Computational Knowledge and Information Management in Veterinary Epidemiology

Svitlana Volkova, William H. Hsu
Laboratory for Knowledge Discovery in Databases
Department of Computing and Information Sciences
Kansas State University
234 Nichols Hall, Manhattan, Kansas 66506
Email: {svitlana, bhsu}@ksu.edu

***Abstract*—Monitoring of infectious animal diseases is an essential task for national biosecurity management and bioterrorism prevention. For this purpose, we present a system for animal disease outbreak analysis by automatically extracting relational information from online data. We aim to detect and map infectious disease outbreaks by extracting information from unstructured sources. The system crawls web sites and classifies pages by topical relevance. The information extraction component performs document analysis for animal disease related event recognition. The visualization component plots extracted events into GoogleMaps[1] using geospatial information and supports timeline representation of animal disease outbreaks in SIMILE[2].**

I. Introduction

Infectious animal diseases can spread at a rapid rate and have a severe negative impact on international travel [1], economies and trade [2]. To conform to national security regulations, officials need an efficient way to determine what potential threats can possibly affect the health and welfare of the citizens, especially in light of recently increased concerns about bioterrorism. Infectious disease informatics (IDI) is the area of studying data collection, sharing, modeling and management tasks in the domain of infectious diseases [3].

Several free online services for tracking disease outbreaks have recently been available. They collect data from news and allow users to monitor information about disease outbreaks. We give an overview of web resources that report infectious diseases outbreaks in Section II. Also, there are some systems which are manually maintained by state and federal governmental agencies. Such health organizations provide user-friendly interfaces for access to their data and analytical tools that described in Section II-A. We discuss the automated epidemic surveillance web interfaces that are similar to our system in Section II-B. We then present the overall system description for disease-related event detection from unstructured web documents in Section III. An overview of system functionality is given in Section III-A, web-crawling, information extraction, event recognition components in Sections III-B, III-C, III-D respectively. In Section IV we conclude with a discussion about our preliminary results and define directions for future work.

[1]GoogleMaps API - http://code.google.com/apis/maps/
[2]SIMILE API - http://www.simile-widgets.org/timeline/

II. Animal Disease Monitoring Systems

A. Manually Supported Web Interfaces

The World Organization for Animal Health (OIE)[3] is the one of the most important sources that report about animal health situations at the international level using *e.g.* the *World Animal Health Information Database (WAHID) Interface*[4]. The World Health Organization (WHO)[5] provides users with an interactive information mapping system, the *WHO Global Atlas of Infectious Diseases*[6]. The Animal Production and Health Division at Food and Agricultural Organization of United Nations[7] allows monitoring infectious disease outbreaks within a map and timeline view using the *Emergency Prevention System (EMPRES) for Transboundary Animal and Plant Pests and Diseases*[8]. The Department for Environment Food and Rural Affairs (DEFRA)[9] provides users with consistent information about animal health and welfare in United Kingdom.

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

- The U.S. Department of Agriculture (USDA)[10] manages a data system for animal diseases (*e.g.*, foot and mouth disease, rift valley fever);
- The U.S. Geological Survey (USGS) administer database for wildlife diseases through its National Wildlife Health Center (NWHC)[11];
- Centers for Disease Control and Prevention (CDC)[12] provide users with data about infectious diseases;
- Iowa State University Center for Food Security and Public Health (CFSPH)[13] website supplies users with information about infectious animal diseases, vaccines, disease fact sheets, image databases for diseases, and other useful resources for producers and veterinarians.

