strategy is best.

Active versus Passive Selection

Figure 4.3a shows the SSL classifier performance under different active (ALU, ALR, ALC) and passive (PLU, PLC, PLR) sample selection strategies. We first examine which select strategy would lead to a better performing classifier for our task. Passive selection was used previously in Chapter 3. As mentioned in Section 2 the main difference between active learner and a passive learner is the querying component, QU that determines which samples will be drawn from the pool of unlabeled data to include as candidates in the next round of semi-supervised learner (SSL). The passive query strategy selects only from most *confident* instances at each iteration. From these confident instance we select a random subset (PLR); a top-N subset (PLC) or the bottom-N subset (PLU). In contrast to a passive learner, an active learner more aggressively seeks the most *productive* instances and presents them to an oracle in the form of queries, who then resolves their true labels. The true labels of these productive instances are then used input for the next round of training. In our case the most productive instances, are considered to be those that lie close to the decision boundary of the SVM Section 4.3.5.

At iteration zero we assume an initial set of seeds were given to start the training process, likewise in all experiments we assumed 10 positive and 10 negative initial sees



(a) Learning Curve Accuracy **(b)** F-Measure,Precision,Recall

Figure 4.3 Fig (a) shows the accuracy of classifier at iteration of the semisupervised learning process. Fig (b) shows the corresponding average F-measure, precision and recall.

were used, and we disregard the cost of obtaining these initial labels. First we notice, that compared to an active strategy, the passive strategy, has a considerable lower accuracy (no more than .60%, which does improve when using more training instances (higher number of iterations). In contrast, we notice a marked improvement in the accuracy over the passive strategy for any of the active learning strategy. When taking only an active learning strategy into account - we notice a strategy which selects most confident instances and, presents them to the user for labeling under-performs other AL strategies.; whereas the uncertainty sampling strategy outperforms all others.

Initially, selecting instances at random (ALR) and presenting them to the user for labeling performs almost as well as a strategy which selected the most uncertain ones (ALU). However we notice that as the active learner proceeds, the gap between ALR and ALU widens; the uncertain strategy performance outstripping the performance of the random strategy. At a cost factor of roughly 30 queries, we are able to achieve an accuracy of 80% and at a cost of 55-60 queries, we reach an accuracy of 90%. We seek to find out if how much we can reduce these costs even further from additional, knowledge that has been obtained in an unsupervised manner. The most uncertainty instances always outperforms the random and most confident strategies; in the remaining experiments, we take ALU, ALR and ALC as baselines.

We see from the above results that acquiring the labels for the most uncertain samples has a high utility. How uncertain are these same uncertain samples for ALU without any human intervention? When the accuracy of predicted samples are measured in the absence of any type of resolution, at best, we achieve level of accuracy of %65. This is a clear indication that the utility gained from resolving the labels of an ALU can leads to an optimal classifier (good accuracy and faster convergence), when compared with the other baseline strategies. In the section that follows we consider

the extent to which we can achieve a comparable level of accuracy by "resolving" as many of the ALU labels as possible without human intervention.

We also notice in Figure 4.3b the corresponding F_1-measure, precision and recall curves for the passive strategies show that a precision as high as .90% is achievable, without user involvement in the labeling training instances. Consistent with the results in the previous chapter for short sentential text. This approach to limited supervision is, in an of itself, quite a promising result for EI system designers. We also notice however, that the passive strategy is recall gated and we never achieve a recall more than .55%. On the other hand, the situation is quite different for the active learning selecting strategies. We notice that if we are able to obtain the true labels from the most uncertain instances, then we are able not only achieve a much higher overall accuracy, but also a high precision (90% for ALU) as well. In the remainder of these experiments we consider the extent to which we can use and active learning with uncertainty sampling (ALU) and reduce the number of queries that we must present to the oracle for their resolution.

Ngram Features

Since Twitter message are limited in length by 140 characters, the feature vectors representing them are very sparse. In this experiment, we are interested in knowing which features lead to a better classifier. Intuitively, we one would like to enrich the feature vectors of tweets, with a more robust set of features. Particularly for cases where the target concept can be correctly labeled with sequences of n words. This is useful to capture the occurrence of phrases that consist of multiple words, e.g., "bieber fever" or "ice-cream headache".

As shown in Figure 4.4, our results show that when using uncertainty sampling, we are less sensitive to the type of features used. Using 2-grams did not achieve improved performance over using 1-grams; and in fact, adding n-grams with $n > 2$ decreases the performance of the classifier. For this reason we opted to using 1-grams.

4.4.5 Results II: Classifier Performance and Costs

In these set of experiments, we use the baseline classifier from Part I, and seek to determine how good a classifier we can build by automatically resolving the labels of uncertain samples using label-aligned clusters LAC for which background topics have been eliminated and the topics, used as features to retrain a secondary classifier. We compare the LAC against a baseline strategy $MuxA$ in which global clusters are using and mutual exclusion is enforced.We assume that batch of uncertain samples (and corresponding seeds) has been produced apriori, by aggregating the set of uncertain samples that were selected at each iteration.

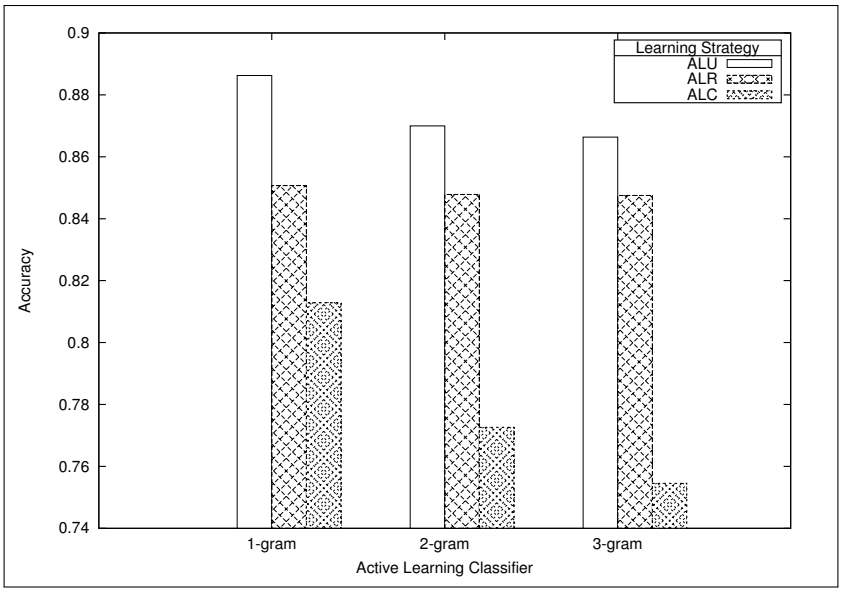
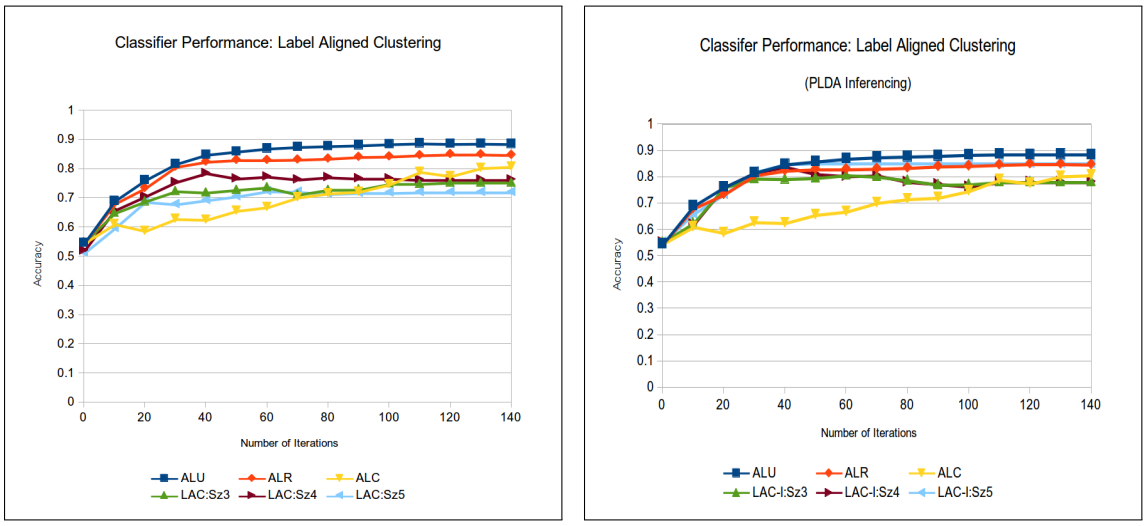


Figure 4.4 Average accuracy for Active and Passive Learners.

Classifier Performance Figure 4.5a shows the classifier performance for Topic Driven Resolution, which uses Label-Aligned Clustering (LAC) topic features, and background topic elimination. We show the results for samples sizes, $beta = 3, 4, 5$: denoted as LAC:Sz3, LAC:Sz4, and LAC:Sz5, respectively.

We first notice that although *ALU* and *ALR* strategies outperform the *LAC* strategy, the difference in performance is not so great, considering that *ALU* and *ALR* learning strategies are fully supervised. It is also interesting to note that all the of *LAC* strategies outperform *ALC*, up to a point. *ALC* overtakes the LAC:Sz3 and LAC:Sz5 after 70 iterations. With respect to *ALC*, the *LAC* strategy is already providing to a cost of: 420 queries (3 x 2 x 70) for sample size of 3; and a cost of 700 queries (5 x 2 x 70) for sample size of 5. For sample size of 4 *ALC* does not begin to overtake the LAC:Sz4 until roughly 110 iterations. These results suggest that in the absence of obtaining true labels from and oracle an intermediate strategy of semisupervised clustering such as PLDA is a reasonable option for obtaining labels at least during the initial stages of the training (up to about iteration 40 for where the LAC:Sz4 peaks). Also, we find that for the first iterations of the bootstrap a warm up period is needed for all strategies until the number of training instances reaches an amount to build a classifier that does more than purely guessing (accuracy exceeds .50% which is around iteration 10).

Figure 4.5b shows the classifier performance for Inference Driven Resolution in which, a bootstrapped purely uses inferencing to resolve the labels of uncertain samples. When only the PLDA inferencing is used, this is actually equivalent to a zero cost strategy, since all the uncertain labels are resolved automatically. The results are rather encouraging. For much of the early iterations (up to 40 iterations) the



(a) Topic Driven Resolution **(b)** Inference Driven Resolution

Figure 4.5 Learning curves for PLDA Topic Driven Resolution in which topics are used as classification features (4.5a). PLDA Inference Driven Resolution: learning curves for Label-Aligned Clustering using only PLDA inferencing (zero cost) label resolve (4.5b).

performance is actually competitive with *ALU* and *ALR*. Only after around 80 iterations do we start to see the pure inferencing strategy degrade and level off. This degradation is due to the fact that at some point, as the size of the training set grows, we are no longer capturing an appropriate number of topics per class and background to model the growing size of the seeds. Also, if poor inferences are made, they start to impact the classifier performance over time. Nonetheless, the degradation experienced here is still acceptable, and does not worse than the Topic Driven Resolution at larger iterations.

Cost Savings Figures 4.6a and 4.6b shows the cost associated using *ALU*, *ALR* and *ALC* and *LAC*, respectively. We actually do save using *LAC* as a selection strategy and fundamental reason is that unlike the basic active learning strategies (ALU, ALR and ALC) which have fixed cost, the *LAC*has a variable cost. Specifically, if assume that unity represents a cost equivalent to requesting labels for each instance, then the results in Figure 4.6b show the percent of the total that we would actually request per iterations. Since we only intermittently request labels from the oracle the cost percentage curve exhibits a step function. In comparison, the basic strategies, which follow a nearly linear shape for their costs.

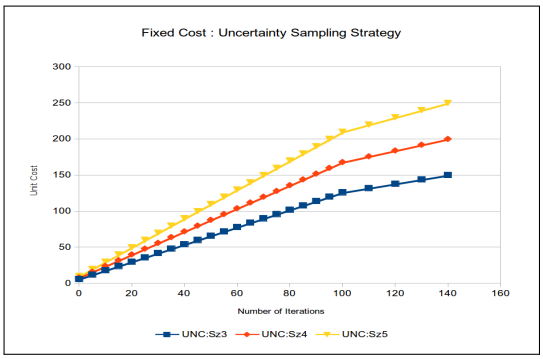
Figures 4.6c, 4.6c,4.6d,4.6e, and 4.6f shows the cost associated with a mutual exclusion aware (Mux-A) strategy. One can notice that in comparison to the cost of the *LAC* strategy the percentage of cost for global labeling scheme is much higher.

The main reason for this is that although we are able to label instances with a reasonable level of confidence, many uncertain instances can be covered by the mutual exclusion criteria - which allows us to automatically assign the label in the first place.

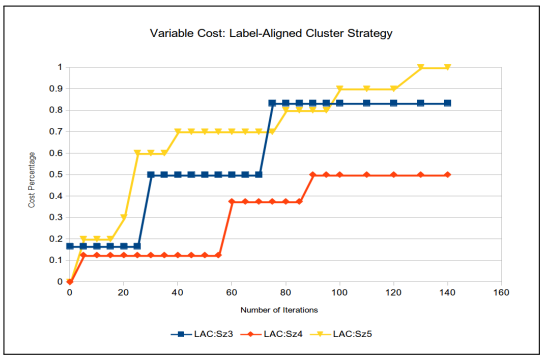
Figure 4.7 shows the average percentage of uncertain instances that are not able to be resolved with the mutual exclusive aware strategy per cluster size and threshold. As as, we notice a clear tradeoff between the percentage of documents that we can assign to a mutual exclusive cluster and the confidence we have that a document belongs to a topic. The high the confidence level we require, the more documents remain unassigned that we must eventually offload to an oracle, hence the higher the cost of the the sampling strategy. This suggests that there is some benefit to using a principled approach to modeling the background topics with the PLDA as opposed to enforcing mutual exclusion among a seeds of a global topic model.

Mux-Aware Labeling Quality Figure 4.8 shows the average hit rate among seeds and uncertain instances compared with true labels for a mutual exclusion-aware clustering strategy, i.e., one that using global topics and enforces mutual exclusion per iteration (the local level). We notice that independent of the cluster size chosen, we are able to achieve an average hit rate of at least 70%. Hit rates at 80 % and above are obtained using threshold values in the range of .55 to .75. We believe is due to the fact that this range represents the densest part of the threshold distributions. Many probabilities values beyond .75 % although extremely confident are few in number.

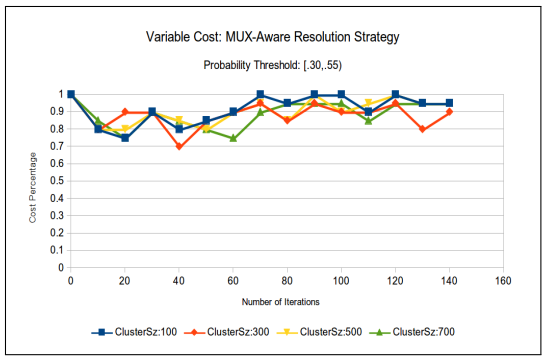
Mux-Aware Classifier Performance Finally, when seek to determine how the bootstrapped classifier would perform using the Mutual-Exclusion Aware topics from a global clustering. Figure 4.9 shows the at each of 140 iterations of the bootstrap the classifier either toggles back and forth between pure guessing and perfect accuracy! Such a learning is not reliable since it uses so few instances to train its a matter of chance if we find the seeds that happen to lead to a highly accuracy result on the test set. These results lead us to conclude that a global strategy is an inexpensive way to obtain a few reliable labels with limited supervision, but when more labels are needed a LAC is a viable option. Even though the quality of LAC labels diminishes, we are able to build a stable classifier during the early stages of the training. When even more training data is needed to build the classifier the LAC becomes less reliable and mixture of global clustering in conjunction with labels from an oracle are then needed. We limited the reclassification to the set of seeds for the current iteration, we believe that improved results can be obtained by using not only the seeds for the current iteration, but a random subset of past previously generated seed when training the PLDA.



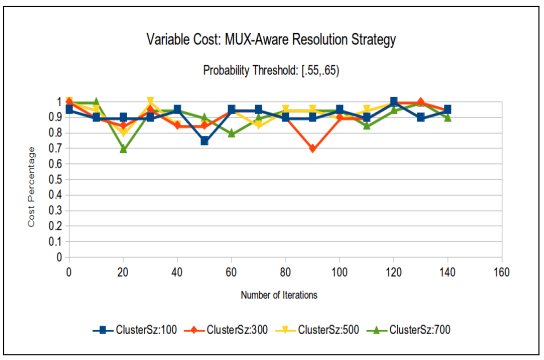
(a) Cost: UNC Sampling



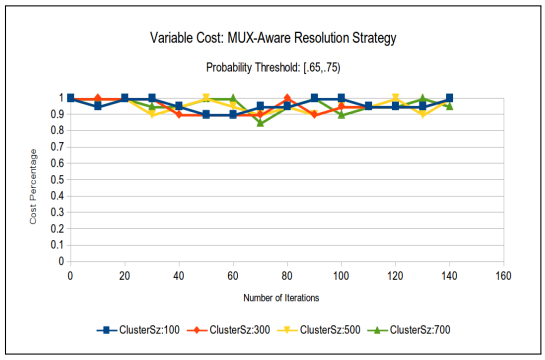
(b) Cost: LAC Sampling



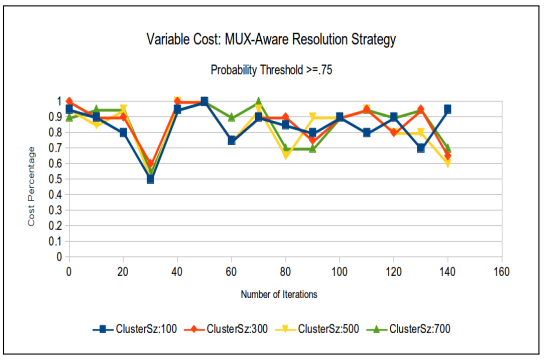
(c) Cost: Mux-A, α =[.30..55)



(d) Cost: Mux-A, α =[.55..65)



(e) Cost: Mux-A, α =[.65..75)



(f) Cost: Mux-A, $\alpha \geq .75$

Figure 4.6 Fixed cost of basic uncertainty sampling strategy (4.6a); variable cost of using a strategy based on Labeling-Aligned Clustering (4.6b). Cost of global clustering with Mutual Exclusive-Aware strategy, for threshold probability,α =, for: α =[.30..55) (4.6c); α =[.55..65)] (4.6d); α = [.65..75)] (4.6e); and ($\alpha \geq .75$) (4.6f).

