The Center For Language and Speech Processing
at the Johns Hopkins University

Automated Event Extraction and Named Entity Recognition in the Domain of Veterinary Medicine

K-State Lab for Knowledge Discovery in Databases

Svitlana Volkova, PhD Student, Johns Hopkins University

MOTIVATION

Global epidemic surveillance is an essential task for national biosecurity management and bioterrorism prevention.

Animal Infectious Disease Outbreaks

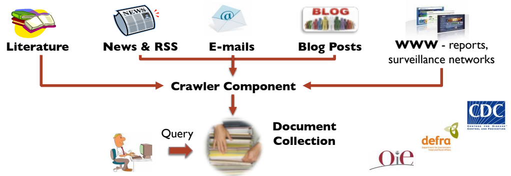
influence on international travel and trade

cause economic crises, political instability

can cause loss of human life (61% of animal disease)

The goal is to protect the public from major health threads by developing the **framework for epidemiological analytics** that allows automated data collection, sharing, management, modeling and analysis in the domain of emerging infectious diseases.

DATA



PROBLEM FORMULATION

- Introduce the following functionality to the framework for epidemiological analytics:
  - Domain-specific and domain-independent named entity recognition: **ontology-based and using syntactic features**:
    - ✓disease names (e.g. "foot and mouth disease");
    - ✓viruses (e.g. "picornavirus") and serotypes (e.g. "Asia-1");
    - ✓species (e.g. "sheep", "cattle");
    - ✓locations (e.g. "United Kingdom", "eastern provinces of Shandong and Jiangsu, China" – different level of granularity);
    - ✓dates in different formats including special cases (e.g. "last Tuesday", "two month ago").
  - Automated animal **disease event extraction and classification** from unstructured web data.

RESEARCH QUESTIONS

How do we construct an ontology of animal disease names, their synonyms and corresponding viruses and learn semantic relationships between them?

How should we resolve location disambiguation "Rabies in Isle of Wight", geo-tag in Virginia, USA or UK?

How should we merge extracted entities into corresponding event tuples?

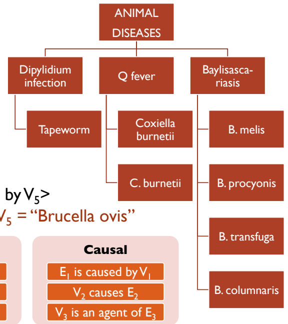
How do we classify extracted event tuples in order to reason about event confidence?

ONTOLOGY-BASED RECOGNITION

Synonym relationships
<E_1 is a kind of E_2>
E_1 = "swine influenza" is a kind of
E_2 = "swine fever"

Hyponym relationships
<E_3 and E_4 are diseases>
E_3 = "anthrax", E_4 = "yellow fever" are diseases

Causal relationships <E_5 is caused by V_5>
E_5 = "Ovine epididymitis" is caused by V_5 = "Brucella ovis"

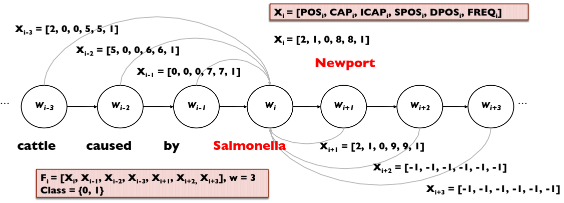


Synonymic	Hyponymic	Causal
E_1 is a E_2 and E_3	D such as E_1, E_2	E_1 is caused by V_1
E_1 also known as E_2	D for example E_1	V_2 causes E_2
E_1 is also called E_2	D including E_1, E_2	V_3 is an agent of E_3

APPLYING SYNTACTIC FEATURES

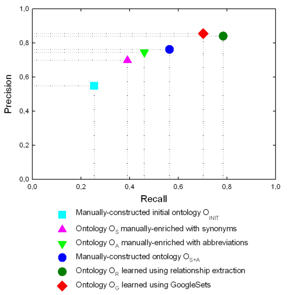
"Severe disease in dairy cattle caused by Salmonella Newport"

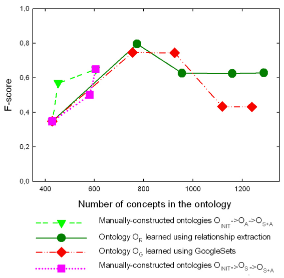
POS = [NNP, IN, NNS, VBN, …] = [2, 0, 2, 5, …]


X_i = [POS_i, CAP_i, $ICAP_i$, $SPOS_i$, $DPOS_i$, $FREQ_i$]
X_{i-3} = [2, 0, 0, 5, 5, 1]
X_{i-2} = [5, 0, 0, 6, 6, 1]
X_{i-1} = [0, 0, 0, 7, 7, 1]
X_i = [2, 1, 0, 8, 8, 1]
Newport
…
w_{i-3}
w_{i-2}
w_{i-1}
w_i
w_{i+1}
w_{i+2}
w_{i+3}
…
cattle
caused
by
Salmonella
X_{i+1} = [2, 1, 0, 9, 9, 1]
X_{i+2} = [-1, -1, -1, -1, -1, -1]
X_{i+3} = [-1, -1, -1, -1, -1, -1]
F_i = [X_i, X_{i-1}, X_{i-2}, X_{i-3}, X_{i+1}, X_{i+2}, X_{i+3}], w = 3
Class = {0, 1}