[3]OIE - http://www.oie.int/eng/en_index.htm
[4]WAHID Interface - http://www.oie.int/wahis/public.php?page=home
[5]WHO - http://www.who.int/en
[6]WHO Atlas Interface - http://diseasemaps.usgs.gov/index.htm
[7]FAO - http://www.fao.org/ag/againfo/home/en/index.htm
[8]EMPRES - http://www.fao.org/EMPRES/default.html
[9]DEFRA - http://www.defra.gov.uk
[10]USDA - http://www.usda.gov/wps/portal/usdahome
[11]NWHC - http://www.nwhc.usgs.gov
[12]CDC - http://www.cdc.gov
[13]CFSPH - http://www.cfsph.iastate.edu

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

- *BioSurveillance Portal* at the University of Arizona, maintained by its Artificial Intelligence Laboratory[14] is a web-based IDI system that provides access to distributed health data for several major infectious diseases;
- *Foot-and-mouth disease (FMD) BioPortal*[15] is developed for global FMD surveillance based on news monitoring and maintained by FMD Surveillance and Modeling Laboratory at the University of California UC Davis. *FMD BioPortal* uses crawlers that regularly collect FMD-related news from the Internet. Relevant news is stored in the database after keyword-based filtering from a large document collection [4].

In addition, there are several specific online resources for highly pathogenic animal diseases, *e.g.,* the *Reference Laboratories Information System*[16] for the OIE/FAO Foot-and-Mouth Disease Reference Laboratories Network. The necessity of human analysis and manual/semi-automated maintenance is a major drawback of the above discussed online systems for animal disease outbreaks tracking.

B. Automated Web Services

The BioCaster Global Health Monitor[17] is an online web-based system for detecting and mapping infectious disease outbreaks from news [5]. The system follows 1500 RSS feeds hourly that deal with a taxonomy of 4300 named entities (50 disease names, 243 country names, 4025 province/city names, and latitudes and longitudes for all locations). It is able to provide information on about 40 infectious diseases at up to 25-30 locations per day. *BioCaster Global Health Monitor* provides functionality such as: multilingual information extraction from news limited to English, French, Spanish, Chinese, Thai, Vietnamese, Japanese; their classification of documents as topically relevant or not; and plotting events on a Google Map [6], [7].

HealthMap[18] aggregates articles from *Google News* and *ProMED-Mail*[19] portal. It is a manually maintained Internet-based system that publishes reports generated by public health experts. The system allows tracking infectious diseases and locations related to the outbreak. It covers 2300 locations and 1100 disease names and identifies between 20-30 outbreaks per day. Since *HealthMap* is manually supported system, it supports processing text in multiple languages (English, French, Spanish, Portuguese, Russian, Chinese, Arabic) [8].

The information retrieval system *MedISys*[20], supported by the European Union is a part of the Europe Media Monitor (EMM)[21] product family, and was developed for searching web-based resources and producing quantitative summaries of the latest epidemics reports. This system includes the information extraction subsystem (the *Pattern-based Understanding and Learning System, PULS)*[22] that allows automated recognizing of the metadata and structured facts related to the disease outbreaks in text. *MedISys* currently collects an average 50000 news articles per day from about 1400 news portals from commercial news providers and from about 150 specialized Public Health sites. Moreover, *MedISys* allows data aggregation from multiple sources approximately on 43 languages about health-related topics such as: epidemics, nuclear, chemical/radiological, bio-terrorism, *etc*. The current ontology contains 2400 disease names, 400 organisms, 1500 political entities and over 70000 location names including towns, cities, provinces. During the information retrieval phase, the system performs real-time news clustering and filtering by matching 3000 patterns (*e.g.,* multi-word terms and their combinations), then classifies sources into 750 categories. During the information extraction phase, additional metadata is extracted such as: language, source country, download time, source site *etc.* from documents previously converted to Unicode [9].

The main advantage of *EpiSpider*[23] is the ability to combine emerging infectious disease data from *ProMED-Mail* with similar information from other sites *e.g., The Global Disaster Alert Coordinating System (GDACS)*[24]. In addition, *EpiSPIDER* extracts this information from the Central Intelligence Agency (CIA) Factbook[25] and the United Nations Human Development Report[26] sites.

The main differences between these intelligent systems and our framework include the system purpose (disease surveillance *vs.* research or epidemiological analytics), targeted audience (public *vs.* domain experts and analysts) and processed data (news *vs.* medical literature, blogs, e-mails *etc.*).