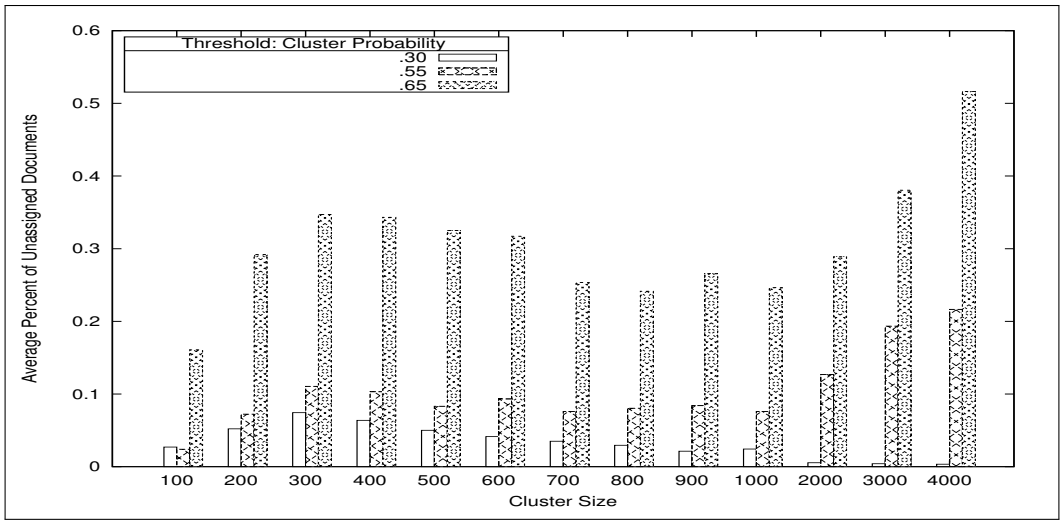


Figure 4.7 Average percentage of documents that remain unresolved after enforcing mutual exclusion among seed instances.

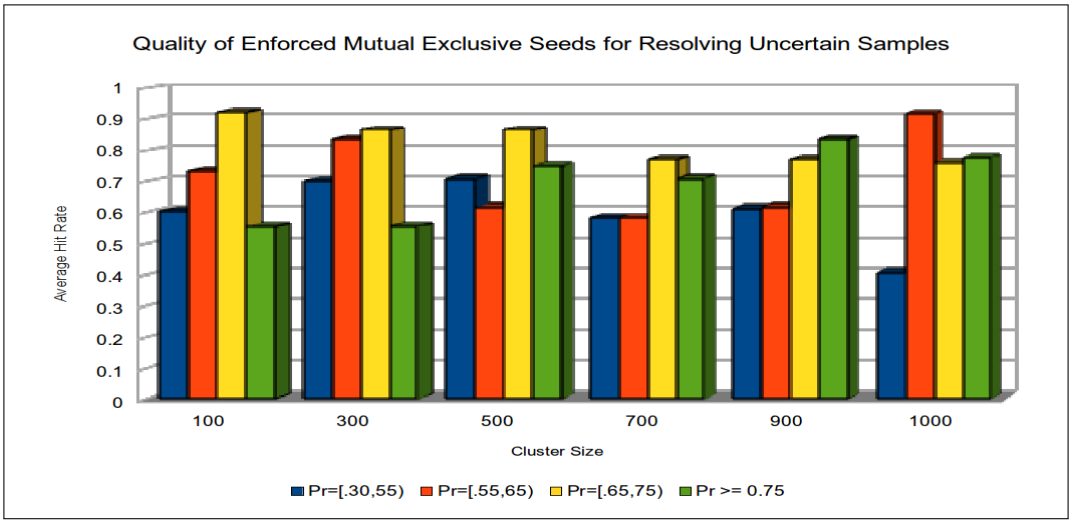


Figure 4.8 Average Hit Rate among seeds and uncertain instances showing the Quality of Mutual Exclusive Seeds when using global topics and enforcing mutual exclusion at the local level.

4.4.6 Results III: Expert Assessment and Interpretation

In this chapter we relied upon an oracle to safely handle those instances that we could not label with confidence in an automatic way. Up to now we assumed this oracle was a subject-matter expert. What if we outsourced the task of labeling to a non-expert or human intelligence worker, could we trust the labels they provide, for our task? In this case study we examine the answer to this question by evaluating the quality of labels obtained by crowdsourcing again the labels assigned by a domain expert.

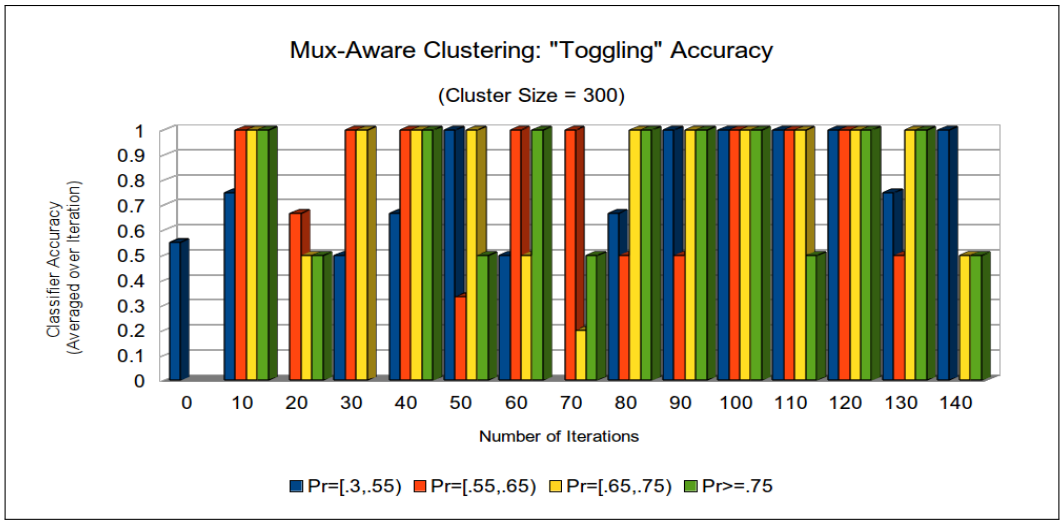


Figure 4.9 Classifier accuracy among seeds and uncertain instances when using global topics and enforcing mutual exclusion at the local level. An unreliable classifier is obtained since at each of 140 iterations of the bootstrap, the classifier toggles back and forth between pure guessing and perfect accuracy.

Crowdsourcing versus Experts. To build a crowdsource set, a total of 1,500 tweets was created by randomly sampling 500 tweets from each of the calendar weeks 15, 16 and 17, and presented to workers of the CrowdFlower platform. Judgments given by untrusted workers were eliminated by computing a trust score for each worker, based on the their agreement with a gold standard set of tweets. The agreement among the 43 HIT workers (minimum of 3 workers for each task) was 93.89%. We chose 130 tweets according to the agreement of the crowdsourcing annotators to include a representative amount of "easy" tweets (perfect agreement) and more "difficult" tweets (low agreement). The agreement between five domain experts was 89.33%. To control the quality of the crowd sourced labels, two actions are taken. First, a set of "golden" tweets with known labels is added to the unlabeled tweets that are randomly shown to the workers. A trust value is computed for each worker based on the number of correctly labeled "golden" tweets. If this trust value is below a fixed threshold, the workers labels are removed from the task. Second, each tweet was labeled by multiple workers. Then, the tweet is assigned to that category that received the most votes weighted by the workers' trust values. The agreement of multiple annotators for the label of a tweet is measured by the percentage of annotators that agreed on a tweet relative to the total amount of annotations on the tweet. The agreement on the crowd sourced data sets was 93.89% with an average accuracy on the injected gold labeled tweets of 92%. Out of the 1,500 tweets labeled by the crowd: 1,114 tweets had a perfect agreement of 100 295 tweets had an agreement between 66% and 100%; and 91 tweets had an agreement between 50% and 66%.

For the 130 tweets labeled by both the public health experts and the HIT workers, there was a percent agreement of 87.69%. When measuring the classifier performance individually for each group based on a 10-fold cross validation and equal percentages of positive and negative examples, the classifier performance was 75% for the experts and 83% for the HIT workers.

Our results suggest that people outside the public health domain are able to accurately judge the relevance of tweets, when given a simple set of criteria. Thus, once M-eco has been detected feature change, it is also feasible to outsource the novel tweets, as part of a separate feature change handling procedure in an *Active Learning* setting. It took roughly 6 hours for the CrowdFlower workers to label the set of 1,500 tweets and the entire M-eco pipeline runs around the clock, restarting roughly every 4 hours, depending on the amount of incoming tweets. Although the classifier performance was much less for the expert labels, than the crowd labeled data, we believe this is due to the fact that in practice, when a tweet is relevant for an expert depends on several factors as outlined in our previous case study in Section 3.5.9. In our case, relevance depends on different time periods of an outbreak, (before, during or after); or on the task and role of the expert with respect to an epidemic investigation. Nonetheless, the crowd can still serve to help filter label instances that are clearly off topic.

To get a better understanding of the impact of detected feature change on the classification accuracy, a larger set of expert labeled tweets for experimentation would be useful to further improve the significance of the results. Nonetheless, doing so, would still not address the need to experts to relabel each time feature change was detected and in practice, the overhead of such a task is too expensive and not timely enough. We propose instead, that after tuning expert labeled examples with a good inter-annotator agreement, expert labels examples could be used as gold standard to filter out HIT workers who trust value is too low.

The fact that our classifier performed better on the labels provided by the HIT workers suggests that there are advantages in being able to use more workers than expert labelers. Importantly, being able to build in a trust value that allowed us to filter instances and/or workers, can greatly help with dynamically labeling and maintaining long-term tweet classification accuracy. Such considerations are even more crucial for more complex criteria, for which the crowdsource performance can greatly diminish.

Table 4.5 lists the 16 tweets where experts and CrowdFlower workers disagree. On those 16 tweets, the average crowdsourcing agreement is 83.65% and the average agreement of the experts is 72.37%.

4.4.7 Discussion

In this chapter, we propose a label resolution strategy for semi-supervised active learner, which aims to reduce the labeling cost for an active learner while simulta-

Table 4.5 Tweets with different expert (E) and crowdsourcing (C) labels

Tweet	Label		Agreement	
	E	C	E	C
How you can inadvertently make a headache worse http://dld.bz/gCqr	-	+	50%	64%
Yay the headache is gone! Way to go, bed.	+	-	73%	66%
3 hours of bell ringing? I'd have such a headache by the end.. haha	-	+	91%	66%
Lmfao I just caught a headache laffn at dis message from @STRONGDEUCE...@BOBdaBuildER3	-	+	73%	66%
Doing my income taxes gave me such a migraine headache, that it kept me from going to work today. We need tax reform with the Fair Tax.	-	+	50%	70%
Want a cough drop?	+	-	63%	71%
I'm so thankful when I wake up headache-free!! Happy Tuesday, it's going to be a busy one! #kidmin	+	-	55%	74%
lol! @SILENT_CHAOS706 it's hunger headache ; not hungry headache ; lol!	-	+	90%	78%
@Phil_KnowsBest aha dam i gotta headache. ma mom is buggn me.	-	+	55%	84%
doing d bad thing just make some headache...argghhh but i cant help it,,	-	+	50%	100%
I accidentally took nighttime cough medicine during the day and slept through half of it.	-	+	60%	100%
@HeartChloe Its been a headache lately, hopefully here too lol u get some green tonight?	-	+	60%	100%
So tonight should be interesting at work. Trainee + Internet issues = headache. — RT @StrangeHand	-	+	90%	100%
@SexGodKatherine Killing Klaus sounds like a plan. *smirks* I just want Elena gone.She gives me a headache *rubs my head*	-	+	100%	100%
If u use my microwave R Urs n stop it early but Don't clear it I get So PISSED watching Hoarders gives me a headache n shortness of breath	-	+	100%	100%
@felee92 Ooooh cool!!! So many people are changing their DPs to sideways like jaejoong also, give me headache only @__@	-	+	100%	100%

neously, seeking to minimize the side-effect of more mislabeling when automatically attempting resolving uncertain with additional context from unsupervised probabilistic clustering. Also, we provide a novel approach to learning using Partially Labeled

Dirichlet Allocation (PLDA). PLDA was to address the problem of ambiguity (non-separability) that arise from seeds instances with overlapping context.

A cost reduction is obtained in pro-actively selecting instances and creating a mutual exclusive context when re-classifying dubious instances (Topic Driven Resolution) and using pure inferencing (Inference Driven Resolve) afforded by the PLDA model. In both cases, we were able to minimizing human annotation effort, when compared with a fixed cost strategy. Thus allows a human annotator to focus on those instances which are most difficult or require world knowledge or additional features that are not part of the current training instances.

In these experiments, we limited the type of features to 1-gram, since they showed better performance over 2-and 3-grams. Still other types of features could be considered. Given the advent of reliable of parts-speech-taggers for Tweets (Ark-Tweet) and the speed with which its performs, it is feasible to include POS tags as features; and we considered this in our future work. For selecting the number of topics per class and the number of background topics we used a simple heuristic based on the proportion of documents in each class. We also restricted the number of topics per document to 1. Experiments with three topics, showed poorer results when compared to using a single topic; more experiments should be undertaken to decide for the optimal set of parameter in each case.

Can active learning be effectively combined with semi-supervised and unsupervised learning? It is known from the rich area of semi-supervised learning that unlabeled data can suggest learning biases (e.g., large margin separators, low dimensional structure) that may improve performance over supervised learning, especially when labeled data are few. When these biases are not aligned with reality, however, performance can be significantly degraded; this is a common but serious criticism of semi-supervised learning. A basic observation is that active learning provides the opportunity to validate or refute these biases using label queries, and also to subsequently revise them. Thus, it seems that active learners ought to be able to pursue learning biases much more aggressively than passive learners.

Another aspect not considered in this work is the use of multi-view active learning. Given the fact that we have can resolve labels either by PLDA inferencing or classifier retraining with PLDA topics, the question arises as to whether these judgments could be combined to improve the perform; we consider this in our future work.

4.5 Chapter Summary and Outlook

In this chapters we explored semisupervised learning to detecting disease reporting mentions in short and sparse text. The basic technique we explored to tackling the label bottleneck was bootstrapping. That is, an initial set of seeds where provided and more labeled were obtained through from a classifier that labeled its own examples.

A growing number of researchers are feasibly using *Human Intelligence Tasks*

(HITs) or crowdsourcing to improve the performance of automated systems [DDCM12]. In a similar work, Paul and Dredze [PD11b], also created an annotated Twitter data set using Mechanical Turk for an EI task similar to the one described in this work, but they do not consider HIT workers in the context of an feedback loop for maintaining the accuracy of their system. We believe that the results presented in this case study show the potential and feasibility of employing human intelligence workers to handle budgeted labeling under limited supervision learning.

In the final chapter on limited supervision approaches to detecting disease reporting mentions, we explore a much different variety of learning in which no labeled training data at all is required: the unsupervised learner.

Unsupervised Dection of Disease Reporting Mentions

[1]

In Chapters 3 and 4 we respectively tackled the label bottleneck problem using forms of weak labeling, and budgeted labeling. In both cases, the classifier algorithm was based on semi-supervised learning (SSL). Using SSL, only a small amount of labels were given for the task and the classifier incrementally built its own training instances through repeated training-inferencing cycles. If we choose to by-pass any form of supervised learning altogether and opt for an unsupervised approach to

[1] Image under License from Fotalia http://de.fotolia.com/

learning, for which no labeled data is required, can we then discover patterns that are representative of disease reporting mentions, which are meaniful to an investigator?

In this chapter we explore the answer to the aforementioned question. Generative models for clustering are a popular tool for the unsupervised analysis of text, and provides a latent topic representation of the corpus. It is used to check models, summarize the corpus, and guide exploration of its contents. Notable, with the exception of work done by [PD11b], no other work has been done in using generative models in EI. In addition to the fact that an oracle need not provide labels, another advantage of an unsupervised approach is that it has the potential of detecting disease related events that were not explicitly under surveillance. This serendipity can be useful for early detection systems if an emerging disease has no known name, or can only be characterized by symptoms.

Although more generic and flexible, an unsupervised approach to detecting disease reporting mentions using generative models can lead to very complex results. This complexity poses a significant challenge for an epidemic investigator, given the number of potential clusters (or latent topics). Additionally, since the pattern is not labeled apriori, the significance and meaning of the pattern must be interpreted, a common problem [BGCG+09]. In order to ensure that the unsupervised methods produce results that are valuable for the human users, it is crucial that EI systems consider a user-centric approach which emphasizes both: an assessment of the cluster quality, and their representations - which is the focus of this work.

The problems to be addressed are: 1) detecting patterns that are meaningful as epidemic events within EI, in an unsupervised manner; 2) presenting the detected patterns to a human effectively; and thereby allowing domain experts to easily interpret any epidemic patterns that are mined; and 3) enabling domain experts to analyze the epidemic intelligence data. To address these problems, we present a novel framework with which field practitioners can interact with the underlying data and assess the results of complex unsupervised algorithms. The **user-centric** pattern assessment framework constitutes: *pattern mining*, *pattern pruning* and a user-centric *pattern evaluation* involving field practitioners. A key aspect of the framework is a feedback interaction loop which involves; tuning a system to better help these users in their epidemic investigation activities.

More specifically, we address the following questions:

1. How can we characterize a Disease Reporting Mention (DRM) as a cluster?
2. How do we measure the quality of Disease Reporting Mention Clusters?
3. When is one DRM cluster better than another?
4. How meaniful are DRM clusters and their word-cloud representations for domain experts?

We first present related work is presented in Section 5.1. In Section 5.2, we present our Field Practitioner-Assisted Assessment framework. Then, in Section 5.3, experimental results are provided. In Section 5.4 we conclude, and provide directions for future work.

5.1 Related Work

5.1.1 Rule-Based Systems

In the health domain, the most prevalent approaches for detecting public health events are rule-based systems [SFvdG+08a, SFvdG+08b, CKJ+06, LNGB12, AHB+93, GHY02, Yan06, KBHT09]. A rule is a conditional of the form: *contextual pattern* → *action*. If the contextual pattern matches an input text, then the action part of the rule fires. A contextual pattern is intended to describe the context in which a pair of entities appear. Pattern matching techniques are used to capture the lexical (surface) structure of a sentence; and are most commonly represented by regular expression. The action part of the rule is used to denote various kinds of tagging or replacement actions such as: assigning an entity type to a sequence of tokens.

Proteus-BIO: Proteus-BIO [GHY02], is a cascaded set of finite-state transducers [AHB+93]. A set of task-specific rules are constructed to capture the possible patterns that are used to express an outbreak. For example, the pattern *[disease killed victim]* would match the string "cholera killed 7 inhabitants". When a pattern matches a string within the text, the result is consider to be a public health event. To simplify the pattern construction process, preprocessing steps are employed to first identify the major linguistic constituents such as: names and dates; noun and verb groups; and noun phrases. The pattern base is then built using these constructs.

