

Can ProMED-mail Bootstrap Blogs? Automatic Labeling of Victim- reporting Sentences

Avaré Stewart, Kerstin Denecke

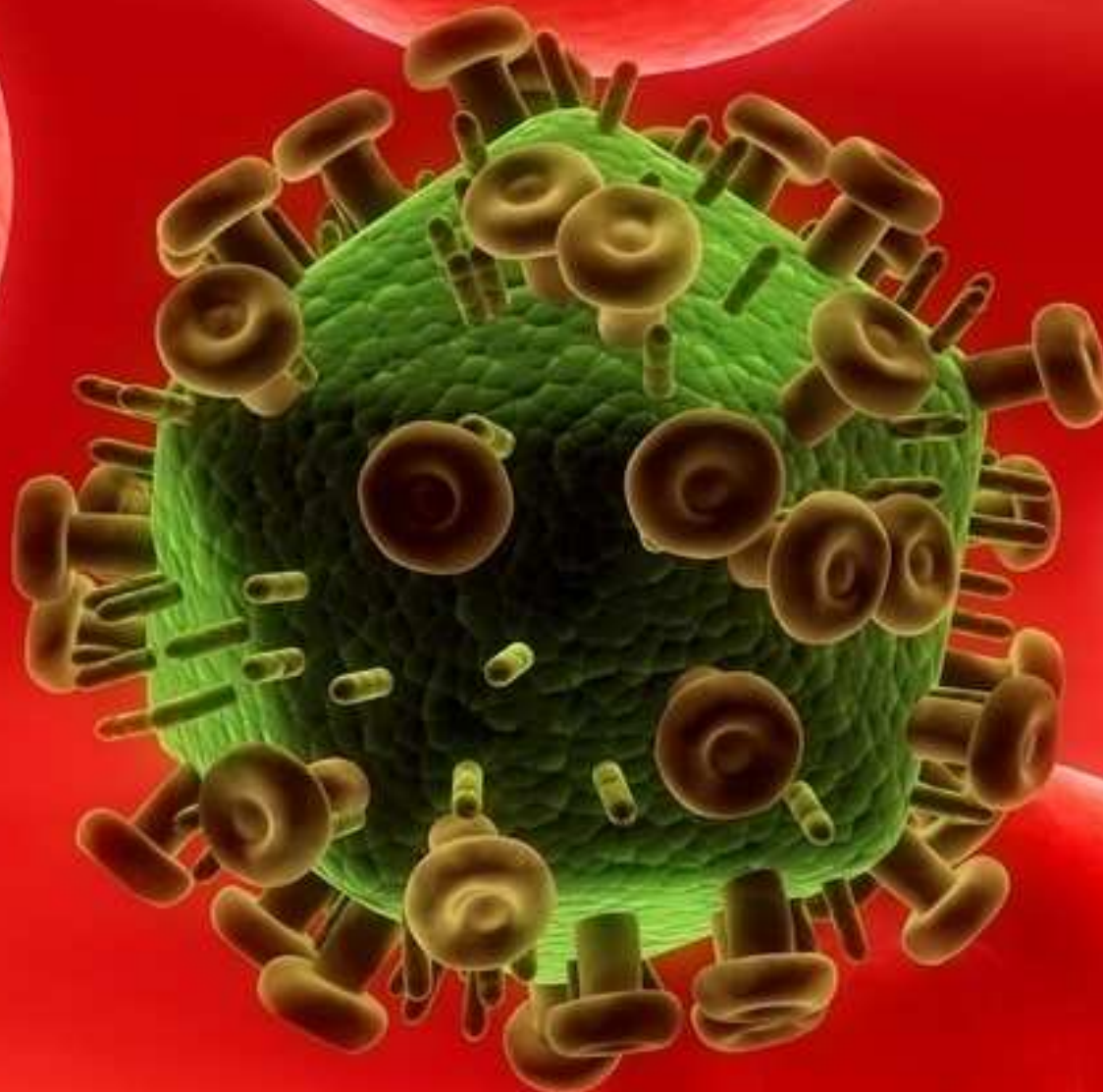
26.04.2010

Outline

- Motivation
- Challenges of Social Media Data
- Approach
- Experiments
- Conclusion & Next Steps

Emergence of Infectious Disease

21st Century



- **346 infectious diseases**
- **220 countries**
- **3 new diseases every 2 years**
- **new infecting animal every week**

Medical Bloggers

PKIDs Blog
Parents of Kids with Infectious Diseases

.....The pediatrician took one look at Matthew and asked me this question: “*Was your son vaccinated ?* “

The answer was a resounding NO!

....we decided that **if we could help one family** not go through what we went through, **we would do it.**

EQUID BLOG

INFORMATION & INSIGHT ON EQUINE INFECTIOUS DISEASES

Scott Weese : **Associate Professor** and Public Health & Zoonotic disease **microbiologist**

Global Infectious Diseases and Epidemiology Online Network

twitter



GIDEON

Hey there! **gideononline** is using Twitter.