III. Framework for Epidemiological Analytics

The goal of this paper is to present an intelligent assistive technology overview for tracking animal infectious disease related events and discuss the preliminary results of the entity extraction and disease related event recognition components.

A. An Overview of System Functionality

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

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

[14]BioPortal - http://biocomputingcorp.com/bpsystem.html
[15]FMD BioPortal - https://fmdbioportal.ucdavis.edu
[16]ReLaIS - http://www.foot-and-mouth.org
[17]BioCaster - http://biocaster.nii.ac.jp/
[18]HealthMap - http://healthmap.org/en
[19]ProMED - www.promedmail.org
[20]MedISys - http://medusa.jrc.it/medisys/homeedition/all/home.html
[21]EMM - http://emm.jrc.it/overview.html
[22]PULS - http://sysdb.cs.helsinki.fi/puls/jrc/all
[23]EpiSpider - http://www.epispider.org/
[24]GDACS - www.gdacs.org
[25]CIA - https://www.cia.gov/library/publications/the-world-factbook/
[26]UNDHDR - http://hdr.undp.org/en

Users access system components using a web interface, search crawled documents, retrieve relevant information from the data storage, perform domain-specific entity extraction, recognize animal disease related events and visualize them on the map and within timeline, as shown in Figure 1.

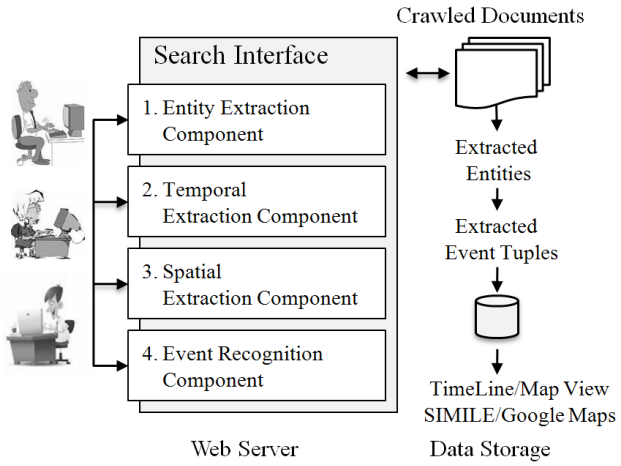

Fig. 1: System components for search, retrieval, extraction, event recognition and visualization functionality.

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

1) data **collection** using crawler components;
2) information **sharing** through the web interface;
3) query-based **search** using a Lucene-based[27] ranking component;
4) data **analysis** using entity extraction and event recognition components;
5) event **visualization** on a map (*GoogleMaps*) and within a timeline (*SIMILE*).

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

Algorithm 1 Information Retrieval Functionality (1 - 3)

Input: Set S of seeds $s_p \in S$ and set T of terms $t_i \in T$, set of topics K.
Output: collection D of documents d_j, set of documents R^q relevant to query q, and $R^q \subset D$.

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

[27]Lucene Search Engine API - http://lucene.apache.org/java/docs/

B. Data Collection using Web Crawling

For data collection we periodically crawl the web using Heritrix[28] crawler with customized set of seeds (*e.g.,* ProMed-Mail, DEFRA *etc.*) and terms (infectious animal disease names from the ontology). Figure 2 demonstrates that, by contrast with systems which use only news sources and do not digest refereed articles, we do not focus on specific sources.

After crawling, we perform an additional processing of web pages for meaningful entity extraction using domain-specific and domain-independent knowledge. Towards this goal, Weninger [10] developed a text-to-tag ratio-based method for content extraction from web pages.

Then, we perform document classification by topics using Naïve Bayes Classifier [25]. Finally, within the set of documents that are classified as disease-relevant, we allow users to perform search by disease and/or location entities in addition to general query-based keyword search.