MedISys / PULS MedISys is an example of a rule-based system [SFvdG+08a]. The work done in the PULS system [Yan06] also uses a pattern-based approach to identify the disease and the location of a reported event. PULS is integrated into MedISys. Pattern matching includes the use of keywords and string deformations rules to capture the morphorlogical variations of a term in the text. An example rule used by MedISys is denoted in the Table 5.1.

BioCaster: Another prevalent rule based system is BioCaster [CDK+08]. BioCaster rules are performed by using regular expressions with the Simple Rule Language (SRL) parser and is capable of matching co-occurring entity mentions. Example SRL named entity rules that defines two instances of a $DISEASE$ event are given in Table 5.2.

Table 5.1 Example rules for detecting the mention of Avian Influenza Medical Condition from unstructured text. In the table, a '+' symbols represents any amount of white space; a '%' represents zero or more characters; a '_' represents exactly one character.

Pattern	**Example String Matches**
avian+flu	avian flu
avian+influenza	avian influenza
bird+flu	bird flu
bird+influenza	bird influenza
gripa+porcin_	gripa porcina, gripa porcină
schweinegrippe%	schweinegrippe-h1n1
svinjsk%+grip%	svinjski grip!

Table 5.2 Example SRL rules and the patterns that are matched. The SRL construct "words(,1)" defines at least one white space. The SRL variables `birdName` and `fluName` define two sets corresponding, respectively, to a set of bird names and a set of flu names, for example: %birdName=("avian","bird") and %fluName=("flu","influenza").

Pattern *DISEASE*	**Example String Matches**
gripa words(,1) porcin?	gripa porcina
list(%birdName) list(%fluName)	avian flu, avian influenza, bird flu, bird influenza

Using SRL, simple inferred relationships can be modeled. For example, in Table 5.2, the SRL expression `list(%birdName) list(%fluName)` would tag any phrase that contained a word from the set `birdName` followed by any word from the set `fluName`, as a `DISEASE` event. More patterns are obtainable through syntactic generalization, for example, by specifying base forms of verb. As one can note from the examples in both Tables 5.1 and 5.2, all variations on an entity name, including its representation in different languages, can be modeled.

In general, the advantages of rule based systems are that they can be updated on-the-fly, for example, to include a previously unknown disease name [CDK+08]. One of the major drawbacks of rule-based approaches is in building (and maintaining) the pattern base. Also, data on the Web is not clean, and in the presence of "dirty" data, rule-based systems do not work well. Additionally, it is costly and labor intensive to build a pattern base for rule systems: writing a few rules is easy, but writing lots of rules to capture all contexts is hard. Even if rules are learned, at some point additional rules actually hurt the performance of the system. The optimizations of the system are difficult to manage, since the order in which rules were applied impacts the performance. Also recall tends to be low if a large enough pattern base is not constructed when relying upon and "all enumeration" approach to pattern detection.

We seek to go beyond these limitations by considering an unsupervised approach to public health event detection and compare our work to the well established rule-based system of MedISys.

5.1.2 Supervised and Unsupervised Systems

Numerous supervised classifiers exist for detecting public health events within unstructured text [CCD09, KBHT09, Zha08]. A limitation however is that they all also use manually labeled data to build their models. Although automatic labeling is exploited in the work of Stewart et al. [SN11a], this approach has some limitations since the full sentence parsing techniques is expensive, when it comes to extracting parse tree classifier features.

Perhaps the most similar EI work to our unstructured approach is that of Paul et al. which uses symptoms and treatments to define health related topics in an unsupervised manner (Ailment Topic Aspect Model). Using messages that have been labeled for relevance with regard to health, as well as a generative model, Paul et al. actually bridge the gap between a fully supervised and unsupervised approach [PD11b, Dre12].

Also in the work of Nowcasting [LC12], the authors analyze correlated burst behavior of tokens to infer the existence of public health events.

However, many types of features that are relevant for public health event detection (i.e. symptoms, victims, or medical conditions) are not modeled in the Nowcasting system. Also, unlike previous work, we focus on assessing the results of using a generative model for detecting disease reporting clusters from the perspective of the

domain experts, which is largely ignored in all related works in EI.

5.2 Field Practitioner-Assisted Assessment

The underlying intuition behind our Field Practitioner-Assisted Assessment approach is that with an appropriate feedback interaction loop, we can gain a better understanding of how end users in the domain of epidemic intelligence assess the patterns that have been detected in an unsupervised manner. Further, mechanisms can be built-in to help chart the success of the applications that have been built with unsupervised event detection. In this section, we present our framework for the Practitioner-Assisted assesment of unsupervised disease reporting events. An overview of the framework is depicted in Figure 5.1 and each stage is presented in the discussion that follows.

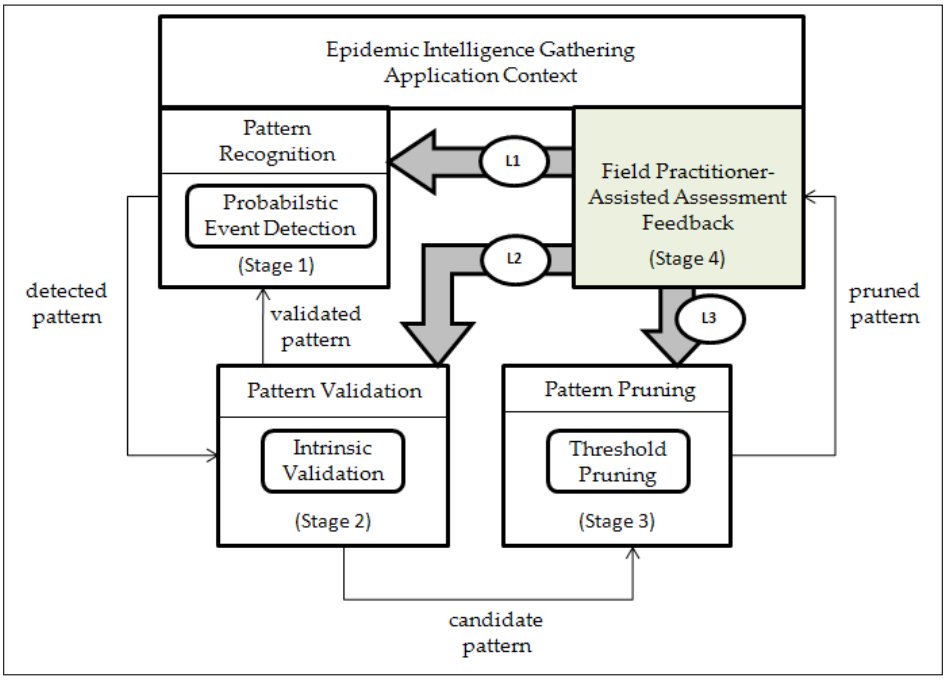


Figure 5.1 An overview of the Field Practitioner-Assisted Assessment Framework.

5.2.1 Pattern Recognition

Starting with a finite set of text articles, $\mathcal{A}$, we process the raw text; transforming each article into a surrogate, vector format that consists of only relevant terms. The relevance of a term is determined by the extraction of entities according to a type system, $\mathcal{T}$. The set of allowable EI entity types we consider are: Location, Medical Condition,

and Victim. Each entry in the vector representation corresponds to the frequency with which an entity, of the given type, appears in the bag-of-words representation of the article. The document surrogates for the set of articles, $D_{\mathcal{A}} := [|\mathcal{A}|][|T|][|V_T|]$ is thus finally created from the vectors, V_T.

How can we characterize a Disease Reporting Mention (DRM) as a cluster? We consider an unsupervised disease reporting event, to be a type of pattern that is recognizable by an event detection algorithm. In our framework, a `Probabilistic Event Detection` algorithm is used. The type of pattern we consider is a (probabilistic) clustering of documents; where each cluster represents and event. We define an event, more formally as:

Definition 2 *Unsupervised Event (Cluster): An unsupervised event, $\mathcal{I}_j$, is considered an unlabeled class, c_j, that represents a clustering of documents, which has been detected by an probabilistic clustering algorithm, Φ_K. The algorithm produces a weight matrix, $w_{ij} \in W_d$, such that each document d_i, where $i = \{1,...N\}$, is assigned to every unlabeled class, c_j, where $j = \{1,..,K\}$, with some weight $w_{ij} \geq 0$; where N is the total number of documents, and K is the total number of or events (i.e., patterns or clusters).*

Numerous techniques exist for detecting events in an unsupervised way. In this work, we base our unsupervised event detection algorithm on the Retrospective Event Detection [LWLM05] algorithm of Li et al. This algorithm for event detection, provides a framework for handling the multiple entity types; and we extend it to handle those defined by our EI-entity types and refer to the resulting clusterings as Disease Reporting Mention Cluster ($DRMC$). In general, the approach assumes that a document contains the textual mention of real-world, temporal events. Events are considered to be hidden variables whose likelihood is inferred by a generative model from the vector representation of the article's observable content (distribution of its features). The generative model produces events using Multinomial distributions over features of multiple EI entity types. The articles of the collection are clustered by relying on the iterative EM algorithm.

For more details regarding the use of $DRMC$ in the domain of public health, including algorithm tuning; feature pruning; the selection of the number of events; as well as other real-world case studies evaluating RED, the reader is referred to the work of Fisichella et al. [FSDN10, FSCD11].

5.2.2 Pattern Validation

`Pattern Validation` allows the different parameters of a pattern recognition algorithm to be iteratively tuned. For example, in unsupervised event detection, the number of classes, K, is required to be given as input. A value for K can determined using one of several intrinsic cluster validation metrics [HBV10, NJT07]. Different

types of validations, other than the choice of K, may be required. For example, using precision and recall, the quality of the generated clusters, or Reference set (Ref), can be assessed with respect to a manual clustering, or Response set (Res) [BB98] as followings:

$$Precision(Ref, Res) = \frac{\sum_{c_i \in C_{Res}} |c_i| - overlap(c_i, Ref)}{\sum_{c_i \in C_Res} |c_i| - 1} \quad (5.1)$$

$$Recall(Ref, Res) = \frac{\sum_{c_i \in C_{Ref}} |c_i| - overlap(c_i, Res)}{\sum_{c_i \in C_Ref} |c_i| - 1} \quad (5.2)$$

5.2.3 Pattern Pruning

When is one *DRMC* better than another? As noted in Definition 2, many patterns may be produced, since the event detection algorithm associates each document with every class, with some weight. In practice, associating every documents to a class will not provide meaningful results for everyday tasks, so some of the clusters produced by the `Pattern Validation` stage, are eliminated before being presented to the user. Clusters are pruned at three levels of granularity: term-level and document-level (intra-pattern); and cluster level (intra-pattern). The definitions for inter- and intra- pattern pruning are given in Definitions 3 and 4, respectively.

Definition 3 *Inter-Pattern Pruning: A pattern, I_j, is pruned if $Quality(\mathcal{I}_j) \leq \alpha$, for some given quality measure; and $0 \leq \alpha \leq 1.0$.*

Definition 4 *Intra-Pattern Pruning: Given a weight matrix, W_d, and a cluster, $\mathcal{I}_j$; a document d_i , is pruned if $w_{ij} \leq \beta$. Likewise, a term is pruned from the cluster if $w_{jk} \leq \gamma$, where $w_{jk} \in W_t$; $0 \leq \beta \leq 1.0$; and $0 \leq \gamma \leq 1.0$;*

In Definitions 3 and 4, the values of α, β and γ represent the cluster, document, and term threshold probabilities. The $Quality(\mathcal{I}_j)$, and values for α, β and γ are chosen according to task and algorithm-specific criteria. In our experiments we select the threshold probabilities for α, β and γ based on the quartile distributions [Lan06]. The quartiles of a ranked set of probability values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data.

5.2.4 Practitioner-Assisted Feedback

The `Field Practitioner-Assisted Assessment` stage takes pruned patterns as input for the user to assess their quality. The feedback interaction loops (denoted by

$L1$, $L2$, $L3$ in Figure 5.1) signify that: the features of the algorithm in the `Pattern Recognition` stage ($L1$), the validation technique in the `Pattern Validation` stage ($L2$), or the pruning criteria in the `Pattern Pruning` stage ($L3$) are all subject to adaptation, based on feedback from the user assessment. This is intended to improve the quality of disease reporting event, as well as the documents and words that make up a $DRMC$. In the next section, we describe the experimental results when applying this framework.

5.3 Experiments

5.3.1 Experimental Goals

The objectives of the experiments is threefold. First, in Section 5.3.3, we address the question: *How do we measure the quality of Disease Reporting Mention Clusters?* We assess the quality of our $DRMC$ extrinsically, by comparing them against a set of clusters whose events have been extracted with a state-of-the-art, rule-based system. Our $DRMC$ have been detected with the Retrospective Event Detection (RED) of Li, et al. [LWLM05] that has been extended to use EI-Entity types.

Second, in Section 5.3.4, we address the question: **How meaniful are DRM clusters for domain experts?** We assess the quality of our $DRMC$ with respect to field practitioners, for a given pruning strategy, which eliminates a subset of $DRMC$ from consideration. The chosen pruning criteria allows us to gain a better insight into the question: **When is one set of $DRMC$ better than another?** Our pruning strategy is based on a choice for an intrinsic validation metric; and threshold values of α (cluster threshold value); β (document threshold value); and γ (term or feature threshold value). We also examine if different combination of weights properly group documents into logical clusters, and whether these clusters are meaningful to the users.

Third, in Section 5.3.5, we address the question: **How meaniful are $DRMC$ word-cloud representations for domain experts?** We are interested in knowing whether the representation of the patterns as word clouds are actually useful to the users and consider how $DRMC$ can be better represented for experts, based on their remarks.

5.3.2 Data Sets Used

In our experiments, two types of data sets were used: news and blogs.

News Data Set

The News Data Set was used for the experiments in Section 5.3.3 to make and extrinsic evaluation of our clusters against a set of clusters whose events have been previously filtered with a state-of-the-art, rule-based system. To build our document collection, we downloaded the web pages for each urls listed in source column of the PULS fact base [YvES08], for the period from, January 1 - December 31, 2009. Of the 2,587 documents collected, we used the 1,280 documents for which the PULS date column could be automatically represented as a timestamp. For the same time period, we also collected the cluster labels (medical condition-location pairs) present in the PULS fact base, to use as benchmark clusters. The benchmark system aggregates facts into the same group, or equivalence class, if they share the same medical condition and location. Based on this criteria, the clusters of the PULS fact base that we collected, yielded a total of 379 clusters of which we used those clusters that constrained at least 10 documents; this amounted to 70 clusters.

Blog Data Set

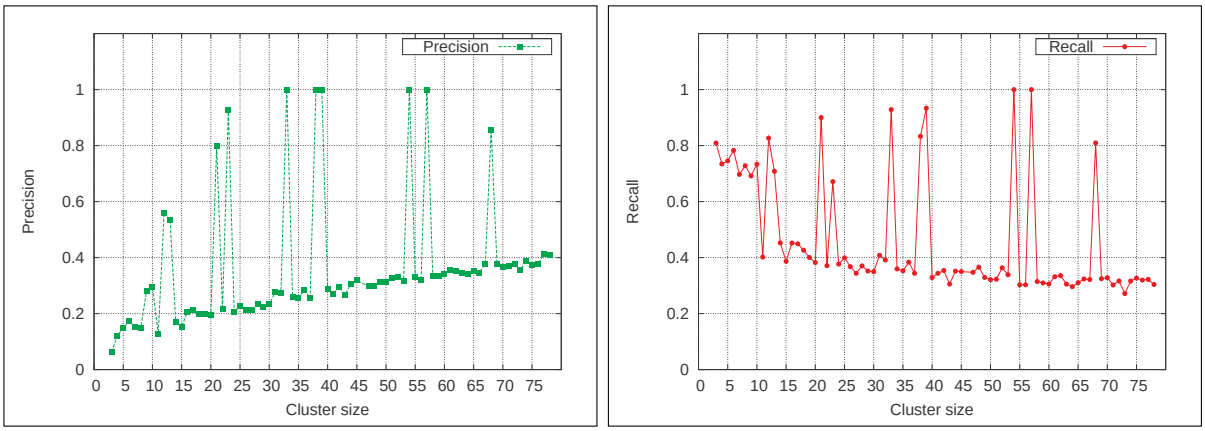
The Blog Data Set was used for the experiments in Sections 5.3.4 and 5.3.5 to assess the quality and representation of our clusters with respect to a field practitioner. The blog data set consists of medical blogs from MedWorm, a moderated blog medical blog aggregation, were collected via RSS during an eight month period from May 2009 through January 2010. This period is known to coincide with the 2009 Swine Flu pandemic. We used this corpus so that users could have some familiarity with the events detected, and used blogs so that the data set would contain noise from events that were not prefiltered by any algorithm, as in the case of our News Data Set.

In total, 30,822 documents were collected by retrieving documents from the sub-categories under the heading of *infectious disease*. Since the RSS text contained only summary short text, the urls from the RSS were used to crawl the website. The raw html was processed by stripping all boilerplate and markup code using the method introduced by Kohlschütter et.al. [KFN10]. A subset of the data collected and used for experimentation consisted of the 2,532 documents, that contained relevant EI entities types of medical condition and location.

5.3.3 Results I: Comparison with State-of-the-Art

The goal of this experiment was to compare the quality of $DRMC$ i.e., clusters that were discovered with our unsupervised, Retrospective Event Detection (RED) with EI-Entity types (the Response Cluster Set) for various cluster sizes, K, ranging from 1 to 70, against the 70 clusters of a state-of-the-art rule-based method, MedISys (the Reference Cluster Set). To accomplish the comparison among these clustering, we

used precision and recall computed according to Equations 5.1 and 5.2. Figure 5.2 shows the precision and recall for various clusters sizes.



(a) Precision over the computed clusters

(b) Recall over the computed clusters

Figure 5.2 Comparison of Precision and Recall for a document clustering based on Retrospective Event Detection (RED) with EI-Entity types (Response Cluster) against the Rule-based Event Detection Clustering of MedISys (Reference Cluster).

The first observation is that that there are several values of K for which precision and recall are above 0.8%, as denoted by the large spikes in the graph, over different values of K. Based on these values, we can see that a statistical $DRMC$ produces good quality clusters when compared with clusters that have been constructed from the inferred events of a hand-crafted rule-based detection system. Upon closer inspection, of these points we notice that when the precision or recall reaches a maximum value of 1.0%, the *victim* entity type made a much smaller contribution. Example events detected for the spikes are shown in Table 5.3.