F_i = [2, 1, 0, 8, 8, 1, 0, 0, 0, 7, 7, 1, 5, 0, 0, 6, 6, 1, 2, 0, 0, 5, 5, 1, 2, 1, 0, 9, 9, 1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1], Class = [1]

RESULTS





Classifier	W_i+1	W_i-1	W_i+/-1	W_i+/-3	W_i+5	W_i-5	W_i+/-5	W_i+/-7
Random Forest	0.771	0.773	0.782	0.775	0.764	0.771	0.757	0.745
AdaBoost	0.758	0.759	0.759	0.758	0.761	0.761	0.761	0.761
Naïve Bayes	0.700	0.706	0.685	0.661	0.662	0.600	0.647	0.639
Logistic	0.738	0.739	0.739	0.739	0.734	0.736	0.753	0.735

Acknowledgements: William H. Hsu, Doina Caragea, Chris Callison-Burch

EVENT EXTRACTION

Type 1: Emergent Outbreak-Related Events

- "On 2 Jun 2010, a total of 35 individuals infected with a matching strain of salmonella"

Type 2: Non-Emergent Outbreak-Related Events

- "The US saw its latest FMD outbreak in Montebello, California in 1929"

Type 3: Disease Outbreak Non-Related Events

- "A meeting on foot and mouth disease was held in Brussels on Oct 17, 2007"

Types 4 & 5: Hypothetical Events or Negation of the Events

EVENT TUPLE

$Event_i$ =< disease; date; location; species; status >

Class 1 – Susceptible Status								
healthi	popul	open	vulner	expos	respons	sign	separ	contamin
Class 2 – Infected Status								
outbreak	infect	report	confirm	affect	sick	diagnos	readi	inciner
Class 3 – Recovered Status								
destroi	burn	erad	dispos	dead	buri	slaughter	elimin	cull

Step 1: Entity Recognition

Foot-and-mouth disease$_{[DIS]}$ on **hog$_{[SP]}$** farm in **Taoyuan$_{[LOC]}$**

Taiwan's TVBS television station reports that agricultural authorities confirmed **foot-and-mouth disease$_{[DIS]}$** on a **hog$_{[SP]}$** farm in **Taoyuan$_{[LOC]}$**. On **9 Jun 2009$_{[DT]}$**, the farm's owner reported symptoms of **FMD$_{[DIS]}$** in more than 30 **hogs$_{[SP]}$**. Subsequent testing confirmed **FMD$_{[DIS]}$**. Agricultural authorities asked the farmer to strengthen immunization. The outbreak has not affected other farms. Authorities stipulated that the affected **hog$_{[SP]}$** farm may not sell pork for 2 weeks.

Step 2: Sentence Classification

+ 1. **Foot-and-mouth disease$_{[DIS]}$** on **hog$_{[SP]}$** farm in **Taoyuan$_{[LOC]}$**.
+ 2. Taiwan's TVBS television station **reports** that agricultural authorities **confirmed** **foot-and-mouth disease$_{[DIS]}$** on a **hog$_{[SP]}$** farm in **Taoyuan$_{[LOC]}$**.
+ 3. On **9 Jun 2009$_{[DT]}$**, the farm's owner **reported** symptoms of **FMD$_{[DIS]}$** in more than 30 **hogs$_{[SP]}$**.
+ 4. Subsequent testing **confirmed** **FMD$_{[DIS]}$**.
− 5. Agricultural authorities asked the farmer to strengthen immunization.
− 6. The outbreak has not affected other farms.
− 7. Authorities stipulated that the affected **hog$_{[SP]}$** farm may not sell pork for 2 weeks.

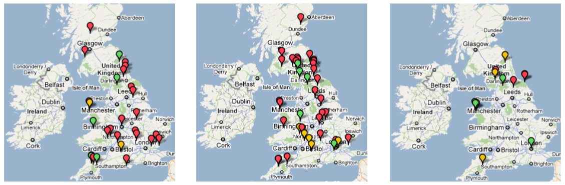
Step 3a: Tuple Generation

E_1 = <Foot-and-mouth disease, ?, Taoyuan, hog, ?>
E_2 = <Foot-and-mouth disease, ?, Taoyuan, hog, confirmed>
E_3 = <FMD, 9 Jun 2009, ?, hog, reported>
E_4 = <FMD, ?, ?, ?, confirmed>

Step 3b: Tuple Aggregation

E = <disease, date, location, species, status> =
<Foot-and-mouth disease, 9 Jun 2009, Taoyuan, hog, infected>

EVENT VISUALIZATION



2001 foot-and-mouth disease outbreak over time in United Kingdom: February, March, April