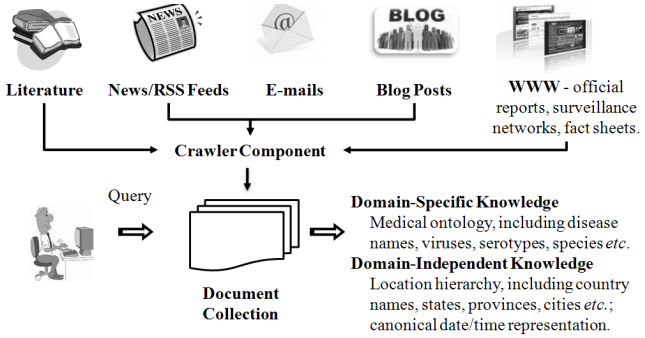

Fig. 2: Collection of animal disease related web documents

C. Domain-specific Entity Extraction

After collecting the data, we are focused on an entity extraction task that is an automatic extraction of structured information about animal disease-related events from unstructured crawled web documents. More precisely, we seek to locate and classify atomic elements in text into predefined categories as shown in Figure 3:

- disease names (*e.g., "foot-and-mouth disease"*);
- viruses (*e.g., "FMDV"*), serotypes (*e.g. "SAT-1"*) - N/A;
- species (*e.g., "cattle"*) and quantities - N/A;
- locations (*e.g., "China"*);
- dates (*e.g., "Friday, Dec 13"*);
- organizations (*e.g., "Agriculture Ministry"*).

We developed several tagging tools including disease and species exractors[29] for automated domain-specific entity extraction. For animal disease extraction, we constructed an initial ontology for the complete set of diseases and viruses using publicly available lists of animal disease names such as: CFSPH[30], DEFRA[31], OIE[32], Wikipedia[33].

[28]Heritrix Crawler - http://crawler.archive.org/
[29]KDD DSEx - http://fingolfin.user.cis.ksu.edu:8080/diseaseextractor/
[30]CFSPH - urlhttp://www.cfsph.iastate.edu/diseaseinfo/animaldiseaseindex.htm
[31]DEFRA - http://www.defra.gov.uk/foodfarm/farmanimal/diseases/atoz/
[32]OIE - http://www.oie.int/eng/maladies/en_alpha.htm
[33]Wikipedia - http://en.wikipedia.org/wiki/Animal_diseases

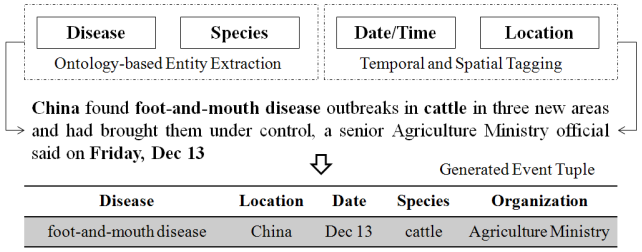

Fig. 3: Information extraction component functionality for tagging disease names, species, dates, locations and organizations.

For evaluation purposes, we collected 100 domain-specific web documents including pdfs that report animal disease outbreaks. The size of the collection that we used for evaluation is constrained by the effort required for manual annotation of the animal disease entities.

Experimental results for ontology-based entity extraction approach are presented on Figure 4 for different feature combinations such as: 1a - using initial ontology (G), no synonyms $(\neg S)$, no abbreviations $(\neg A)$ and no capitalization $(\neg C)$, 1b - $(G, \neg S, \neg A, C)$, 2a - $(G, S, \neg A, \neg C)$, 2b - $(G, S, \neg A, C)$, 3a - $(G, \neg S, A, \neg C)$, 3b - $(G, \neg S, A, C)$, 4a - $(G, S, A, \neg C)$, 4b - (G, S, A, C). For a detailed discussion of animal disease entity extraction, ontology learning based on synonymic, hyponymic and causal relationship extraction, we refer the reader to [11].