The second observation we make is that overall, precision and recall are fairly low for most of the clusters, the majority of the values falling in the range of .02 to .04. This suggest that the some of the events detected differ significantly among the two approaches. This can be explained in part by the fact that the MedISys benchmark set used only two entity types (disease and location) for clustering, whereas we used three entity types (disease, location, and victim). Also the $DRMC$ uses a soft clustering, whereas MedISys uses hard clustering. In cases where the information about an outbreak is spread across multiple sentences, an event may also not be detected with the rule-based system. In constrast, $DRMC$ is more sensitive and forms clusters based on latent co-occurances, between entities that are not explicitly in the same sentence. Some example events for which the precision and recall are below 0.4 shown in Table 5.4.

Table 5.3 Example events detected using unsupervised event detection for which precision reaches a maximum value of 1. The columns respectively show: the extracted terms, number of documents and brief description of the real-world events.

Event Id	Event Terms	No. Docs	Event Description
E_8	manila, afghanistan, disease, infection, people, father	57	In October 2009 there was an outbreak of leptospirosis in Manila, the capital of Philippines. Further, there was a typhoon which led to an increase of several infectious diseases.
E_9	africa, united states, flu, disease, female, people	44	In December 2009 a deadly outbreak of cholera in north-western Kenya took place.
E_{10}	delhi, united states, swine flu, dengue, children, people	44	In November 2009 there was an outbreak of Dengue in Delhi.
E_{15}	surrey, london, infection, flu, children, animals	50	In September 2009, there was an outbreak of E. coli in England. Mainly children were concerned that visited the some farm in Surrey.

Upon observation, we notice that although these clusters represent events, statistically, they do not call for action on the part of any public health official. These type of events, which contain relevant mentions of EI-entities, but do not **semantically** represent a threat to public health are not possible to filter out automatically without additional criteria for defining the role of the entity. Two possible solutions to automatically filter such events are with: (1) additional feature analysis and pruning of terms apriori; or (2) constructing a hybrid Supervised-Unsupervised system in which irrelevant documents are first eliminated based on their relevance for EI, using techniques outlined in Chapters 3 and 4 and then applying $DRMC$.

We do not automatically tackle the semantic problem within $DRMC$. In the remaining sections we present $DRMC$ to domain experts for a semantic and qualitative assessment in order to better understand the extent to which these "noisy" documents impact the usefulness of $DRMC$ for domain experts. As we shall see, the use of meaniful visualizations, such as word clouds, is in practice, is one useful way to help experts better navigate to document clusters containing semantically relevant disease reporting mentions.

Table 5.4 Example events detected using unsupervised event detection for which precision have a value of 0.2%. The columns respectively show: the extracted terms, number of documents and brief description of the real-world events.

Event Id	Event Terms	No. Docs	Event Description
E_7	united states, russia, swine flu, disease, people, female	24	Several news articles provide comparison of swine flu statistics for various countries, comparing mainly cases happening in Europe, Russia and U.S.
E_{13}	nigeria, russia, disease, infection, children, people	31	Several studies found out that babies and children in Africa die from infections (September 2009). Further, there was a measles campaign in South Africa (October 2009).
E_{14}	japan, tokyo, disease, cholera, people, children	44	In October 2009, Japan started with swine flu vaccinations. Another event reported in the documents assigned to this cluster are outbreaks of Cholera in several parts of Africa.

5.3.4 Results II: Expert Assessment of Cluster Quality

In the previous section, we compared the clusters of the unsupervised event detection algorithm with the clusters obtained from a rule-based extraction system. In this section we present *DRMC* to experts in order to address the question of: **How meaniful are DRM clusters and their representations for domain experts?**

Pattern Validation: In order to determine the number of clusters to use as input for the algorithm, we run the event detection algorithm for values a $K = \{10...100\}$ and selected the K for which the cluster cohesion was the highest, this corresponded to a value of $K = 93$. Since the detection algorithm uses a random initialization, the results vary for each trial. We selected the best trial by running the algorithm for 100 trials and selected the trial having the highest log-likelihood.

Pattern Pruning: The 93 clusters, were pruned by using intra-patten and inter-patten pruning. The threshold values for pruning the clusters (α), documents (β), terms (γ) were selected based on a quartile distribution of their probabilities as described by Table 5.5.

In order to determine the impact of using quartiles as a pruning criteria, we selected two different quartile ranges for the documents and clusters. We used the

Table 5.5 Pruning Criteria based on a quartile probability distribution, where the first quartile (L) contains the lowest probability values and the fourth quartile (H) contains the highest probability values. HH contains cluster with the least noise; HL contains cluster having mixed noise; and LH contains clusters with the most noise.

Pruning Criteria	Description
HH	both the cluster and document probability belong to the fourth quartile range
HL	cluster probability belong to the fourth quartile range and the document probabilities belong to the first quartile range
LH	cluster probability belong to the first quartile range and the document probabilities belong to the fourth quartile range

notation L and H to correspond to the range of probabilities in the first quartile (low probability values), and the fourth quartile (high probability values), respectively. Using a combinations of these quartiles, three pruning criteria: HH, HL, and LH, were used to prune the clusters and documents. The ordering: HH, HL and LH, reflects the increasing noise, with respect to the probability values. All term probabilities were taken from the H quartile range and in total, 9 of the 93 clusters (3 clusters for each pruning criteria), were presented to the users for evaluation.

Experimental Setting

User Groups: Two user groups were asked to assess the quality of the 9 clusters, and their representation as word clouds. The first group consisted of five practitioners in the field of epidemiology. We label this group as the "Expert" group. The second group, as a basis for comparison, consisted of non-practitioners, five individuals with backgrounds in the area of user centered design; we label this group as the "Non-Expert". Figure 5.3 shows an example of two words clouds, and a document snippet that was presented to the users. Each cluster was presented to the users as a set of 5 documents; each document consisting of an ordered list of 4 *representative sentences*. A sentence was considered representative, if it contained a term, whose probability was in the fourth quartile range a preference was given to those sentences appearing towards the top of the document.

Table 5.6, shows each of the questions that was posed to the user and the order in which there were asked. For each of the nine clusters, they users were instructed to read the sentences for a set of documents within a single cluster, and then, complete each question.

Metrics Used: Three metrics were used in for obtaining the quantitative results. For the questions that required a user rating, a 5-point Likert scale and the percent agreement metrics were used.

Figure 5.3 Example words clouds, and a document snippet that was presented to the users during evaluation.

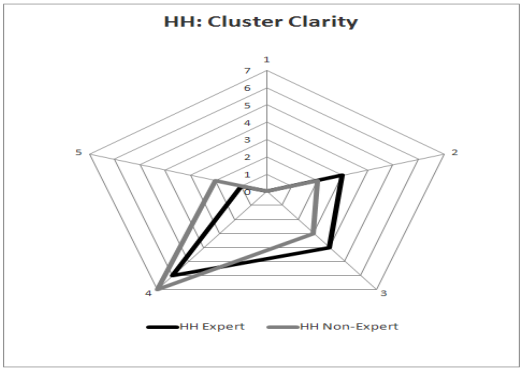
Table 5.6 Assessment Questions posed to the domain experts to assess the quality of the disease reporting clusters.

Id	Assessment Question
Q1:	Write 3 to 5 keywords that best describe the set of documents for the group
Q2:	Extent to which the set of documents for the group makes sense to you. Scale: 1=confusing,5=clear
Q3:	Indicate the word cloud that best describes the set of documents for the group Choice: Term Frequency, Named Entities
Q4:	Extent to which each document fits the group. Scale: 1=does not fit at all,5=fits very well
Q5:	What other representations would meet your expectations

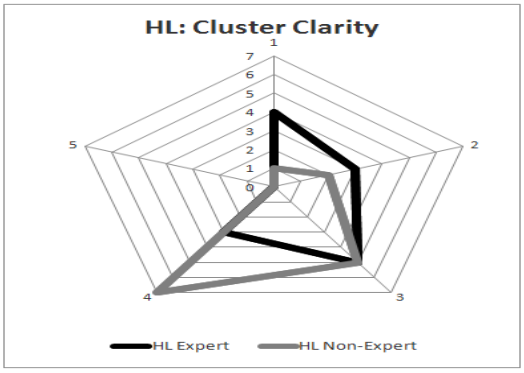
Clustering Clarity

[**Question 2**]: Figure 5.4, depicts the results for Assessment Question-Q2. We were interested in knowing whether the high probability clusters correspond to clusters that actually make sense to the users. In order to do this, each cluster was evaluated by

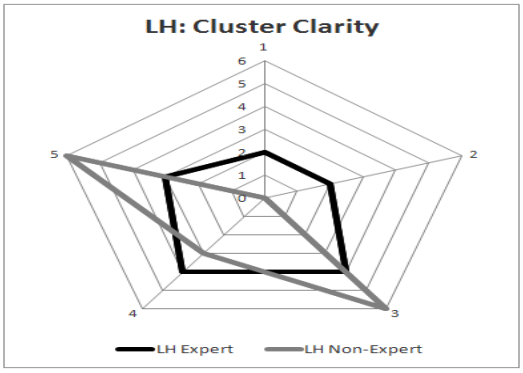
the user for its overall clarity. It shows the frequency of each rating, for each pruning criteria among each user group. Quite noticeable for the HH-pruning criteria, is that the shapes of the two graphs for the experts and non-experts are quite similar. Both user groups found the clusters within the HH-pruning criteria to be rather clear (see Figure 5.4a).



(a) HH



(b) HL



(c) LH

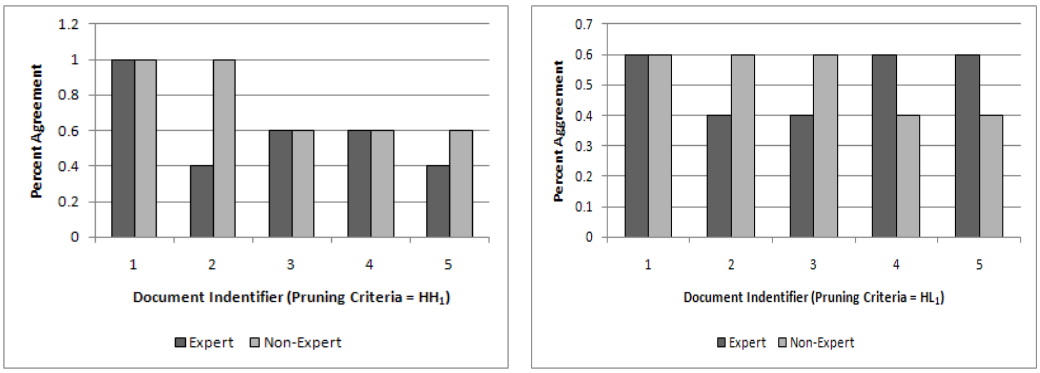
Figure 5.4 Overall Clarity for Pruning Criteria HH (5.4a); HL (5.4b); and LH (5.4c) based on the extent to which the set of documents for the group makes sense to the user; using the scale: 1=confusing,5=clear.

In contrast, for the LH-pruning criteria in Figure 5.4c, the experts seem to be quite neutral and do not rate strongly for, or against the clarity of the clusters. This suggests that in the presence of lower probability clusters, in the LH category, the experts are more unclear about what constitutes a pattern, unlike the non-expert users. For the HL-pruning criteria (Figure 5.4b) one can notice that there is a clear overlapping (that splits around the most neutral rating of 3). This suggests that given the mixture of low probability documents with high probability clusters, the users are completely mixed, about the clarity of the clusters. Finally, we also notice in the Figures 5.4b and 5.4c, that in the presence of increasing noise, the expert users are more conservative about their ratings, unlike the non-experts, who still rate the

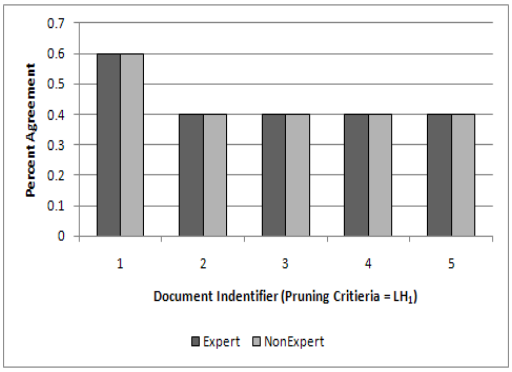
clarity of the cluster comparatively higher (in the range of $3 \cdots 5$), than the experts. The Expert group, assesses the quality of the clusters differently than the Non-Expert group in the presence of more noise (i.e., lower probability).

Document Fit within a Cluster

[**Question 4**]: Figure 5.5, depicts the results for Assessment Question-Q4. In this experiment, we are interested in assessing the quality of clusters, with respect to the documents contained therein. The users were asked to rate the extent to which the five documents in each of the three pruning criteria fit the cluster. Figure 5.5a, 5.5b and 5.5c represent the results of the percent for trial 1 using the HH, HL and LH pruning criteria, respectively.



(a) HH **(b)** HL



(c) LH

Figure 5.5 Percent agreement for the extent to which the documents of the HH, HL, and LH pruning criterial fit the cluster.

In Figure 5.5a, we notice that the percent agreement is is stratified across the values of (1.0, 0.6 and 0.4) for both user groups. This stratification suggests that users saw distinct sub-clusters of documents as belonging to the same equivalence

class. In Figure 5.5b, we see a similar stratification. In contrast, however to the HH criteria, here we see that the peak stratification is significantly lower: at a value of 0.6 compared with 1.0 in Figure 5.5a. This suggests that, again, users perceive sub-clusters, yet, they were less confident about the meaningfulness of the document with respect to the cluster. This can be explained by the fact that the documents that were included, were taken from the low probability range of the quartile. This means that low probability documents do, in fact, lead to less meaningful and less clear clusters from the users perspective. Quite remarkably, in Figure 5.5c, we notice that for nearly four out of the five documents, both user groups, show the least agreement about the fit of the document within the cluster. This lead us to believe that when using high probability documents with a low probability cluster, the users do not perceive that the documents fit the clusters well. In our instantiation of the framework, we relied up the cohesion of the documents with the cluster to validate the cluster. We believe that alternative metrics should be considered. The silhouette metric, for example, takes into account, not only the cohesion of the document within the cluster, but also its separation, with respect to the other clusters. Validating the clusters according to other criteria, would bring further insights into the task of assigning documents to a cluster.

Practitioner Assisted Feedback Loop (L2): *Based on the stratification in the percent agreement scores we propose that another cohesion metric for intrinsic validation in Pattern Validation stage, be considered.* We conclude that the HH Pruning Criteria is meaningful for the users, whereas the HL and LH clusters are less clear and meaningful.

5.3.5 Results III: Expert Assessment of Cluster Representation

[Question 3]: In order to ensure that the unsupervised methods produce results that are valuable for the human users, it is crucial that systems also provide a means to represent the clusters so users can interpret them with ease. We consider a cluster representation based on two different types of word clouds and seek to know which one the users find useful, if any. Figure 5.6 shows the results when users, were asked to choose which type of word cloud they thought best represents the cluster. The results overwhelmingly show that the users preferred the term frequency word cloud over the named entity word cloud. The term frequency clouds was constructed using the frequency for all terms that were present in the documents for, the given cluster. In contrast, the named entity word clouds were constructed by using the named entities with a threshold probability as determined by the fourth quartile of the term probability distribution.

Practitioner Assisted Feedback Loop (L2): *Since the term frequency word cloud was preferred to the named entity word cloud, we propose that the choice of features for the event detection algorithm should not only contain entity types such as*

medical condition and location, but also include non-entity terms.

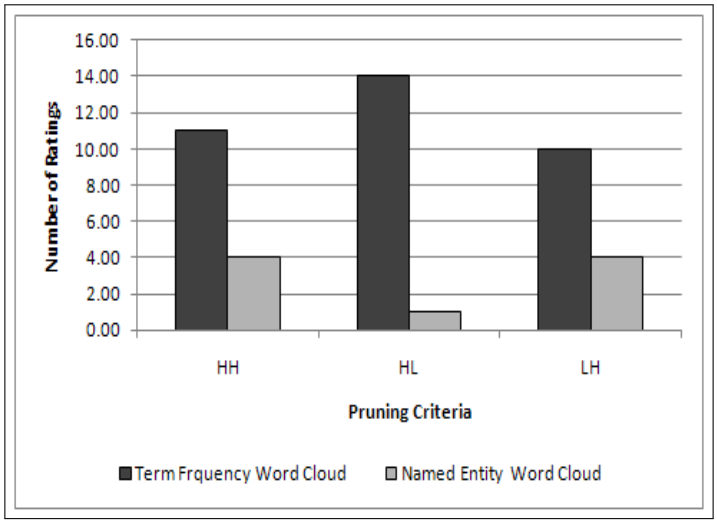


Figure 5.6 Number of Ratings indicating the word cloud that users thought best describes the set of documents for the group. The choice of words cloud representations where: Term Frequency and Named.

User's Description of a Cluster

[Question 1]: The overlap between the terms in the word cloud and the keywords entered by all users for each pruning criteria was computed. The results show that, although the event detection algorithm was capable of suggesting words that the user would have selected to represent the cluster, there is still some room for improvement. A closer examination showed that the types of terms entered by the users, consisted of entities types that we did not include as input into the algorithm.

Practitioner Assisted Feedback Loop #1***:*** *Since the overlap between the entity types in the word cloud and the keywords entered by all the users is low, we propose that the selection of features for the event detection algorithm should contain additional entity types such as victims and symptoms.*

User Remarks and Feedback

[Question 5]: In this section, qualitative feedback based on the users remarks are presented. All users found the word clouds effective. However, when asked what other representations would meet their expectations, they thought it useful to have more robust and *interactive word clouds*. Some user's comments for more interactivity, include: a) the use of color coding to identify different entity types in the cloud; b) interactive clouds to remove words and redraw word clouds; c) the use of an ontology

to collapse semantically related terms that are in the cloud; and d) the elimination of very small cloud terms.

Practitioner Assisted Feedback Loop (L3): *Since the users wanted to eliminate very small terms and collapse semantically related terms: we propose that: 1) more pruning should be considered for the term pruning phase, and 2) the terms that are semantically related should be aggregated before they are used as input into the pattern recognition algorithm.*

Although the experts in the study are not trained in the use of Web 2.0 tools, they were quite clear about the ways in which the system can help them to better manage and manipulate the complexity of the outputs for unsupervised disease reporting events, in a Web 2.0 style.

Practitioner Assisted Feedback Loop (L1): *Based on the desire for the users to have more control over the terms that appear in the word cloud, more interactivity should be produced to help users manipulate and digest the content.*

Finally, users thought other types of representations could be considered. These representations dealt mostly with alternative ways to see the same underlying data. Users even suggested having a toggle button, so that they could decide which alternate representation, they could see, for a given situation.

In summary our results show that:

1. Field practitioners can identify clear cluster that have been produced by an unsupervised event detection algorithm,

2. patterns with a high cluster probability and high document probability are better suited for field practitioners, in digesting and interpreting the meaning of the pattern,

3. the use of term frequency word clouds can help field practitioners to distinguish patterns with respect to their quality.