Rubor Dolor Calor Tumor

Commentaries on Infectious Diseases (with Rat Holes) by an Inveterate Persiflagger

Mark Crislip MD : practicing Infectious **Disease specialist**; **Chief** of Infectious Diseases

M-Eco
Medical EcoSystem
27/04/10



Challenges of Social Media Data

- Huge amount of data available
 - Irrelevant information vs. relevant
- Subjective content
 - Information vs. opinion
- Use of specific language
 - Medical language vs. consumer health vocabulary
 - Common language
- Different styles of writing
 - Abbreviations, writing errors, emoticons..

Detecting Health Events

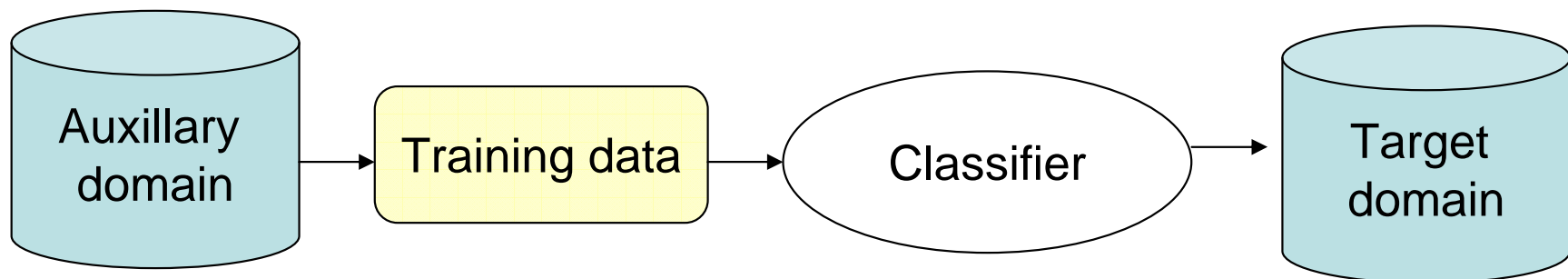
- Annotation
 - Identifying relevant sentences
- Event Extraction
 - Identifying relevant facts
- Event Aggregation
 - Aggregating information on the same event

Subproblem: Annotation

- Classification Task:
 - Identifying relevant sentences in informal media
- Problem:
 - Labeled data is unavailable
 - Manual annotation is expensive.
- Proposed Solution:
 - Semi-supervised learning approach
 - Use for structural data for producing training data
 - Exploit Bootstrapping algorithm to reduce noise

Approach

1. Collect training data from documents of auxiliary domain (weakly labeling)
2. Build classifier in a bootstrap process
3. Apply classifier to documents of a target domain



1. Weakly labeling

Bovine brucellosis - Fiji Islands

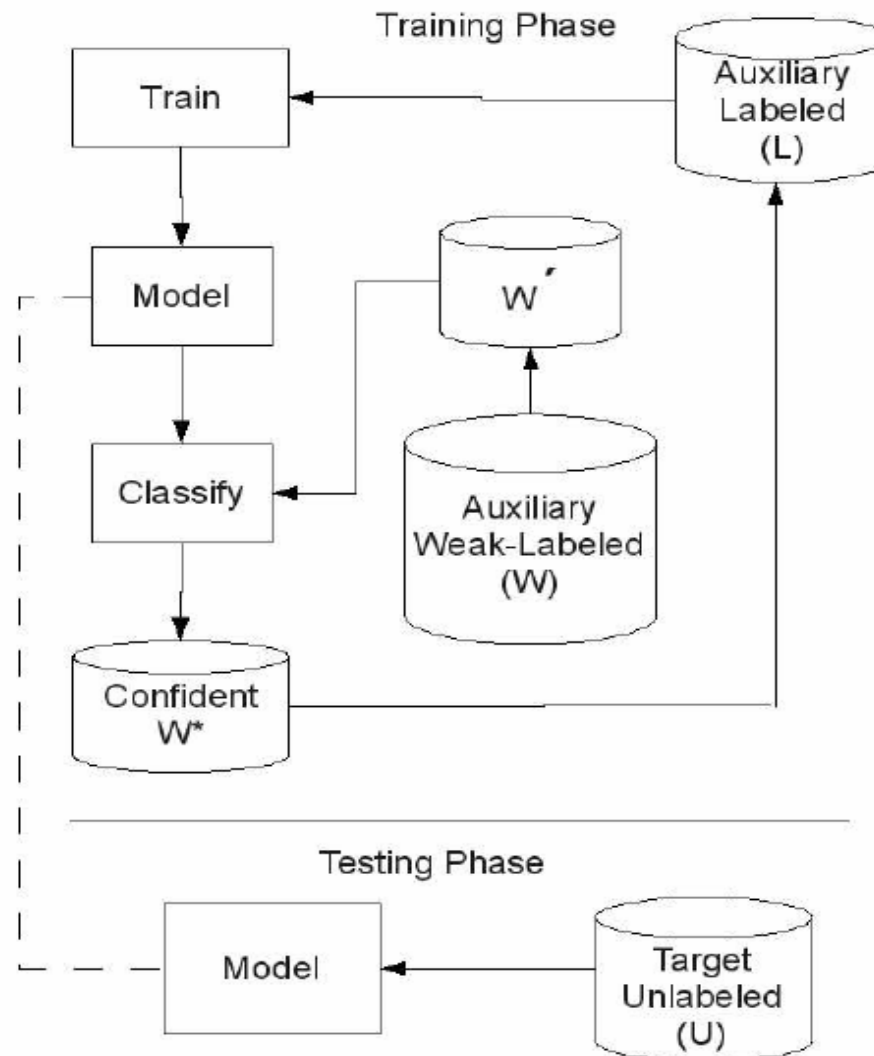
A syndrome of abortion, weak calves and increasing occurrence of retained afterbirth in a number of dairy cattle herds on the main island of Fiji (Viti Levu) has been investigated by the Fiji Ministry of Primary Industries. Reports indicate the syndrome has been occurring since February 2009.