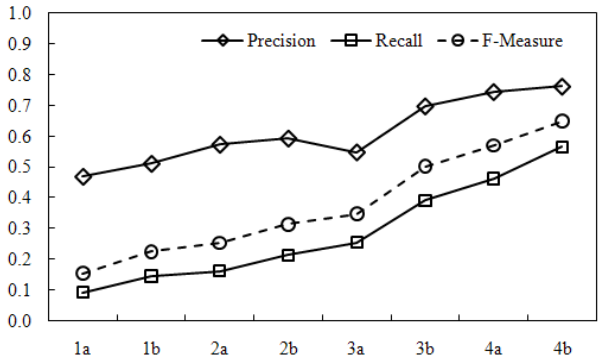

Fig. 4: The precision, recall and F-measure values for animal disease extraction using different ontologies, for example, 1a. $|(G, \neg S, \neg A, \neg C)|$ = 429 concepts, 2a. $|(G, S, \neg A, \neg C)|$ = 581 concepts, 3a. $|(G, \neg S, A, \neg C)|$ = 453 concepts, 4a. $|(G, S, A, \neg C)|$ = 605 concepts.

For boosting animal disease extraction results in future, we plan to enrich semantically and extend our initial ontology G by extracting semantic relations (including synonymic, hyponymic and causative) between concepts. For semantic relation extraction approach we use syntactic pattern matching in combination with Part-of-Speech (POS) tagging[34]. For example, if we know a disease D and do not know the virus that causes it, we can learn the right-hand side patterns of relationship entailed in the text, such as E_i where "D *is caused by* E_i" [12], [13].

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

For location and organization named entity extraction, we used Stanford Named Entity Recognition (NER) tool[35]. It is based on conditional random fields approach developed by Lafferty [14].Moreover, we refer to GEOnet Names Database (GNS)[36] for location disambiguation and getting latitude/longitude values. For date extraction, we perform pattern matching using regular expression-based rules. For species extraction we use pattern matching on a stemmed dictionary of animal names from Wikipedia.

D. Animal Disease-related Event Recognition

The event recognition functionality is based on the entity extraction component which is described using an example in Figure 3. As can be seen, the extracted entities can be possibly augmented in event tuple in form $[disease, location, date, species]$, where the main event descriptors are disease, date, location and species. Additionally, we can extract organization that reports an outbreak.

More precisely, we describe how the entity extractors discussed in Section III-C produce an event tuple for an example sentence in Figure 5. Initially, each document is tokenized into sentences; then disease, location, species and dates taggers are applied in addition to a confirmation status extractor which relies on the set of specific verbs for event recognition. For example, the sentence "*Foot and mouth disease is[V] a highly pathogenic animal disease*" is not disease related event, and by using constrained sets of a confirmation status verbs, we are able to eliminate this sentence.

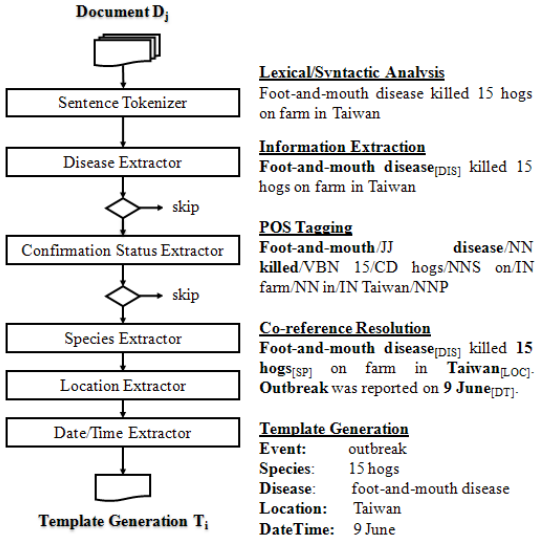

Fig. 5: Event recognition component for event tuple generation using extracted entities and confirmation status verbs.

As can be seen from Figure 5, we perform a sentence-based event recognition. Therefore, the future work requires co-reference resolution on the document level for generating the most complete event tuple and disambiguation events with missing attributes [15], [16].

[35]Stanford NER - http://nlp.stanford.edu/ner/index.shtml
[36]*GNS - http://earth-info.nga.mil/gns/html/*

We designed an experiment for evaluation of the event recognition approach which is described in details in [17]. For that purposes, we used Google to retrieve 100 documents. Furthermore, we used the Pyramid scoring method [18] for automated comparison of extracted events with summaries constructed for each of 100 documents. In accordance to Pyramid, extracted events can get score in range [0 - 1], where 1 denotes a perfectly extracted event.