5.3.6 Discussion

These results suggest that an unsupervised approach to detecting disease reporting clusters can, at least, align with clusters that have been detected with a rule based approach. Also, based on an extrinsic qualitative evaluation, we would prune clusters that have a precision and recall below 80%. In the $DRMC$, we produce many more clusters than in the rule-based approach. Work in this direction includes an extensive case study to explore this question by comparing the unsupervised approach with the rule-based approach of MedISys[FSN]. Finally we note that numerous systems exist to detect disease reporting events [HNW+09, LSW+09]. None of these existing EI systems use an unsupervised event detection approach. As such, they do not allow for disease reporting events to be identified in the absence of predefined matching keywords or linguistic rules.

One limitation of this work is that it is an instance-based approach. Since no model is built that can be reused on new data, we must re-run the event detection for each new set of results. In future work, we plan to consider an on-line alternative to this instance based approach. Finally, given the exploratory nature of EI, better mechanisms are needed to support the aggregation of events for statistical models, as well as navigation for epidemic intelligence gathering. On the one hand, public health officials are only interested in receiving a limited number of events per session; yet on the other hand, they want to be adequately informed, and not miss potentially relevant ones. Identifying the balance for this trade-off is a challenge and currently not handled.

In future work, a more detailed evaluation of the proposed algorithm will be undertaken. This includes additional measures, such as the B-Cube for computing precision and recall. Also, it should be noted that many factors influence the quality of disease reporting clusters. For example, the existing prevalence levels of a disease or even the personal preference, of the information seeker: such as their geographical location or occupation. Assessing the quality of an clusters based on such factors requires a more robust qualitative evaluation with input from domain experts. We plan this as future work.

5.4 Chapter Summary and Outlook

We realized the approach by discovering clusters in an unsupervised manner and further present a framework to evaluate the their quality. We also presented a novel framework with which field practitioners can assess the results harnessed from the EI for intelligence gathering to detect public health events from unstructured text. We presented formalizations for characterizing disease reporting events and how this is embedded in the user-centric framework. We presented a study including over 30,000 documents and numerous domain experts to validate the both the components and the underlying unsupervised event detection algorithm and the feedback interaction loop between the two. We have shown that 1) field practitioners are able to find clear clusters that have been produced by an unsupervised event detection algorithm, 2) patterns with a high cluster probability and high document probability are better suited for field practitioners, in digesting and interpreting the meaning of the pattern and 3) the use of term frequency word clouds can help field practitioners to distinguish patterns with respect to their quality.

In this work, we draw attention to some of the problems faced in the area of Epidemic Intelligence. Particularly, we shed a new light on the task of disease reporting clustering and posit that disease reporting events can be detected in an unsupervised manner, but it requires a filtering mechanisms and additional support that help experts interpret and navigate the clusters that are produced. The impact of such work in practice and research is two-fold. First, is an improved understanding of the types

of visualization and representations that are useful to domain experts in the areas of epidemiology. Second, is bridging the gap between system mining and filtering, and the domain experts in the field, who must rely upon a summarized interpretation and an elucidation of facts for Social Media-Based Epidemic Intelligence systems. The practical impact is that such frameworks as the one proposed in this work, will allow field practitioners to gain a better trust for the outputs of Epidemic Intelligence systems. This also implies that a mechanism for ensuring users trust in the results can be envisioned.

Future work will include incorporating the information from the practitioner assisted feedback loops. We also plan as extension of the social media sources to, more noisy data, in a streaming setting, to further stress test the ability of our framework to represent this complex data in a human understandable way. Finally an interesting area for future research is to include official sources of information from EI as a baseline such as the WHO or ProMED-mail database statistics, to enable an automated correlation between human-centric data and other EI data.

Summary and Open Directions

Laboriously annotating training data for building supervised learners is a common problem faced in many domains. In this thesis, we explored various approaches to Limited Supervision learning with the intension of tackling the label bottleneck problem in the domain of Epidemic Intelligence (EI). Our aim was to provide a means of filtering disease reporting mentions from unstructured web text without relying upon large volumes of labeled training data to build the classifier.

We described three important types of limited supervision approaches, namely: (1) *Semi-Supervised Learning with Weak labels*, in which we exploited the properties of an auxiliary domain to acquire labels for relevant patterns, that could then be propagated to bootstrap a self-trained classifier within a desired, target domain; (2) *Active Learning with Label Resolution*, with the aim of reducing the budget associated with requesting labels from an oracle; and (3) *Unsupervised Dection of Disease Reporting Mentions* to detect disease reporting mentions without using any labels at all.

We first summarize our major contributions with respect to the three above mentioned domains, provide a scenario in a more global context to show the benefits of an entire system in which limited supervision is used; and then we discuss some issues which remain open for future investigation.

6.1 Summary of Contributions

(1) Semi-Supervised Learning with Weak labels. In Chapter 3, we explored the extent to which semi-supervised learning (SSL) could be used to filter disease reporting mentions for the short text of sentences. SSL have the advantage of only requiring small amounts of labeled data, and assumes the existence of much larger quantities of unlabeled data. We found a SSL to be quite effective when compared with a traditional learner, that relies upon a much larger amounts of labeled data apriori.

Weak Labeling and Cross-Corpora Learning. Though a SSL only requires

a small amount of labeled data, the question still remains, from where do we obtain even these small labeled data? Inspired by the work done in the domain of distant supervision, we address the problem of building a semi-supervised learner from EI knowledge bases in an automated manner, i.e., with weak labels. The motivation for doing so is that huge volumes of disease mentioning patterns exist in these knowledge bases; but these patterns are themselves unlabeled. We were able to exploit a simple set of heuristics, which allowed us to bootstrap a auxiliary short text classifier for filtering similar types of mentions in a comparable corpus.

Structural Kernels for Tackling Recall Gating. A key aspect of our approach is the use of structural features, and a kernel-driven method for classification. Kernel-based methods allow linguistic structures to maintains their discreet, structural properties during classifier training. We found structural representations to be useful in overcoming the recall-gating problem in short text cross-corpora bootstrapping.

Named Entity Recognition. In addition to the use of structural features for overcoming the recall-gating problem, we obtained significant performance boosts in recall when using entity bearing features within the short text. Although a number of tools exist for detecting basic entity types (person, location, organization), we found that an appropriate infectious disease and organism tagger to be lacking. Our organism tagging extracts Organism-by-Family, Organism-by-Occupation and Organism-by-Geography and animals. The medical condition tagger focuses specifically on infectious diseases, their pathogens and associated symptoms. We provide to the community a dictionary of terms that we constructed, with the help of domain experts, for building these taggers.

Expert Feedback. We notice with the exception of a few systems, most EI filtering systems do not employ the assessment of the domain experts to judge the final quality of the results. We solicited the feedback from prominent experts in the domain and was able to draw many conclusions about the feasibility of our approach. One of the main insights is the need to build an ensemble of classifiers as opposed to a single, monolithic one. This is motivated by our assessment results and the fact that relevance for a domain expert depends heavily upon their specific task, which can change, depending on their investigative needs.

(2) Active Learning with Label Resolution In Chapter 4, we approached the label bottleneck problem from the perspective of a limited budget strategy, or active learning (AL). We used even sparser text than sentences, namely, Twitter messages (or tweets). In AL, the labels are *hidden*, and each of them can be revealed only at a cost. The active learner is allowed to pro-actively select the most productive training instances, without having to label and supply the learner with more data than necessary. The label bottleneck is overcome by only asking the oracle for advice when the utility of doing so is high.

Label-Aligned Clustering in Active Learning. The most common approach to using clustering in active learning is the consider a cluster over the entire set of

unlabeled documents. These approaches have been known to suffer when: (i) no obvious clustering exists; (ii) clusterings exist, but are at an unknown granularities; (iii) the classifier labels themselves are not be aligned with the active learner clusters (label-cluster alignment problem). In this thesis we tackle the label-alignment problem with semi-supervised clustering, using the machinery of Partially Labeled Dirichlet Allocation (PLDA). PLDA offer us the ability to incrementally cluster and labeled the most uncertain (i.e., productive) instances in a quantitative manner. In this way we resolve many labels automatically, and only defer to the the oracle when automated predictions are below a certain level of confidence.

Mutual Exclusion. A key aspect of our approach is the feature space used in handling the mutual exclusion between training instances. Using the labels of the current seeds, the Partially Labeled Dirichlet Allocation (PLDA) models the context that overlaps among the seeds as background knowledge that is commonly shared by all training instances for the current iteration, while maintaining their polarity. A new feature space for the once overlapping view is constructed by excluding common background clusters. Exploiting such a model allows us to resolve more of the uncertain instances automatically with less human intervention than a global clustering strategy.

Domain Expert and Crowdsourcing. Active Learning assumes that an oracle, one who is preferably a domain-matter expert, is present to provide labels. In practice, however, domain experts in EI are far to busy to regularly provide labels for filtering disease reporting mentions, particularly at the volume and speed that might be required for maintaining a classifier. For this reason, we assessed the potential of for-hire or Human Intelligence Task Workers to provide relevance judgments for building disease reporting mention classifier. We found that compared to expert judgments, the crowdsourcing judgments are comparable to those provided by an expert. We find that the benefit of using crowdsourcing outweighs any limitations; since we are able to obtain a: high volume, with rapid results, and multiple judgments from a crowd, which is simply not possible from domain experts. Moreover, the potential cost of employing HIT workers can be attenuated when combined with label resolution strategies presented in this chapter.

(3) Unsupervised Dection of Disease Reporting Mentions In Chapter 5, we tackled the label bottleneck problem from the perspective of unsupervised learning. In unsupervised learning, one seeks to find salient patterns in the data, which are above and beyond what would be considered pure unstructured noise. In particular, we focused on the use of generative models (mixture models) for clustering documents that are represented as a sparse feature vector of entities relevant to our task. Generative models represent a hypothesis about how instances are generated for each class (cluster). Little or work has been done in filtering disease reporting mentions in this way.

Domain Expert Interpretable. One of the challenges faced in producing clusters, in general, is that it can leads to very complex results. This complexity poses

a significant challenge for an epidemic investigator, given the number of potential clusters. Additionally, since the pattern is not labeled apriori, the significance and meaning of the pattern must be interpreted. We have shown that: 1) field practitioners are able to find clear clusters that have been produced by a retrospective, unsupervised event detection algorithm; 2) patterns with a high cluster probability and high document probability are better suited for field practitioners in digesting and interpreting the meaning of the pattern; and 3) the use of term frequency word clouds can help field practitioners to distinguish patterns with respect to their quality. We showed that experts were in fact able to use the results to check outbreaks, summarize the corpus, and guide exploration of its contents.

In conclusion, we proved using limited supervision approaches for detecting disease reporting mentions from unstructured short and sparse text is possible. We showed three viable options for building a limited supervision classifier and assessed our results from the perspective of prominent domain experts in the field. The implications for the work that has been done here is that if we could employ suitable mechanisms for tackling the label bottleneck, then we could ease the costs of constructing and maintaining a classifier in the domain of EI, thereby enabling investigators in detecting potential threats from unstructured text on the Web.

6.2 Limited Supervision Learning in Context

Throughout this thesis, numerous results have been presented to show the extent to which the aforementioned limited supervision approaches of: (1) Cross-Corpora Learning; (2) Label-Aligned Clustering; and (3) Unsupervised Learning, can help to tackle the label bottleneck problem. In this section, we provide a more global context, to show the benefits of how the results presented here could be used within an entire EI system that is devoted to building and maintaining a trainable classifier for ensuring the veracity of unstructured text documents. Veracity underpins a web-based intelligence system because it emphasizes the importance of screening out data that is not useful for intelligence gathering.

In order for us to ensure veracity, we need to rely upon automatic, machine learning approaches to handle message filtering via classification and clustering. Also, we assume a near-real-time intelligence gathering scenario, for which the EI pipeline results are not available immediately, but can be expected to be updated several times a day, within a 24 hour period, and made available within the M-Eco GUI. In the absence of large volumes of training data to support the data veracity for message filtering module, we rely upon the limited supervision approached discussed in this thesis.

6.2.1 Data Sources and Variety

In Chapter 3, we assume that large volumes of revelant mentions of historical events exist, from which we can glean labels. Numerous sources exist for the domain of public health, particularly citizen journalists, hobbyists (Flutrackers, RSOE) who serve as aggregators of public health related events. Using the large volumes of past instances from these types of auxiliary sources, we are able to collect labels to build filters for our desired target domain of unlabeled data, such as RSS news feeds and blogs, with reasonable success. We accomplished this by wrapping the Support Vector Machine with an incremental, bootstrapped, (or Semi-Supervised) learner. In fact, in the presence of limited amounts of training data for the desired target domain, we showed, in Section 3.5.6 Recall Boosting Strategy, that the steps taken to generalize the structural features for training the auxiliary classifier, allowed us to build a classifier for the target domain, using seeds from the auxiliary domain, without overfitting, as demonstrated with by an F1-measure of 89.99% (Table 3.9).

It should be noted that not all algorithms are well suited for all varieties and sources of data. For example, although the unsupervised approach to limited supervision, (Chapter 5) is language agnostic, it does not perform as well on the short sparse text of Tweets due to the limited context. Likewise, the cross-corpora approach is not well suited for application to a target domain such as Twitter, even when the topics are devoted to the same subject.

6.2.2 Document Label Acquisition Time

In Chapter 3, the time taken to acquire an initial set of labeled documents from the auxiliary domain for seeding the semi-supervised learning is only limited by the ability to process a given volume of relevant instances of disease reporting mentions. When done once, the application of a model built from these training instances on the target domain, is negligible. Since full-text models are not applicable for the short and sparse text of Twitter, we address this issue of limited supervision learning for tweets by relying upon crowdsourcing. In an Active Learning setting a number of benefits are achievable. In our experiments with crowdsourcing, it took roughly six hours for the CrowdFlower workers to label a set of 1,500 tweets (with any completed jobs being available for use at any time). Even though this labeling time, using crowdsourcing, is much greater than the time to obtain ad-hoc labels from the auxiliary domain, the benefits makes a budget labeling for tackling the label bottleneck worthwhile for a number of reasons: (1) a human (albeit non-expert) could at least make an assessment on the validity of a label; (2) we were able to get multiple human judgments on a single data instance; (3) some measure of quality control over the labels among crowdsource workers could be realized; (4) the quality control over the labels, with respect to the domain experts was comparable; and (5) we could process a larger variety of data sources.

The ability of the crowd to properly make an assessment, heavily depends on the type of question posed. In our case, we were able to fine-tune the crowdsource question posed, with the help and input of domain experts. Also, with respect to handling quality control, among workers themselves, we interspersed gold labels among the raw data and workers who failed to correctly label a minimum number of gold instances could be excluded. In addition to filtering based on minimum number of correct gold instances, worker agreement was also used to filter workers. Armed with labels from the crowd, we showed that we could successfully grow a much larger labeled set of data, with an accuracy of up to 90 percent.

For less sparse text, such as sentences, where deeper linguistic processing is possible for representing the context of named entities, there is clearly a trade-off in terms of accuracy and document processing time. Although we do obtain a performance boost with a composite kernel (structural and non-structural features), compared to vector-based features alone, in practice, for a time-critical deployment strategy, using tokens features is better trade-of, since a full sentential parsing can be costly.

6.2.3 Balancing Accuracy versus Batch Processing Time

The time series data used for signal generation can be noisy, incomplete and sparse, in part, from propagation of errors within the message filtering stages, or simply due to the nature of the data. For example, noise within the time series data can be due to spurious events in which an entity is correctly detected, but its role is not relevant with respect to a public health threat. Incomplete or sparse time series data implies that instances of an event are missing or under-reported. This may occur due to: 1) the presence of processing errors - an acronyms or abbreviations not recognized as EI entity type, 2) the fact that people who are actually suffering do not make a social report of their condition, 3) the documents which contain these mentions have not be collected by the system - i.e., based on the imbalance between the genre collected (personal versus news), and 4) the minimum required entity types are not present. Sparse time series data refers specifically to low aggregation counts - which impact the anomaly detection algorithm.

In practice, when dealing with unstructured web text, no amount of data cleansing can completely guarantee a veracity-pure system. Yet, in spite of the inherent uncertainty, the data still contains valuable information. One step to handling uncertainty can be to improve the named entity recognition particularly for symptoms. The tradeoff of doing so, must be balanced against the cost. To help manage uncertainty, data fusion, i.e., combining multiple (less reliable sources); or using multiple languages, creates redundant, and often more accurate results.

The results presented in this thesis are important when is comes to dealing with fundamental issues of a document filtering system, namely: 1) utilizing large amounts of documents labelled as irrelevant (and non-relevant) in order to construct accurate trainable classifier models; 2) detecting when a given classification model is no longer

adequate; 3) acquiring new labelled data for updating a classifier, when needed; and 4) validating the results of a trainable classification models with respect to field practitioners and domain experts. These aforementioned issues are fundamental challenges faced by any long-term, automatic text filtering systems, not just one devoted to the task of Epidemic Intelligence.

6.3 Open Directions

Effectiveness for Early detection. An EI system is only effective when it can actually trigger a response as part of an Epidemic investigation. In order to measure the effectiveness of our system under these conditions, the results presented in this thesis would need a long-term commitment on the part of domain experts to monitor and interact with an operational system. As part of the M-Eco project the EI system we proposed, as whole, has not reached all of its full potential. In general, it seems that the undertaking to build an, integrated, end-to-end system proved to be more challenging than we had initially envisioned. Nonetheless, we feel that many of the lessons learned as part of this work can provide guidance for similar endeavors. We present a synopsis of these issues below.

Multi-Linguality. Public health is of concern to many individuals, as such is it necessary to provide multi-lingual support for filtering. In this work, we have restricted our analysis to the English language. Specifically for the weak labeling and Cross-Corpora approach, we relied heavily upon the EI knowledge bases for bootstrapping the learner for news and blogs. Different auxiliary domains, or machine translation techniques would be needed to support mappings for languages other than English.

Visual Support. Given the exploratory nature of EI, better mechanisms are needed to support the aggregation of cluster for statistical models. On the one hand, public health officials are only interested in receiving a limited number of clusters per session; yet on the other hand, they want to be adequately informed, and not miss potentially relevant ones. Identifying the balance for this trade-off is a challenge and currently not handled.

Domain Expert Feedback. The collaborations with different domain experts as part of this work has been inspiring. All to often, systems are built without the actual needs of end-users in mind, understandable, since it is a costly endeavor. In this work we have only scratched the surface and more extensive end user evaluation is needed and under more rigorous conditions. One such extension would be to use many more domain experts and ensure that they are trained such that a robust inter-annotator agreement among them (e.g., Cohen-Kappa) is achieved.