Preliminary analysis for brucellosis of 2429 sera collected from 9 herds in June 2009 by Rose Bengal test identified 236 strongly reactive sera from 6 of the herds tested. A sub-sample of RBT reactive sera was subsequently analysed with the assistance of the Queensland Department of Primary Industries in Australia by both CFT and ELISA for Brucella. Serology results were 35/50 sera positive by CFT (titres ranging from 1/16 to 1/4096) and 48/50 sera positive by ELISA. Culture results are currently pending. Further serological analysis is to be conducted. Pre-emptive slaughter of reactor cattle has been undertaken in a number of herds. Investigation of the source of the outbreak is ongoing. Fiji conducted a test and slaughter program for bovine brucellosis in the 1980s and subsequently declared freedom from the disease in 1989. Fiji is believed to be free from Brucella melitensis. Brucella suis is known to be present in some sectors of the pig population but recent survey information is not available.



2. Build Classifier, 3. Classify Data

Bootstrapping
process



Typical Bootstrapping Process

- Inferring labels from Seed: used to automatically label data.
- *Small* Amount high-quality labeled examples
- *Larger* amount of unlabelled data.
- Classifier incrementally learns a model:
 - Model is applied to unlabeled data;
 - Retaining most confidently classified examples
 - Added to the growing pool of labeled data

Experiments

- Objectives
 - Assess quality of the classifier
- Data
 - Auxillary Domain: ProMed-Mail, WHO
 - Target Domain: AvianFluDiary

Source	Years	No. of Documents	No. of Sentences
AvianFluDiary	2006-2009	4249	100890
ProMed-Mail	2002-2009	13369	22170
WHO	1996-2009	1531	16213

Experiment: Quality

- *Scenario 1:* ProMed as auxiliary data, state of the art settings for the bootstrap
- *Scenario 2:* applied to the bottom-1 sentences additionally filtered (i.e.: eliminating http containing sentences)
- *Scenario 3:* WHO as auxiliary data, standard settings
- *Scenario 4:* Filter based on presence of named entities (medical condition, location, etc.)

Experimental: Results

Scenario	Precision	Recall	Accuracy
1	.77	.45	.57
2	.71	.66	.69
3	.75	.22	.34
4	.80	.40	.53

- Best Precision (Scenario 4):
 - Use of named entity to filter sentences
- Best Accuracy (Scenario 2):
 - Filter the “bad negative” examples
- Best Tradeoff : (Scenario 2):
 - F-Measure = .68

Conclusions

Can ProMED-mail Bootstrap Blogs for Automatic Labeling of Victim-reporting Sentences ?

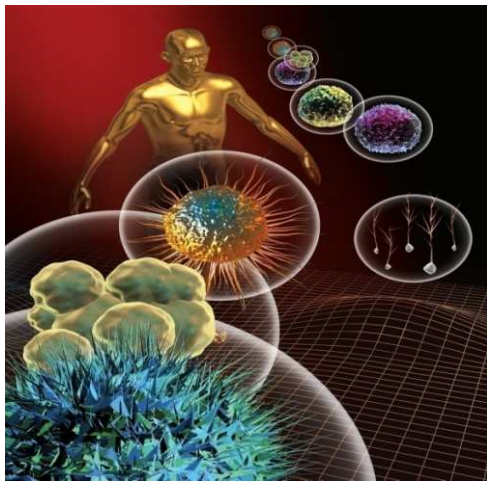
1. Yes: achievable precision = .80 ,accuracy = .60
2. No human effort labeling a training set: through ***weak labeling*** (position of sentence within a document)
3. No Feature Engineering: Kernel Based Support Vector Machine

Future Work

- Robust experiments
 - more blogs
 - increasing values for top-N
- Comparison to other semi-or unsupervised learning approaches

Can ProMED-mail Bootstrap Blogs?

Automatic Labeling of Victim-reporting Sentences



Thank you for your attention!