The experimental results for disease-related event recognition are shown in Figure 7 for different sets of confirmation status verbs such as: INS - initial set unstemmed, GNS - *GoogleSets*[37] unstemmed, IS - initial set stemmed and GS - *GoogleSets* stemmed. As can be seen, the list of verbs extended using *GoogleSets* and then stemmed GS demonstrates, on the one hand, an increasing numbers of recognized events in comparison to results obtained using initial list unstemmed INS, but on the other hand, the these results are similar to the results obtained from initial list stemmed IS. Moreover, both the *GoogleSets* stemmed GS and initial list stemmed IS allow us extracting events within high score range [0.71 - 1]. It shows the feasibility of the proposed animal disease-related event recognition approach with accuracy 65% for both initial list of verbs and extended using *GoogleSets*.

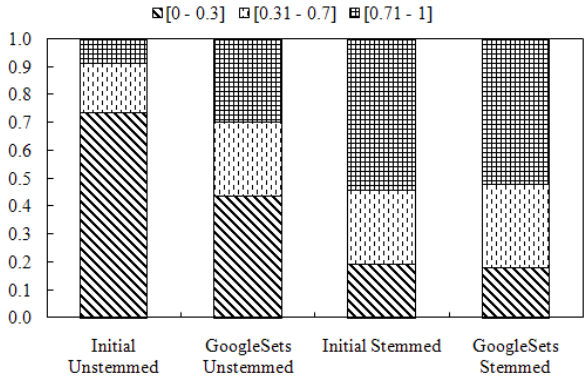

Fig. 6: The event recognition scores for initial list of verbs and extended with *GoogleSets* calculated using Pyramid method

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

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

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

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

[37]GoogleSets - http://labs.google.com/sets

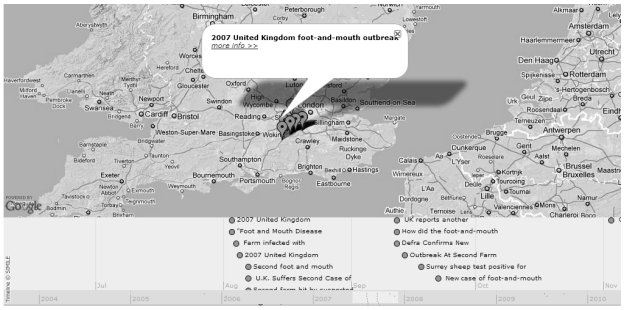

Fig. 7: Temporal and spatial visualization of the extracted events using *GoogleMaps* and *SIMILE* respectively

IV. Summary and Future Work

We have presented results from continuing research on entity extraction and animal disease-related event recognition in a veterinary medical intelligence domain. These preliminary results demonstrate the effectiveness of the ontology-based approach for domain-specific entity extraction and the sentence-based event recognition approach.

Working on the project we encountered several opened research questions including: managing the contextual specificity of blogosphere [19]; processing biomedical literature [20], [21], [22]; mining news content vs. official health reports [23].

Similarly to other systems, we applied an existing approaches for data collection using web crawling [24] and document by topics classification [25]. For domain-specific entity extraction, we proposed an ontology-based extraction method and semantic relation learning approach for ontology expansion [11]. Using these techniques for animal disease extraction, we obtained precision value as high as 76% and recall value as high as 56%. For event recognition, we suggested to apply event tuple generation using extracted entities such as: disease, location, date, species together with the confirmation status verb compared to the "disease-location" pair used in other systems. We received preliminary results in terms of accuracy as high as 65%. In order to increase the recall of event recognition approach, we intend to use more structured sources such as *WordNet* [26] instead of *GoogleSets* for confirmaton status verb list expansion.