Temporal Dimensions. We also plan to take up the temporal aspects of dynamic stream classification with limited supervision. In doing so, a number of new challenges are faced. This includes: (1) detecting feature changes; as well as (2) handling these changes once they are detected. The static classification methods presented here have

not been tuned to consider temporal aspects such as those required for long-term and time sensitive surveillance activities. Detecting changes within the classification stream is needed so that we known when and if, it is time to retrain the classifier. Specifically, when it is time to include new features; and remove outdated ones. An important consideration for detecting feature change is that the processes for extracting features is not too costly.

Data Velocity and Volume. Finally, as part of ongoing work, we consider the aspects of our system that need to be adapted to handle large volumes of data (i.e., greater then 100 million training instances). The techniques presented in this work, simply are not suitable for such data volumes. One promising step in this direction is to tackle the label bottleneck by a large scale, and distributed redesign of the Partially Labeled Dirichlet Allocation (PLDA). PLDA allowed us to automatically acquire labels, via inference, with a label-aligned aware clustering model. Additionally, large scale redesign of both, the classifier and clustering algorithms, would allow hybrid predictors to be built, such that the additional types of features obtained from one algorithm, could be used to enhance the other.

Given the scale of information within today's Web, it is has becoming increasingly difficult to adequately maintain a fully supervised learner - adapting the techniques presented in this work, not only for EI, but also other domains as well, will help bring us another step closer to better supporting the information needs of users in the World of Web Science.

Dictionary of Terms Use for Named Entity Extraction

A.1 Organism Entities

In this section, we provide the complete list of terms used to construct each of the organism named entities dictionaries, namely: (i) Non-Human Organisms; (ii) Persons-by-Geography; (iii) Persons-by-Population; and (iv) Persons-by-Occupation. Each of aforementioned types of organisms entities were extracted with a simple dictionary based approach using LingPipe http://ir.exp.sis.pitt.edu/ne/lingpipe-2.4.0/

Table A.1 Non-Human Organisms refers to the textual mention of a non-human animal (e.g., swine, horse).

heifers	heifer	animals	animal	insectivorouses
insectivorous	horses	horse	aardvarks	aardvark
albatrosses	albatross	alligators	alligator	alpacas
alpaca	amphibians	amphibian	amurs	amur
anacondas	anaconda	anemones	anemone	ants
ant	anteaters	anteater	antelopes	antelope
apes	ape	armadillos	armadillo	arthropods
arthropod	asses	ass	audaxes	audax
aye-ayes	aye-aye	baboons	baboon	badgers
badger	bandicoots	bandicoot	bangles	bangle
barnacles	barnacle	barracudas	barracuda	basilisks
basilisk	basses	bass	bassets	basset
bats	bat	bears	bear	beavers
beaver	bees	bee	beetles	beetle
belugas	beluga	billies	billy	birds
bird	bisons	bison	blackbacks	blackback
blackbirds	blackbird	blowfish	boas	boa

Non-Human Organisms: continued on next page

Table A.1 Non-Human Organisms refers to the textual mention of a non-human animal (e.g., swine, horse).

boars	boar	bob-cats	bob-cat	bobcats
bobcat	herds	herd	bonoboes	bonobo
boobies	booby	boomers	boomer	bovinaes
bovinae	boxers	boxer	bucks	buck
budgies	budgie	buffaloes	buffalo	bugs
bug	bulls	bull	bunnies	bunny
butterflies	butterfly	buzzards	buzzard	caimen
caiman	calves	calf	camels	camel
canaries	canary	canids	canid	canines
canine	caribous	caribou	carnivores	carnivore
carp	cats	cat	caterpillars	caterpillar
catfish	cattles	cattle	centipedes	centipede
cetaceans	cetacean	chameleons	chameleon	chamois
chantelles	chantelle	cheepers	cheeper	cheetahs
cheetah	chicks	chick	chickens	chicken
chihuahuas	chihuahua	chimpanzees	chimpanzee	chinchillas
chinchilla	chipmunks	chipmunk	chordates	chordate
chrysalises	chrysalis	chulengoes	chulengo	chupacabras
chupacabra	clams	clam	cobs	cob
cobras	cobra	cocks	cock	cockatiels
cockatiel	cockatoos	cockatoo	cockers	cocker
cockerels	cockerel	cockroaches	cockroach	cod
codfish	codlings	codling	cohoes	coho
collies	collie	colts	colt	cormorants
cormorant	cossets	cosset	cougars	cougar
cows	cow	coyotes	coyote	crabs
crab	cranes	crane	crawfish	crays
cray	crias	cria	crickets	cricket
crocodiles	crocodile	crocodilians	crocodilian	crows
crow	cubs	cub	cuckoos	cuckoo
cuttles	cuttle	cygnets	cygnet	dacshunds
dacshund	dalmations	dalmation	dams	dam
damsels	damsel	darts	dart	dears
dear	deer	mule deer bucks	mule deer buck	deguses
degus	dik-diks	dik-dik	dingoes	dingo
dobermen	doberman	dodoes	dodo	doe
dogs	dog	dogfish	dogues	dogue
dollies	dolly	dolphins	dolphin	donkeys
donkey	dormice	dormouse	doves	dove

Non-Human Organisms: continued on next page

Table A.1 Non-Human Organisms refers to the textual mention of a non-human animal (e.g., swine, horse).

dragons	dragon	dragonflies	dragonfly	drakes
drake	drones	drone	ducks	duck
duckbills	duckbill	ducklings	duckling	dugongs
dugong	eagles	eagle	eaglets	eaglet
earthworms	earthworm	earwigs	earwig	echidnas
echidna	eclectuses	eclectus	eels	eel
efts	eft	egrets	egret	elands
eland	elephants	elephant	elks	elk
emus	emu	ephyras	ephyra	equines
equine	ernes	erne	ewes	ewe
eyases	eyas	falcons	falcon	falconiformes
falconiforme	farrows	farrow	fawns	fawn
felidaes	felidae	felines	feline	ferrets
ferret	fillies	filly	finches	finch
fingerlings	fingerling	fireflies	firefly	fish
flamingoes	flamingo	flappers	flapper	flatworms
flatworm	fledglings	fledgling	flies	fly
foals	foal	fowls	fowl	foxes
fox	frogs	frog	froglets	froglet
fries	fry	ganders	gander	gastropods
gastropod	gaurs	gaur	gazelles	gazelle
gerbils	gerbil	giraffes	giraffe	gnats
gnat	gnus	gnu	goats	goat
gobblers	gobbler	geese	goose	gophers
gopher	gorillas	gorilla	goslings	gosling
grasshoppers	grasshopper	grilses	grilse	groundhogs
groundhog	hogs	hog	grouses	grouse
guanacoes	guanaco	guillemots	guillemot	guineas
guinea	gulls	gull	hakes	hake
hammerheads	hammerhead	hamsters	hamster	hares
hare	harts	hart	hatchlings	hatchling
hawks	hawk	hedgehogs	hedgehog	hembras
hembra	hens	hen	herons	heron
herrings	herring	hinds	hind	hippoes
hippo	hippopotamuses	hippopotamus	hobs	hob
hoglets	hoglet	hornets	hornet	hounds
hound	hubs	hub	hummingbirds	hummingbird
huskies	husky	hyenas	hyena	hyraxes
hyrax	ibises	ibise	ibis	iguanas

Non-Human Organisms: continued on next page

Table A.1 Non-Human Organisms refers to the textual mention of a non-human animal (e.g., swine, horse).

iguana	iguanodons	iguanodon	impalas	impala
inchworms	inchworm	insects	insect	irrawaddies
irrawaddy	jacks	jack	jackals	jackal
jackrabbits	jackrabbit	jaguars	jaguar	jakes
jake	jays	jay	jellyfish	jennies
jenny	kangaroos	kangaroo	keets	keet
kittens	kitten	koalas	koala	komodoes
komodo	kookaburras	kookaburra	koupreys	kouprey
krills	krill	kudus	kudu	lamas
lama	lambs	lamb	lambkins	lambkin
lancelets	lancelet	larks	lark	larvas
larva	larvaes	larvae	leeches	leech
lemurs	lemur	leopards	leopard	leverets
leveret	lias	lia	lices	lice
lions	lion	lionfish	lizards	lizard
llamas	llama	lobsters	lobster	loriss
loris	louse	lynxes	lynx	maggots
maggot	magpies	magpie	mallards	mallard
mammals	mammal	manatees	manatee	mantises
mantis	mares	mare	marmots	marmot
marsupials	marsupial	mastiffs	mastiff	meerkats
meerkat	minks	mink	moles	mole
mollusks	mollusk	mollies	molly	mongooses
mongoose	monkeys	monkey	mooses	moose
mosquitoes	mosquito	mice	mouse	mules
mule	muskoxes	muskox	muskrats	muskrat
narwhals	narwhal	nenes	nene	newts
newt	nightingales	nightingale	nutrias	nutria
nyalas	nyala	nymphs	nymph	ocelots
ocelot	octopuses	octopus	okapis	okapi
opossums	opossum	possums	possum	orangutans
orangutan	orcas	orca	oryxes	oryx
ospreys	osprey	otters	otter	owls
owl	owlets	owlet	oxen	ox
oysters	oyster	pandas	panda	pangolins
pangolin	panthers	panther	parrs	parr
parrots	parrot	partridges	partridge	peachicks
peachick	peacocks	peacock	peafowls	peafowl
peahens	peahen	peeps	peep	pelicans

Non-Human Organisms: continued on next page

Table A.1 Non-Human Organisms refers to the textual mention of a non-human animal (e.g., swine, horse).

pelican	penguins	penguin	pigs	pig
pigeons	pigeon	piglets	piglet	pikas
pika	pinschers	pinscher	planulas	planula
platypuses	platypus	polliwogs	polliwog	polyps
polyp	ponies	pony	porcupines	porcupine
poriferas	porifera	porpoises	porpoise	pottoes
potto	poults	poult	prawns	prawn
pronghorns	pronghorn	przewalskis	przewalski	puffins
puffin	puggles	puggle	pumas	puma
pups	pup	pupas	pupa	pupaes
pupae	puppies	puppy	quails	quail
queleas	quelea	quetzals	quetzal	rabbits
rabbit	raccoons	raccoon	rails	rail
rams	ram	rats	rat	rattlers
rattler	ravens	raven	stingrays	stingray
rays	ray	reeves	reeve	reindeer
reptiles	reptile	reynards	reynard	rhinos
rhino	ringworms	ringworm	robins	robin
rodents	rodent	roos	roo	rooks
rook	roosters	rooster	roundworms	roundworm
ruffs	ruff	salamanders	salamander	salmon
sandpipers	sandpiper	sapsuckers	sapsucker	scallops
scallop	scorpions	scorpion	scorries	scorrie
seafoals	seafoal	seahorses	seahorse	seals
seal	seastallions	seastallion	seastars	seastar
servals	serval	sharks	shark	wolves
wolf	sheep	shoats	shoat	shoebills
shoebill	shrews	shrew	shrewlets	shrewlet
shrimps	shrimp	snakes	snake	sires
sire	skates	skate	skunks	skunk
sloths	sloth	slugs	slug	smolts
smolt	snails	snail	snakelets	snakelet
sows	sow	spaniels	spaniel	spats
spat	spiders	spider	sponges	sponge
spoonbills	spoonbill	sprags	sprag	sprats
sprat	squabs	squab	squamates	squamate
squids	squid	squirrels	squirrel	stags
stag	stallions	stallion	starfish	steers
steer	stinkbugs	stinkbug	storks	stork

Non-Human Organisms: continued on next page

Table A.1 Non-Human Organisms refers to the textual mention of a non-human animal (e.g., swine, horse).

studs	stud	sucklings	suckling	swallows
swallow	swans	swan	swordfish	tadpoles
tadpole	tamanduas	tamandua	tamarins	tamarin
tapeworms	tapeworm	tapirs	tapir	tarantulas
tarantula	tarpans	tarpan	tarsiers	tarsier
tercels	tercel	termites	termite	terrapins
terrapin	terriers	terrier	terzels	terzel
tiercels	tiercel	tigers	tiger	tigresses
tigress	toads	toad	toadlets	toadlet
tods	tod	turkeys	turkey	tomcats
tomcat	tortoises	tortoise	trout	tuna
turtles	turtle	uakaris	uakari	ungulates
ungulate	urchins	urchin	urutus	urutu
vardens	varden	vervets	vervet	vicunas
vicuna	vipers	viper	vixens	vixen
voles	vole	vultures	vulture	wallabies
wallaby	walruses	walrus	wapitis	wapiti
warblers	warbler	warthogs	warthog	wasps
wasp	waterfowls	waterfowl	weaners	weaner
weasels	weasel	weevils	weevil	whales
whale	whelps	whelp	whippets	whippet
whoopers	whooper	wildcats	wildcat	wildebeasts
wildebeast	wolverines	wolverine	wombats	wombat
woodchucks	woodchuck	woodpeckers	woodpecker	worms
worm	wormlets	wormlet	wrens	wren
xantuses	xantus	xenops	xenop	xiphactinuses
xiphactinus	yaks	yak	yetis	yeti
zanders	zander	zebras	zebra	zebus
zebu	zorillas	zorilla	tick	ticks
flea	fleas	ruminants	ruminant	livestock
farm animal	farm animals	poultry		

Table A.2 Persons-by-Geography refers to the textual mention of a human by a geographical description (e.g., Moroccans, Brazilians).

Afghanistanians	Afghanistanian	Africans
Algerians	Americans	Angolans
Persons-by-Geography: continued on next page		

Table A.2 Persons-by-Geography refers to the textual mention of a human by a geographical description (e.g., Moroccans, Brazilians).

Argentinians	Armenians	Australians
Austrians	Azerbaijanians	Bagians
Bangladeshians	Bavarians	Belarusians
Belgians	Bolivians	bolognians
Brazilians	Britains	Bulacans
Bulgarians	Burundians	Californians
Cambodians	Cameroonians	Canadians
Cape Verdians	Caribbeans	Chinese
Colombians	Costa Ricans	Croatians
Cubans	East Africans	Ecuadorians
Egyptians	Eritreans	Ethiopians
Europeans	Europeans	Finlandians
Floridians	French	Fujians
Gabonians	Georgians	Germans
Ghanian	Hawaiians	Hungarians
Indians	Indonesians	Iranians
Iraqis	Israelis	Italians
Jamaicans	Japanese	Jordanians
Jordanians	Kenyans	Koreans
Kosovans	Kuwaitis	Macedonians
Malasians	Malaysians	Maldives
Maltese	Mediterraneans	Mexicans
Mongolians	Moroccans	Nairobians
Namibians	Nepalians	Nigerians
Nimbanians	North Americans	Norwegians
Pakistanians	Palestinians	Papua New Guineans
Paraguays	Polish	Romanians
Romans	Russians	Rwandans
Saudi Arabians	Scottish	Senegalese
Serbians	Singaporians	Slovakians
South Africans	South Koreans	Spainards
Sudanese	Swazilanders	Swedens
Switzers	Taiwanese	Tanzanians
Tehranians	Texans	Thailanders
Tibetians	Trinidadians	Tunisians
Turkmenistanians	Ugandans	Ukrainians
Uzbekistanians	Victorians	Vietnamese
Yatengans	Yemenese	Yemens
Yerevans	Yunlins	Yunnans

Persons-by-Geography: continued on next page

Table A.2 Persons-by-Geography refers to the textual mention of a human by a geographical description (e.g., Moroccans, Brazilians).

Zaires	Zambians	Zambians
Zanzibarians	Zhongshans	Zimbabwians
Ziyoratuts		

Table A.3 Persons-by-Population refers to the textual mention of a human by a family relation (e.g., brother, father), or a general population group to which a human belongs (e.g., elderly, group of children).

families	family	mothers	mother	adults
adult	aunts	aunt	babies	baby
boys	boy	brothers	brother	children
child	cousins	cousin	daughters	daughter
fathers	father	fellows	fellow	females
female	gentlemen	gentleman	girls	girl
god-child	godchilds	godchild	grandchilds	grandchild
granddaughters	granddaughter	grandfathers	grandfather	grandmothers
grandmother	grandparents	grandparent	grandsons	grandson
husbands	husband	infants	infant	kids
kid	lads	lad	ladies	lady
males	male	men	man	neonates
neonate	nephews	nephew	newborns	newborn
nieces	niece	offsprings	offspring	parents
parent	siblings	sibling	sisters	sister
sons	son	spouses	spouse	stepbrothers
stepbrother	stepcaddlabelhilds	stepchild	stepdaughters	stepdaughter
stepfathers	stepfather	stepmothers	stepmother	stepparents
stepparent	stepsisters	stepsister	stepsons	stepson
teens	teen	teenagers	teenager	toddlers
toddler	tribes	tribe	twins	twin
uncles	uncle	widows	widow	wives
wife	women	woman	refugee	refugees
tribal	tribals	tribesman	tribesmen	tribesperson
tribeswoman	tribeswomen	tribes man	tribes men	tribes person
tribes woman	tribes women	traveler	travelers	laborer
laborers	staff	staffers	pilgrim	pilgrims
people	soldier	soldiers	resident	residents
refugee	refugees	humans	human	person

Persons-by-Population: continued on next page

Table A.3 Persons-by-Population refers to the textual mention of a human by a family relation (e.g., brother, father), or a general population group to which a human belongs (e.g., elderly, group of children).

guest	individual	tourist	passenger	traveller
villager	laborer	inmate	pilgrim	resident
dweller	migrant			

Table A.4 Persons-by-Occupation refers to the textual mention of a human by their occupation (e.g., pilgrims, mine workers, nurse).

ederly	students	student
workers	worker	individuals
individual	tourists	tourist
passengers	passenger	travellers
traveller	villagers	villager
laborers	laborer	inmates
inmate	pilgrims	pilgrim
residents	resident	dwellers
dweller	migrants	migrant
pupils	pupil	guests
guest	personell	health care worker
healthcare worker	health care workers	healthcare workers
hospital worker	hospital workers	hospitalcare worker
hospitalcare workers	hospital care worker	hospital care workers
officials	staff	staffers
staffer	farm worker	farm workers
hospital worker	hospital workers	teacher
teachers	businessman	businessmen
businesswomen	businesswomen	businessperson
businesspeople	business man	business men
business women	business women	business person
business people	researcher	doctor
nurser		

A.2 Medical Condition Entities

In this section, we provide the complete list of terms used to construct each dictionary for the domain expert supplied medical condition named entities.