Consequently, in comparison to other systems which are designed for mining news and have no functionality for past outbreak tracking (*BioCaster* - 1500 News Feeds; *HealthMap* - *Google News*, *ProMED-Mail*; *MedISys* - 1400 news portals, 150 Public Health sites), perform ontology-based information extraction for limited number of domain-specific entities (*BioCaster* - 50 diseases including synonyms, symptoms, syndromes; *HealthMap* - 1100 diseases; *MedISys* - 2400 disease names, 400 organisms, 1500 political entities), require manual moderation phase (*HealthMap*), limited with geo-entity extraction (*BioCaster* - 243 countries, 4,025 cities; *HealthMap* - 2,300 locations; *MedISys/PULS* - 70,000 locations) and have no timeline visualization (*BioCaster*), our system:

- performs focused crawling of different sources (books, research papers, blogs, governmental sources, *etc.*);

- uses semantic relationship learning approach (including synonymic, hyponymic, causal relationships) for automated-ontology expansion for domain-specific entity extraction (*e.g.,* diseases, viruses) [11];
- recognizes geo-entities using CRF approach and disambiguates them using GNServer;
- extract animal disease-related events with more descriptive event attributes such as: species, dates, event confirmation status [17], in contrast to "disease-location" pairs;
- supports timeline representation of extracted events in *SIMILE* in addition to visualized events on *GoogleMaps*.

One limitation of our system is the ability to process web document only in English compare to other systems (*BioCaster* - English, French, Spanish, Chinese, Thai, Vietnamese, Japanese; *HealthMap* - English, French, Spanish, Portuguese, Russian, Chinese, Arabic; *MedISys* - 43 languages). To address this issue, our future work aims at applying a "wikification" approach and using knowledge from Wikipedia for multilingual information extraction [27] and disambiguation [28].

ACKNOWLEDGMENTS

This work was supported by a grant from the U.S. Department of Defense. We would like to acknowledge KDD Lab alumni: Tim Weninger (crawler deployment) and Jing Xia (rule-based event extraction); KDD Lab assistants: Information Extraction Team (John Drouhard, Landon Fowles, Swathi Bujuru); Spatial Data Mining Team (Wesam Elshamy, Andrew Berggren); Topic Detection and Tracking Team (Surya Kallumadi, Danny Jones, Srinivas Reddy). A collaborative program on information extraction with faculty at the University of Illinois at Urbana-Champaign (ChengXiang Zhai, Dan Roth, Jiawei Han and Kevin Chang) and the 2009 Data Sciences Summer Institute contributed to this research.

REFERENCES

[1] M. E. Wilson, "Travel and the emergence of infectious diseases, taylor and francis group," *Journal of Agromedicine*, vol. 3, pp. 51–66, 1996.

[2] N. M. M'ikanatha, D. D. Rohn, C. Robertson, C. G. Tan, J. H. Holmes, A. R. Kunselman, C. Polachek, and E. Lautenbach, "Use of the internet to enhance infectious disease surveillance and outbreak investigation." *Biosecurity and bioterrorism : biodefense strategy, practice, and science*, vol. 4, no. 3, pp. 293–300, 2006. [Online]. Available: http://dx.doi.org/10.1089/bsp.2006.4.293

[3] Y. Zhang, Y. Dang, Y.-D. Chen, H. Chen, M. Thurmond, C.-C. King, D. D. Zeng, and C. A. Larson, "Bioportal infectious disease informatics research: disease surveillance and situational awareness," in *dg.o '08: Proceedings of the 2008 international conference on Digital government research*. Digital Government Society of North America, 2008, pp. 393–394.

[4] M. Thurmond, A. Perez, C. Tseng, H. Chen, and D. Zeng, "Global foot-and-mouth disease surveillance using bioportal," pp. 169–179, 2009. [Online]. Available: http://www.springerlink.com/content/87g66180875p7172

[5] S. Doan, QuocHung-Ngo, A. Kawazoe, and N. Collier, "Global Health Monitor - a web-based system for detecting and mapping infectious diseases," in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2008, pp. 951–956.

[6] C. Nigel, D. Son, K. Ai, G. R. Matsuda, C. Mike, T. Yoshio, N. Quoc-Hung, D. Dinh, K. Asanee, T. Koichi, S. Mika, and T. Kiyosu, "Biocaster: detecting public health rumors with a web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, 2008.

[7] A. Kawazoe, H. Chanlekha, M. Shigematsu, and N. Collier, "Structuring an event ontology for disease outbreak detection," *BMC Bioinformatics*, vol. 9 Suppl 3, 2008. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-9-S3-S8

[8] C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports." *J Am Med Inform Assoc*, December 2007. [Online]. Available: http://dx.doi.org/10.1197/jamia.M2544

[9] R. Steinberger, F. Fuart, E. Groot, C. Best, P. Etter, and R. Yangarber, "Text mining from the web for medical intelligence," *Mining Massive Data Sets for Security*, 2008.

[10] T. Weninger and W. H. Hsu, "Text extraction from the web via text-to-tag ratio," in *DEXA Workshops*. IEEE Computer Society, September 2008, pp. 23–28.

[11] S. Volkova, W. Hsu, and D. Caragea, "Named entity recognition and tagging in the domain of epizootics," 2009, poster presentation at Women in Machine Learning Workshop (WiML'09).

[12] P. Cimiano and S. Staab, "Learning by googling," *SIGKDD Explor. Newsl.*, vol. 6, no. 2, pp. 24–33, 2004.

[13] R. C. Wang and W. W. Cohen, "Language-independent set expansion of named entities using the web," *Data Mining, IEEE International Conference on*, vol. 0, pp. 342–350, 2007. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2007.104

[14] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[15] E. Bengtson and D. Roth, "Understanding the value of features for coreference resolution," in *EMNLP*, Oct 2008, pp. xx–yy. [Online]. Available: http://l2r.cs.uiuc.edu/~danr/Papers/BengtsonRo08.pdf

[16] X. Li, P. Morie, and D. Roth, "Semantic integration in text: From ambiguous names to identifiable entities," *AI Magazine. Special Issue on Semantic Integration*, pp. 45–68, 2005. [Online]. Available: http://l2r.cs.uiuc.edu/~danr/Papers/LiMoRo05.pdf

[17] S. Volkova, D. Caragea, W. H. Hsu, and S. Bujuru, "Animal disease event recognition and classification," 2010.

[18] A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *HLT/NAACL*, 2004.

[19] A. Neustein, "Sequence package analysis: A new method for intelligent mining of patient dialog, blogs and help-line calls," *Journal of Computers*, vol. 2, no. 10, 2007. [Online]. Available: http://www.academypublisher.com/ojs/index.php/jcp/article/view/02104551

[20] Q. Mei, D. Xin, H. Cheng, J. Han, and C. Zhai, "Semantic annotation of frequent patterns," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 3, p. 11, 2007.

[21] J. Jiang and C. Zhai, "An empirical study of tokenization strategies for biomedical information retrieval," *Inf. Retr.*, vol. 10, no. 4-5, pp. 341–363, 2007.

[22] Y. Lu, H. Fang, and C. Zhai, "An empirical study of gene synonym query expansion in biomedical information retrieval," *Inf. Retr.*, vol. 12, no. 1, pp. 51–68, 2009.

[23] J. Woodall, "Official versus unofficial outbreak reporting through the internet," *International Journal of Medical Informatics*, vol. 47, pp. 31–34, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/B6T7S-3SFN8R0-4/2/1c6803f56112ea42720a3dd6b7155ff3

[24] S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*, 1st ed. Morgan Kaufmann, October 2002. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/1558607544

[25] T. M. Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, March 1997. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0070428077

[26] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998, http://wordnet.princeton.edu/.

[27] A. E. Richman and P. Schone, "Mining wiki resources for multilingual named entity recognition," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 1–9. [Online]. Available: http://www.aclweb.org/anthology-new/P/P08/P08-1001.bib

[28] X. Han and J. Zhao, "Named entity disambiguation by leveraging wikipedia semantic knowledge," in *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2009, pp. 215–224. [Online]. Available: http://dx.doi.org/10.1145/1645953.1645983