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

a sore throat	abdominal cramps
abdominal distension	abdominal pain
abdominal wall blister	abdominal wall numbness
abdominal wall rash	abdominal wall redness
abdominal wall tingling	abrupt watery diarrhoea
absent abdominal reflexes	absenteeism in children
acanthosis nigricans	aches
acute colitis-like symptoms	acute diarrhoea
acute hepatitis	acute ibd-like diarrhea
acute ibs-like diarrhea	acute ibs-like symptoms
acute virus hepatitis	adenoids bleeding
adenoids blister	adenoids rash
adenoids ulcer	adenovirus
adenovirus infection	adenovirus keratitis
adrenal adenoma	adrenal carcinoma
agitation	airway occlusion
altered respiratory pattern	amebiasis
amenorrhea	anal fissure
anal rash	anatomic obstruction
ankle rash	anthrax
anxiety	aphthous ulcer
apnea	appetite changes
ards	arm pain
arm rash	ascites
aseptic meningitis	asparatate aminotransferase elevation
ataxia	atypical bacteria - coxiella burnetti
avian influenza	axillary swelling
bacillus anthracis	back lump
back pain	back rash
bad breath	bilateral adnexal tenderness
bilateral crackles	bilharziose
bladder burning sensation	bladder infection
bladder inflammation	bladder lump
	Medical condition: continued on next page

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

bladder redness	bladder swelling
bladder ulcer	blebs
bleeding gums	bleeding under skin
blood in stool	blood in urine
blood streaked diarrhoea	bloody diarrhea
bloody ejaculation	bloody semen
bloody sputum	bone pain
borrelia	borrelia recurrentis
borreliosis	botulism
bowel obstruction	bradycardia in children
brain tumor	brain tumour
brasilianisches hmorrhagisches fieber (sabia)	brassy cough
breast rash	breathing difficulties
breathing difficulty	bronchial inflammation
bronchial redness	bronchial stiffness
brucella	brucella sp.
brucella spinal abscess	brucellosis
bruising in pregnancy	bubonic plaque
buttock rash	cachexia
calf rash	campylobacter
campylobacter enteritis	campylobacter sp.
catatonia	cervical lymphadenopathy
cervix ulcer	cheek rash
chest infection	chest inflammation
chest pain	chest rash
chest tenderness	chest weakness
chickenpox	chikungunyafieber
chills	chin rash
chinococcus	chlamydia psittaci
choking	cholera
chronic cough	chronic crohns-like symptoms
chronic ibd-like diarrhea	chronic ibd-like symptoms
chronic ibs-like diarrhea	chronic ibs-like symptoms
chronic progressive dysarthria	cirrhosis of liver
cjk	clay coloured stools
clitoris rash	clostridien
clostridium botulinum	clostridium difficile
clubbed fingers	clubbing in children
	Medical condition: continued on next page

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

clubbing of fingers	cold feet
cold-like symptoms	colitis-like abdominal pain
colitis-like symptoms	complete respiratory arrest
congenital rubella syndrome	congenital toxoplasmosis
conjunctivitis	constant diarrhea
constant throat pain	constipation
continuous spine pain	corynebacterium
corynebacterium diphtheriae	cough
coxiella burnetii	creutzfeldt-jakob disease
crohns-like abdominal symptoms	crohns-like diarrhea symptoms
crohns-like symptoms	cryptosporidiosis
cryptosporidium parvum	cystic fibrosis-like symptoms
dactylitis	dandy-fieber
decreased cardiac output	decreased hair growth
decreased oxygen saturation	decreased reflexes
decreased respiratory excursions	delayed puberty
delirium in children	dementia
dengue fever	diabetes insipidus
diaphoresis	diarrhea
diarrhea in children	difficulty walking
diminished breath sounds	diphtheria
diplopia	double vision
drenching night sweats	drooling
dull sounds	dyarthria in children
dysarthria	dysphagia
e.coli	ear rash
early summer meningoencephalitis	ebola fever
ebola virus	ebolavirus
echinococcosis	edema of larynx
ehec	ejaculate blood
elbow rash	elevated sedimentary rate
encephalitis	enlarged lymph nodes
enlarged tonsils	enteric fever
enterohemorrhagic e. coli	enteropathisch
episcleritis	erythema migrans
eschar	escherichia coli
excessive watery diarrhea	exercise symptoms
external os ulcer	eye inflammation
	Medical condition: continued on next page

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

eye pain	eye rash
eyebrow rash	eyelid rash
face swelling	facial paralysis
facial rash	failure to thrive
fast breathing	fatigue
fatigue in children	febris undulans
feet weakness	female infertility
fever	finger clubbing
finger rash	fixed pupils
flu	flu-like symptoms
foot rash	foot weakness
foot-drop in children	forearm rash
forehead rash	foreskin rash
formation of necrosis	foul smelling sputum
francisella tularensis	frequency of urination
fsme-virus	fuo
gangrene	gas gangrene
gastroenteritis	gastrointestinal symptoms
gi infection	giardia lamblia
giardiasis	green stool
groin rash	gum rash
haematuria	haemophilia influenza
haemophilus influenzae	halitosis
hand rash	hanta virus
hantavirus	hantavirus pulmonary syndrome
head itch	head rash
headache	healing symptoms
hemolytic-uremic syndrome	hemoptysis in children
hemoptysis in newborns	hemorrhagic rash
hemorrhagic rashes	hepatic failure
hepatitis	hepatitis a
hepatitis a virus	hepatitis b
hepatitis b virus	hepatitis c
hepatitis d	hepatitis d virus
hepatitis e	hepatitis e virus
hepatitis epidemica	high arched foot
high blood calcium	high blood pressure
high fever	hiv infection
	Medical condition: continued on next page

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

hiv-infektion	hoarseness in children
human spongiform enzephalopathy	hypoactive dtrs
hypoalbuminemia	hypogastric swelling
hyporeflexia	hyporeflexia in children
hypotonia in children	ibd-like abdominal pain
ibd-like symptoms	ibs-like diarrhea
icterus	impaired brain function
impaired motor movement	impaired speaking
increased lactate	increased salivation
increased thirst	indigestion
inflammatory joint effusion	influenza
influenza virus	intermittent bacterial pneumonia
intermittent crohns-like symptoms	intermittent foot weakness
intermittent generalised rashes	intermittent ibd-like diarrhea
intermittent ibd-like symptoms	intermittent ibs-like symptoms
intermittent palmar erythema	intermittent rib pain
internal os ulcer	interrupted urine flow
intestinal obstruction	iritis
itchy rash	jaundice
jaw rash	joint symptoms
knee rash	kneecap rash
knuckle rash	kryptosporidiose
kyphosis	kyphosis in children
lack of urine	lacrimation
large tender liver	laryngitis
laryngospasm	larynx deformity
larynx infection	larynx lump
larynx redness	larynx ulcer
lassa fever	lassa virus
leg paralysis	leg ulcers
legionella	legionella sp.
legionellose	legionellosis
leprosy	leptospira interrogans
leptospirosis	leucocytosis
leucopenia	leukopaenia
limp in children	listeria monocytogenes
listeriosis	liver cancer
liver enlargement	liver failure
	Medical condition: continued on next page

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

liver infection	liver mass
liver redness	low-grade fever
lower back pain	lower jaw rash
lumbar spasm	lung abscess
lung inflammation	lung redness
lung stiff	lymphadenitis in children
lymphocytosis	maculopapular rash
malaise	malaria
male infertility	malnutrition
marburg virus	massive hemoptysis
massive hemorrhaging	measles
measles virus	mechanical intestinal obstruction
mechanical obstruction	melena
memory changes	meningitis
meningococcal disease	meningoencephalitis
meningoenzephalitis	meningomyelitis
menorrhagia	menschliches immundefekt virus
menschliches immunschwche-virus	methicillin-resistant staphylococcus aureus
methicillin-resistenter staphylokokkus aureus	middle back pain
migratory arthritis	mild colitis-like symptoms
mild crohns-like symptoms	mild fever
mild ibs-like symptoms	miliary tuberculosis
mouth infections	mouth lesions
mouth lump	mouth pigmentation
mouth ulcers	mouth white patches
mrsa	mucoid sputum
mucopurulent secretions	mucopurulent sputum
mucous plugs	mucus buildup
mucus in stool	mucus symptoms
multi-organ dysfunction	mumps
muscle atrophy	muscle flaccidity
muscle pain	muscle spasms
muscle weakness	myalgia
mycobacterium	mycobacterium tuberculosis
mydriasis	myocarditis
nasal obstruction	nasopharyngeal tonsil bleeding
nasopharyngeal tonsil blister	nasopharyngeal tonsil rash
nasopharyngeal tonsil ulcer	nausea
	Medical condition: continued on next page

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

neck lump	neck rash
neisseria meningitidis	nerve root irritation
neurotoxic effects	new variant of creutzfeldt-jakob disease
night sweats	no appetite
noncardiogenic pulmonary edema	nonproductive cough
norovirus	norwalk-like virus
nostril ulcer	not feeling hungry
nuchal rigidity	ochropyra
ocular dysmetria	ophthalmoplegia
opisthotonus	optic neuritis
ornithosis	other pathogens of hemorrhagic fever
pain	pallor in children
palm rash	pancreas inflammation
paralysis symptoms	paraplegia
paratyphoid	paratyphoid fever
peripheral neuropathy	peripheral vasoconstriction
persistent cough	persistent high fever
personality changes	pertussis
pestsepsis	petichiae in pregnancy
pharyngeal edema	pharyngeal muscle spasms
pharynx bleeding	pharynx blister
pharynx ulcer	photophobia
plague	pleural effusion
pneumococcus	poliomyelitis
poliovirus	popliteal fossa blister
positive babinski sign	progressive weakness
prolonged fever	proprioception
ptosis in children	pubic area rash
pulmonary inflammation	pulmonary redness
pulmonary stiff	purpura in adults
purulent sputum	pus in stool
pustules	pyelonephritis in pregnancy
pyuria	pyuria in pregnancy
q fever	quadriceps muscle weakness
quadriplegia	rabies
rabiesvirus	raised white cells
rapid respirations	rapid respiratory rate
rash in children	rectal bleeding
	Medical condition: continued on next page

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

rectal mass	rectal ulcer
recurrent fever	recurring bouts of fever
recurring crohns-like symptoms	recurring ibd-like diarrhea
recurring ibd-like symptoms	recurring ibs-like symptoms
red spots	red throat
reduced tendon reflexes	relapsing fever
renal lump	respiratory disorder
respiratory inflammation	respiratory redness
respiratory stiff	rib itch
rib pain	rib rash
rickettsia	rickettsia prowazekii
rose spots	rotavirus
rubella	rubeola rubella
rubula	runny nose
rupture of blebs	sacral spasm
salivary gland pain	salmonella paratyphi
salmonella typhi	salmonellosis
scarlet fever	scarring alopecia
scleritis	scoliosis
scoliosis in children	scrotal rash
secondary infection	sensory ataxia
severe acute respiratory syndrome	severe colitis-like symptoms
severe crohns-like symptoms	severe diarrhea
severe headache	severe ibd-like diarrhea
severe ibd-like symptoms	severe ibs-like diarrhoea
severe ibs-like symptoms	severe weight loss
shigella	shigella infection
shigella sp.	shigellen
shigellosis	shin rash
short stature	shortness of breath
shoulder rash	sick
sinus ulcer	skin lesion
skin ulcer	skull itch
smallpox	sneezing
sole rash	sore joints
sore throat	speaking difficulty
speech abnormalities	spinal blister
spinal itch	spinal rash
	Medical condition: continued on next page

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

spinal spasm	spine pain
spitting blood	spongiform cerebral degeneration
spongiform encephalopathy	sputum
stertorous breathing	stomach itch
stomach rash	stool color
stridor in children	succussion sounds
sudden onset	suspected rabies exposures
swallowing difficulty	sweating
swine flu	swollen bone
swollen lymph glands	swollen nail
swollen spleen	syphilis
tearing in children	temperature symptoms
temple itch	temporal itch
testicle lump	testicle swelling
testicular atrophy	testicular pain
testis rash	tetanus
thigh itch	thigh paralysis
thigh rash	thigh weakness
thoracic blister	thoracic numbness
thoracic rash	thoracic redness
thoracic tingling	thoracic vertebrae blister
thoracic vertebrae itch	thoracic vertebrae rash
thoracic vertebrae spasm	thoracic wall blister
thoracic wall infection	thoracic wall inflammation
thoracic wall weakness	thorax infection
thorax inflammation	thorax weakness
throat bleeding	throat blister
throat infection	throat pain
throat rash	throat ulcer
thumb rash	thyroid enlargement
tingling	tiredness
toe rash	tongue ulcers
toxoplasmosis	tracheal stenosis
travellers diarrhoea	trichinella spiralis
trichinella spiralis infection	trichinosis
trismus	trunk rash
tuberculosis	tularaemia
typhoid and paratyphoid fever	typhus
	Medical condition: continued on next page

Table A.5 Medical condition terms consisting of infectious diseases, their synonyms, pathogens and symptoms in English, which was manually constructed by M-Eco domain experts.

ulcer	underarm rash
unequal motor movement	upper arm rash
urethral discharge	urogenital triangle rash
urosepsis	vaginal rash
varizellen	vertebral blister
vertebral itch	vertebral rash
vertebral spasm	vibrio cholerae o 1 and o 139
viral haemorrhagic fever	viral hemorrhagic fever
virushepatitis a	virushepatitis b
virushepatitis d	virushepatitis e
visible bleeding	vision changes
vocal cord paralysis	voice symptoms
voicebox deformity	voicebox infection
voicebox lump	voicebox redness
voicebox ulcer	vomiting
vomiting in children	vulva rash
vulval area rash	watery diarrhea
watery stool	weakness
weight loss	wet cough
wheezing in children	whooping cough
wrist pain	wrist rash
wrist swelling	yellow fever
yellow fever virus	yersinia
yersinia enterocolitica	yersinia pestis
yersiniosis	

Example Relevant and Non-Relevant Sentences

This Appendix present example sentences and the their corresponding sentence position for sentences that have been labeld with weak labeling for the auxilary domains of ProMED-mail. Additional examples are downloadable from the following web address: http://pharos.l3s.uni-hannover.de/~stewart/.

- Table B.1 shows examples of revelant sentences obtained with weak labeling from ProMED-mail auxilary domain.
- Table B.2 shows examples non-revelant and non-medical condition bearing sentences obtained with weak labeling from ProMED-mail auxilary domain.

Table B.1 Examples of relevant weakly labeled sentences from ProMED-mail without temporal filtering.

Pos.	ProMED-mail Revelant / Non-Relevant Sentences
0	Cap Verde: Cholera has been reported every week since November 1994.
0	The number of cases of dengue and dengue hemorrhagic fever continued to rise in November, and the epidemic spread to new areas in the Americas.
0	Cholera has also reported in Kwara 272 cases\/77 deaths , Niger 304\/97 , Ondo no report , and Oyo 215\/7 .
1	No new cases of Ebola hemorrhagic fever have been reported in Gabon since the death of the last case on 12 March 1996.
1	Introduction: last 17 April 1996, a local newspaper reported 2 monkeys from the Philippines died from Ebola infection in Hazleton, Alice, Texas, USA.
1	An increased number of cases of meningococcal meningitis has been registered since the beginning of February 1996, first in the District of Dioila, followed by the District of Bamako.
2	Other countries reporting cholera in the past week are: Cameroon, Kenya, Liberia and Niger.
1	An epidemic of cerebrospinal meningitis has been reported in Cabo Delgado province in the Northern Region.
1	Since May, Tajikistan has faced an outbreak of typhoid fever, resulting in nearly 4 000 cases notified so far.
2	The health authorities in Cyprus have informed WHO of an outbreak of viral meningitis which has affected a total of 223 persons 193 in children under 14.
1	The Ministry of Health and Social Welfare has announced a steady upward trend in cholera figures in the city of Monrovia and its environs since April 1996.
1	EHEC infection in Sakai City has affected a total of 6 309 schoolchildren and 92 school staff members from 62 municipal elementary schools.
1	Communicable disease clinics in Bucharest have registered an increase in the number of patients with meningoencephalitis since the beginning of August.
1	A press release from the Ministry of Health on 3 September reported an additional 22 cases of meningitis and viral meningoencephalitis hospitalized in the communicable disease hospitals in Bucharest.

Table B.2 Examples of non-relevant weakly labeled sentences from ProMED-mail for the precision boosting strategy and no temporal filtering.

Pos.	ProMED-mail Non-Relevant Sentences
37	Other possible infectious agents include Classical swine fever African swine fever swine influenza anthrax more acute symptoms might be expected and others.
12	The article also mentions 34 outbreaks involving 222 cases in the preceding week.
38	World Wildlife Federation WWF Canada: Toxics: EDCs Canada Web Hormones.
7	This outbreak occurred within the observation zone set up around the zone that was declared – on 22 Apr 2002 – infected with CSF in wild boar see Disease Information 15 17 55 dated 26 Apr 2002.
15	There are also increasing problems with chloroquine resistance in the state and apart from increased transmission as discussed in the article increasing resistance to chloroquine without a switch to 2nd-line drugs like mefloquine Lariam and sulfadoxine\/pyrimethamine Fansidar could also be a factor in the mortalities.
28	In September levels rose to 23 ppm which was above the action level and then from 40 to 94 ppm.
14	Trypanosomiasis African - Dem.
10	But what is needed now is not opinions but good science comprehensive epidemiology and a tremendous amount of work by USDA officials.
23	These birds were kept in a separate building and did not show any symptoms which is why it was regarded as less risky to cull them later.
16	It clearly can occur anywhere around the world as the list of countries where this has happened in swine now includes a pandemic type range of countries including Canada Argentina Australia Singapore United Kingdom Ireland Norway Japan Iceland USA Taiwan Indonesia Finland and now Italy see references below.
7	This is the 4th recall of olives this month March 2007 regarding botulism risk.
55	The name is said to come from comparison with a corpulent female tavern keeper “ ale-wife ”.
19	The mortality reported in the press reports of the current HFMD outbreak in the city of Linyi in Shandung province 26 deaths among 292 children if accurate would be an unprecedented event.
8	Requests to contribute a paper or poster as well as accompanying abstracts should be received by no later than 31 Jan 2005.
42	Viable bacteria can also be found for weeks to months in the carcasses and hides of infected animals and in fomites including grain dust straw water soil and bedbugs.
25	“ We hope these initiatives will set the stage for other countries to adopt similar approaches to the release of Influenza virus sequence data that they manage ” Cox said.
97	Significant points are that during outbreak investigations up to 10 times more cases have been identified than those notified to the local health authorities.
34	Few anglers eat muskies anyway said Greg Van Assche a fishing guide who says he has caught more than 3000 muskies in his lifetime.

Bibliography

[AA91] Steven Abney and Steven P. Abney. Parsing by chunks. In *Principle-Based Parsing*, pages 257–278. Kluwer Academic Publishers, 1991.

[ABK+07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *In 6th Internation Semantic Web Conference, Busan, Korea*, pages 11–15. Springer, 2007.

[AC75] Alfred V. Aho and Margaret J. Corasick. Efficient string matching: an aid to bibliographic search. *Commun. ACM*, 18(6):333–340, June 1975.

[AG00] Eugene Agichtein and Luis Gravano. Snowball: extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, DL '00, pages 85–94, New York, NY, USA, 2000. ACM.

[AHB+93] D Appelt, J Hobbs, J Bear, D Israel, M Kameyama, and M. Tyson. Fastus: a finite-state processor for information extraction from real-world text. In *Proc. 13th Intl Joint Conf. on Artificial Intelligence*, pages 1172–1178, 1993.

[AMM11] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, 2011.

[ANC07] Andrew Arnold, Ramesh Nallapati, and William W. Cohen. A comparative study of methods for transductive transfer learning. In *ICDMW*

'07: Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, pages 77–82, Washington, DC, USA, 2007. IEEE Computer Society.

[AP08] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December 2008.

[Ban00] Michele Banko. *Open Information Extraction for the Web*. PhD thesis, 2000.

[BB98] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, 1998.

[BCS+07] Michele Banko, Michael J. Cafarella, Stephen Soderl, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *In IJCAI*, pages 2670–2676, 2007.

[BE08] Michele Banko and Oren Etzioni. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[BGCG+09] Jordan Boyd-Graber, Jonathan Chang, Sean Gerrish, Chong Wang, and David Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*, 2009.

[BM98] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM.

[BMC11] Zalán Bodó, Zsolt Minier, and Lehel Csató. Active learning with clustering. *Journal of Machine Learning Research - Proceedings Track*, 16:127–139, 2011.

[Bri99] Sergey Brin. Extracting patterns and relations from the world wide web. In *Selected papers from the International Workshop on The World Wide Web and Databases*, pages 172–183, London, UK, 1999. Springer-Verlag.

[CBW+10] Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 101–110, New York, NY, USA, 2010. ACM.

[CC10] Hutchatai Chanlekha and Nigel Collier. Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports. *Journal of biomedical semantics*, 1(1):3, 2010.

[CCD09] Mike Conway, Nigel Collier, and Son Doan. Using hedges to enhance a disease outbreak report text mining system. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 142–143, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[CD01] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14*, pages 625–632. MIT Press, 2001.

[CDK+08] Nigel Collier, Son Doan, Ai Kawazeo, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. "biocaster: detecting public health rumors with a web-based text mining system", bioinformatics, oxford university press, 2008. [90]DOI: 10. 1093 / bioinformatics / btn 534.

[CDKC09] Mike Conway, Son Doan, A. Kawazoe, and Nigel Collier. Classifying disease outbreak reports using n-grams and semantic features. *International Journal of Medical Informatics*, 78(12):e47–e58, 2009.

[CJ09] Zheng Chen and Heng Ji. Can one language bootstrap the other: a case study on event extraction. In *SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 66–74, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[CKCC09] Nigel Collier, Ai Kawazoe, Hutchatai Chanlekha, and Michael Conway. Annotation of public health events in biocaster a concise guide for annotators, 2009.

[CKJ+06] Nigel Collier, Ai Kawazoe, Lihua Jin, Mika Shigematsu, Dinh Dien, Roberto Barrero, Koichi Takeuchi, and Asanee Kawtrakul. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. 40(3):405–413, December 2006.

[CMS07] J. R. Curran, T. Murphy, and B. Scholz. Minimising semantic drift with mutual exclusion bootstrapping. *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, pages 172–180, 2007.

[CR03] Chad Cumby and Dan Roth. On kernel methods for relational learning. In *In Proc. of the International Conference on Machine Learning*, pages 107–114, 2003.

[CSN11] Nigel Collier, Nguyen Truong Son, and Ngoc Mai Nguyen. Omg u got flu? analysis of shared health messages for bio-surveillance. *CoRR*, abs/1110.3089, 2011.

[CST00] Nello Cristianini and John Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA, 2000.

[Cul10] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, 2010.

[DAGSN10] Ernesto Diaz-Aviles, Mihai Georgescu, Avaré Stewart, and Wolfgang Nejdl. Lda for on-the-fly auto tagging. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 309–312, New York, NY, USA, 2010. ACM.

[Das11] Sanjoy Dasgupta. Two faces of active learning. *Theor. Comput. Sci.*, 412(19):1767–1781, April 2011.

[DAS12] Ernesto Diaz-Aviles and Avaré Stewart. Tracking twitter for epidemic intelligence: case study: Ehec/hus outbreak in germany, 2011. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, pages 82–85, New York, NY, USA, 2012. ACM.

[DDAD+11] Kerstin Denecke, Ernesto Diaz-Aviles, Peter Dolog, Tim Eckmanns, Marco Fisichella, Ricardo Gomez-Lage, Jens Linge, Pavel Smrz, and Avaré Stewart. The medical ecosystem [m-eco] project: Personalized event-based surveillance. In *Proc. of International Meeting on Emerging Diseases and Surveillance (IMED 2011), Vienna, Austria, February 4-7, 2011*, 2011.

[DDCM12] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 469–478, New York, NY, USA, 2012. ACM.

[DDS+10] Kerstin Denecke, Peter Dolog, Pavel Smrz, Jens Linge, Wolfgang Nejdl, and Avaré Stewart. Using web data in the medical domain. In *Proc. of 1st International Workshop on Web Science and Information Exchange in the Medical Web, MedEx 2010, Raleigh, NC, USA, April 26, 2010*, 2010.

[DKCC08] S Doan, A Kawazoe, M Conway, and N Collier. Towards role-based filtering of disease outbreak reports. *J Biomed Inform*, Dec 2008.

[dMM08] Marie-Catherine de Marneffe and Christopher D. Manning. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[Dre12] Mark Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84, 2012.

[ECD+04] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 100–110, New York, NY, USA, 2004. ACM.

[EFC+11] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. Open information extraction: The second generation. In Toby Walsh, editor, *IJCAI*, pages 3–10. IJCAI/AAAI, 2011.

[FGM05] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[FKG+09] Ariel Fuxman, Anitha Kannan, Andrew B. Goldberg, Rakesh Agrawal, Panayiotis Tsaparas, and John Shafer. Improving classification accuracy using automatically extracted training data. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1145–1154, New York, NY, USA, 2009. ACM.

[FRP09] Hannes Korte Frank Reichartz and Gerhard Paass. Dependency tree kernels for relation extraction from natural text. *Lecture Notes in Computer Science*, pages 270–285, 2009.

[FSCD11] Marco Fisichella, Avaré Stewart, Alfredo Cuzzocrea, and Kerstin Denecke. Detecting health events on the social web to enable epidemic intelligence. In *SPIRE*, pages 87–103, 2011.

[FSDN10] Marco Fisichella, Avaré Stewart, Kerstin Denecke, and Wolfgang Nejdl. Unsupervised public health event detection for epidemic intelligence. In *CIKM 2010: 19th ACM Conference on Information and Knowledge Management*, New York, NY, USA, 2010. ACM.

[FSN] Marco Fisichella, Avaré Stewart, and Wolfgang Nejdl. Unified approach to retrospective event detection for event based epidemic intelligence (under review). *IEEE Trans. Knowl. Data Eng.*

[Gha04] Zoubin Ghahramani. Unsupervised learning. In *Advanced Lectures on Machine Learning*, pages 72–112. Springer-Verlag, 2004.

[GHY02] Ralph Grishman, Silja Huttunen, and Roman Yangarber. Information extraction for enhanced access to disease outbreak reports. *J. of Biomedical Informatics*, 35(4):236–246, 2002.

[GS07] Mark A. Greenwood and Mark Stevenson. A task-based comparison of information extraction pattern models. In *Proceedings of the Workshop on Deep Linguistic Processing*, DeepLP '07, pages 81–88, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[GSO+11] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[HBV10] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 7(2/3):107–145, 2010.

[HNW+09] D Hartley, NP Nelson, R Walters, R Arthur, R Yangarber, L Madoff, JP Linge, A Mawudeku, N Collier, J Brownstein, G Thinus, and N Lightfoot. The landscape of international event-based biosurveillance. *Emerging Health Threats*, 2009.

[KBHT09] Mikaela Keller, Michael Blench, and et.al. Herman Tolentino. Use of unstructured event-based reports for global infectious disease surveillance. 15(5), May 2009.

[KFN10] Christian Kohlschüter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu, editors, *WSDM*, pages 441–450. ACM, 2010.

[KMN09] Sandra Kübler, Ryan T. McDonald, and Joakim Nivre. *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2009.

[KRS12] Nattiya Kanhabua, Sara Romano, and Avaré Stewart. Identifying relevant temporal expressions for real-world events. In and, editor, *SIGIR 2012 Workshop on Time-aware Information Access (TAIA'2012)*, TAIA'2012, 2012.

[KRSN12] Nattiya Kanhabua, Sara Romano, Avaré Stewart, and Wolfgang Nejdl. Supporting temporal analytics for health-related events in microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2686–2688, New York, NY, USA, 2012. ACM.

[LAJ01] Adenike M. Lam-Adesina and Gareth J. F. Jones. Applying summarization techniques for term selection in relevance feedback. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–9, New York, NY, USA, 2001. ACM.

[Lan06] Eric Langford. Quartiles in elementary statistics. *Journal of Statistics Education*, 14(3), 2006.

[LC12] Vasileios Lampos and Nello Cristianini. Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.*, 3(4):72:1–72:22, September 2012.

[LG94] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 3–12, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[LNGB12] A. Lyon, M. Nunn, G. Grossel, and M. Burgman. Comparison of web-based biosecurity intelligence systems: Biocaster, epispider and healthmap. *Transboundary and Emerging Diseases*, 59(3):223–232, 2012.

[LSW+09] Jens P Linge, Ralf Steinberger, T P Weber, R Yangarber, E van der Goot, D H Al Khudhairy, and N I Stilianakis. Internet surveillance systems for early alerting of health threats. *Eurosurveillance*, 14(13), 2009.

[LT97] Ray Liere and Prasad Tadepalli. Active learning with committees for text categorization. In *In proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 591–596, 1997.

[LWLM05] Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma. A probabilistic model for retrospective news event detection. In *SIGIR*, pages 106–113, 2005.

[MBSJ09] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[MC09] Tara McIntosh and James R. Curran. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 396–404, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[McI10] Tara McIntosh. Unsupervised discovery of negative categories in lexicon bootstrapping. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 356–365, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Mos04] Alessandro Moschitti. A study on convolution kernels for shallow semantic parsing. In *ACL '04*, page 335, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[Mos06] Alessandro Moschitti. Making tree kernels practical for natural language learning. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[MS09] Prem Melville and Vikas Sindhwani. Active dual supervision: reducing the cost of annotating examples and features. In *HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 49–57, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[NHN06] Joakim Nivre, Johan Hall, and Jens Nilsson. MaltParser: A data-driven parser-generator for dependency parsing. In *Proc. of LREC-2006*, 2006.

[NJT07] Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. Using cluster validation criterion to identify optimal feature subset and cluster number for document clustering. *Information Processing and Management*, 43(3):730–739, 2007.

[NS04] Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 79–, New York, NY, USA, 2004. ACM.

[NSC10a] M. Naughton, N. Stokes, and J. Carthy. Sentence-level event classification in unstructured texts. *Information Retrieval*, 13(2):132–156, April 2010.

[NSC10b] M. Naughton, N. Stokes, and J. Carthy. Sentence-level event classification in unstructured texts. *Inf. Retr.*, 13(2):132–156, 2010. key paper.

[PCKC06] C Paquet, D Coulombier, R Kaiser, and M Ciotti. Epidemic intelligence: a new framework for strengthening disease surveillance in europe. *Euro Surveillence*, 11(12):212–214, 2006.

[PD11a] Michael J. Paul and Mark Dredze. A model for mining public health topics from twitter. Technical report, Johns Hopkins University, 2011.

[PD11b] Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Proceedings of ICWSM'2011*, 2011.

[PNH08] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 91–100, New York, NY, USA, 2008. ACM.

[PP04] Amruta Purandare and Ted Pedersen. Word sense discrimination by clustering contexts in vector and similarity spaces. pages 41–48, 2004.

[PR09] Siddharth Patwardhan and Ellen Riloff. A unified model of phrasal and sentential evidence for information extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 151–160, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[PY10] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010.

[RKP10] Frank Reichartz, Hannes Korte, and Gerhard Paass. Semantic relation extraction with kernels over typed dependency trees. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 773–782, New York, NY, USA, 2010. ACM.

[RMD11] Daniel Ramage, Christopher D. Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

[SD10a] Avaré Stewart and Kerstin Denecke. Can promed-mail bootstrap blogs? automatic labeling of victim-reporting sentences. In *Proc. of 1st International Workshop on Web Science and Information Exchange in the Medical Web, MedEx 2010, Raleigh, NC, USA, April 26, 2010*, 2010.

[SD10b] Avaré Stewart and Kerstin Denecke. Using promed mail and medworm blogs for cross-domain pattern analysis in epidemic intelligence. In *Proc. of 13th World Congress on Medical and Health Informatics Medinfo 2010, 12-15th September 2010, Cape Town, South Africa*, 2010.

[SDA12] Avaré Stewart and Ernesto Diaz-Aviles. Epidemic intelligence: For the crowd, by the crowd. In *ICWE*, pages 504–505, 2012.

[SDAN08] Avaré Stewart, Ernesto Diaz-Aviles, and Wolfgang Nejdl. Mining user profiles to support structure and explanation in open social networking. *CoRR*, abs/0812.4461, 2008.

[SDAN+09] Avaré Stewart, Ernesto Diaz-Aviles, Wolfgang Nejdl, Leandro Balby Marinho, Alexandros Nanopoulos, and Lars Schmidt-Thieme. Cross-tagging for personalized open social networking. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09, pages 271–278, New York, NY, USA, 2009. ACM.

[SDN10] Avaré Stewart, Kerstin Denecke, and Wolfgang Nejdl. Cross-corpus textual entailment for sublanguage analysis in epidemic intelligence. In *LREC*, 2010.

[SDSS12] Mustafa Sofean, Kerstin Denecke, Avaré Stewart, and Matthew Smith. Medical case-driven classification of microblogs: characteristics and annotation. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, pages 513–522, New York, NY, USA, 2012. ACM.

[Set09] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

[SFD] Avar Stewart, Marco Fisichella, and Kerstin Denecke. Detecting public health indicators from the web for epidemic intelligence. In *3rd*

International ICST Conference on Electronic Healthcare for the 21st century (eHealth 2010).

[SFvdG+08a] Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter, and Roman Yangarber. Text mining from the web for medical intelligence. 2008.

[SFvdG+08b] Ralf Steinberger, Flavio Fuart, Erik van der Groot, Clive Best, Peter von Etter, and Roman Yangarber. Text mining from the web for medical intelligence. *Mining Massive Data Sets for Security*, 19:295–310, 2008.

[SG07] Mark Steyvers and Tom Griffiths. *Probabilistic Topic Models*. Lawrence Erlbaum Associates, 2007.

[SG09] Mark Stevenson and Mark Greenwood. Dependency pattern models for information extraction. *Research on Language & Computation*, 7(1):13–39, March 2009.

[SG12] Jannik Strötgen and Michael Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 2012.

[SKdQ10] Martin Szomszor, Patty Kostkova, and Ed de Quincey. #swineflu: Twitter predicts swine flu outbreak in 2009. In *Proceedings of eHealth'2010*, 2010.

[SKL11] Martin Szomszor, Patty Kostkova, and Connie St Louis. Twitter informatics : Tracking and understanding public reaction during the 2009 swine flu pandemic. In *Proceedings of WI-IAT'2011*, 2011.

[SKR+] Avaré Stewart, Nattiya Kanhabua, Sara Romano, Ernesto Diaz-Aviles, and Wolfgang Nejdl. Leveraging social media for epidemic intelligence: Challenges and opportunies (under submission). In *ACM SIGIR Workshop on Health Search and Discovery: Helping Users and Advancing Medicine*.

[SKW07] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[SN11a] Avaré Stewart and Wolfgang Nejdl. Exploiting the language of moderated sources for cross-classification of user generated content. In *WEBIST*, pages 571–576, 2011.

[SN11b] Avaré Stewart and Wolfgang Nejdl. Self-supervised learning for medical web disease reporting events detection. In *Proc. of ACM WebSci'11, June 14-17 2011, Koblenz, Germany*, 2011.

[SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of WWW'2010*, 2010.

[SS11] A. Stewart and M. Smith. User centric public health event detection within social medical ecosystems. In *Proceedings of the 5th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2011)*, 2011.

[SSN11] Avaré Stewart, Matthew Smith, and Wolfgang Nejdl. A transfer approach to detecting disease reporting events in blog social media. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 271–280, New York, NY, USA, 2011. ACM.

[TK02] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, March 2002.

[TSZ07] Anthony Tomasic, Isaac Simmons, and John Zimmerman. Learning information intent via observation. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 51–60, New York, NY, USA, 2007. ACM.

[vEHV+10a] Peter von Etter, Silja Huttunen, Arto Vihavainen, Matti Vuorinen, and Roman Yangarber. Assessment of utility in web mining for the domain of public health. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 29–37, Los Angeles, California, USA, June 2010. Association for Computational Linguistics.

[vEHV+10b] Peter von Etter, Silja Huttunen, Arto Vihavainen, Matti Vuorinen, and Roman Yangarber. Assessment of utility in web mining for the domain of public health. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, Louhi '10, pages 29–37, Morristown, NJ, USA, 2010. Association for Computational Linguistics. key paper.

[VHSP10] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of CHI'2010*, 2010.

[VLC94] Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the vc-dimension of a learning machine. *Neural Computation*, 6:851–876, 1994.

[WHW09] Daniel S. Weld, Raphael Hoffmann, and Fei Wu. Using wikipedia to bootstrap open information extraction. *SIGMOD Rec.*, 37:62–68, March 2009.

[WW10] Fei Wu and Daniel S. Weld. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Yan06] Roman Yangarber. Verification of facts across document boundaries. *In Proceedings International Workshop on Intelligent Information Access*, 2006.

[YCB+99] Yiming Yang, Jaime Carbonell, Ralf Brown, Tom Pierce, Brian T. Archibald, and Xin Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14:32–43, 1999.

[YvES08] Roman Yangarber, Peter von Etter, and Ralf Steinberger. Content collection and analysis in the domain of epidemiology. In *International Workshop on Describing Medical Web Resources: the 21st International Congress of the European Federation for Medical Informatics*, 2008.

[yZhL09] Pei ying Zhang and Cun he Li. Automatic text summarization based on sentences clustering and extraction. *Computer Science and Information Technology, International Conference on*, 0:167–168, 2009.

[ZAR02] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 71–78, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[Zha08] Yi Zhang. *Automatic Extraction of Outbreak Information from News*. PhD thesis, University of Illinois at Chicago, 2008.

[Zhu05] Xiaojin Zhu. Semi-Supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005.

[ZNL+09] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. Statsnowball : a statistical approach to extracting entity. *Methods*, pages 101–110, 2009.