Animal Disease Event Recognition and Classification

Svitlana Volkova, Doina Caragea, William H. Hsu, Swathi Bujuru

Department of Computing and Information Sciences
Kansas state University, 234 Nichols Hall, Manhattan, KS, USA 66506
{svitlana,dcaragea,bhsu,swathi}@ksu.edu

Abstract. Monitoring epidemic crises, caused by rapid spread of infectious animal diseases, can be facilitated by the plethora of information about disease-related events that is available online. Therefore, the ability to use this information to perform domain-specific entity recognition and event-related sentence classification, which in turn can support time and space visualization of automatically extracted events, is highly desirable. Towards this goal, we present a rule-based approach to the problem of extracting animal disease-related events from web documents. Our approach relies on the recognition of structured entity tuples, consisting of attributes, which describe events related to animal diseases. The event attributes that we consider include animal diseases, dates, species and geo-referenced locations. We perform disease names and species recognition using an automatically-constructed ontology, dates are extracted using regular expressions, while location are extracted using a conditional random fields tool. The extracted events are further classified as confirmed or suspected based on semantic features, obtained from the *e.g., GoogleSets*[1] and *WordNet*[2]. Our preliminary results demonstrate the feasibility of the proposed approach.

Key words: entity recognition, animal disease, event tuple detection, classification, text mining

1 Introduction

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

Infectious Disease Informatics (IDI) includes tasks such as: data collection, sharing, management, modeling and analysis in the domain of emerging infectious diseases [1]. An enormous amount of data about animal infectious disease-related events is available online in both structured and unstructured formats.

[1] GoogleSets Inteface - http://labs.google.com/sets
[2] WordNet - http://wordnet.princeton.edu/

Structured data is presented to public in official reports by different organizations such as: state and federal laboratories, local health care providers, governmental agricultural or environmental agencies. In addition, a lot of unstructured information can be found in a variety of other contexts *e.g.,* news, e-mails, blogs, which in contrast to the official reports is completely unorganized. In order to exploit this unstructured data, machine learning and text mining techniques can be used to recognize disease-related events, *e.g., "On 12 September 2007, a new foot-and-mouth disease outbreak was confirmed in Egham, Surrey"*. Such techniques could be part of automated systems that can detect, monitor and track responses to animal infectious disease outbreaks (defined as a set of events which are constrained in space and have temporal overlap).

Several automated systems for animal disease monitoring exist [2], [3]. They have the ability to crawl news and use ontology pattern matching approaches to recognize entities such as disease and location of an event. While the existing systems focus on news data and identify emergent diseases, in this paper we describe a system which can be used not only with news data, but also with e-mails, blog posts and scientific web articles. Therefore, our system can identify events in historical data as opposed to identifying only emergent disease events. Specifically, our system extracts event tuples from a variety of web documents. These tuples can be seen as structured summaries of the events specified by attributes such as: disease, location, date, species and confirmation status.

2 Methodology

In this section we describe in detail our methodology for identifying disease-related events and their associated confirmation status. The confirmation status refers to an event being suspected or confirmed. This information is important with respect to the action that needs to be taken. Our approach to the event recognition problem involves three main steps: first, we perform entity recognition from unstructured sources; next, we classify the sentences from which entities are extracted as being related to an event or not; furthermore, if they are related to an event we classify them as confirmed or suspected; finally, we combine entities within an event sentence into structured tuples. Figure 1 illustrates these three steps through an example.

2.1 Entity Recognition

The entity recognition module in our system automatically extracts structured information related to animal diseases from unstructured web documents. To achieve this functionality we associate meta-data in the form of ontologies with documents in our collection. Specifically, the meta-data consists of *domain-independent* location and time hierarchies (including names of countries, states, cities; and canonical dates) and a *domain-specific* medical ontology (including diseases, serotypes, and viruses). Based on these ontologies and pattern matching, we design specialized extractors that locate and classify atomic elements into predefined categories such as:

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

For the animal disease name recognition, we developed an Animal Disease Extractor (DSEx)[3], which relies on a medical ontology, automatically-enriched with synonyms and causative viruses [4]. For species extraction we use pattern matching on a stemmed dictionary of animal names from Wikipedia[4]. Furthermore, we used the Stanford NER[5] tool (which uses conditional random fields) together with NGA GEOnet Names Database (GNS)[6] for location recognition and set of regular expressions for date/time extraction.

The top panel in Figure 1 shows a paragraph where entities recognized by our extractors are highlighted. As an example, the output from our entity recognition module for the sentence *"Taiwan's TVBS television station reports that agricultural authorities confirmed foot-and-mouth disease on a hog farm in Taoyuan"* is shown below:

– animal diseases - *"foot-and-mouth disease"* (recognized by the DSEx);
– locations - *"Taoyuan"* (recognized by the Location Extractor);
– species - *"hog"* (recognized by the Species Extractor).

2.2 Event Sentence Classification

After the entities are recognized in a document, we next extract sentences that contain such entities and classify them as corresponding to true events or false positive events. True events should include a disease name together with a disease-related verb. Furthermore, these events are classified as confirmed or suspected using the Confirmation Status Extractor. This extractor relies on a restricted list of verbs that suggest confirmed events (*e.g., happened*) or suspected events (*e.g., catch*) and their synonyms identified using *GoogleSets*[1] or *WordNet*[2] [5] . For example, the following sentence is classified as corresponding to a confirmed event: *"On 9 Jun 2009, the farm's owner reported symptoms of FMD in more than 30 hogs."*

The initial list of verbs consists of single word verbs (*e.g., kill*) and verb phrases (*e.g., strike out*). The first two columns in Table 1 show the number of initial verbs denoted as $IN\text{-}V$ and verb phrases denoted as $IN\text{-}VP$ for both suspected and confirmed categories. Columns 3 and 4 show similar numbers for the augmented list of verbs obtained using *GoogleSets*[1] (denoted by $GS\text{-}V$, $GS\text{-}VP$ respectively), while columns 5, 6 show these numbers for *WordNet*[2] (denoted by $WN\text{-}V$, $WN\text{-}VP$ respectively).

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

Table 1: Statistics about the restricted list of verbs

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

The list of verbs used to classify sentences as confirmed or suspected is also useful for eliminating frequent, but not event-related sentences such as: "*Foot and mouth disease is[V] a highly pathogenic animal disease*".

The second step in Figure 1 shows more examples of potential event-related sentences and their classification. We first classify sentences as event-related (corresponds to *"YES"*) or event non-related (corresponds to *"NO"*). We then classify event-related sentences as suspected or confirmed based on the restricted list of verbs and verb phrases represented in Table 1.

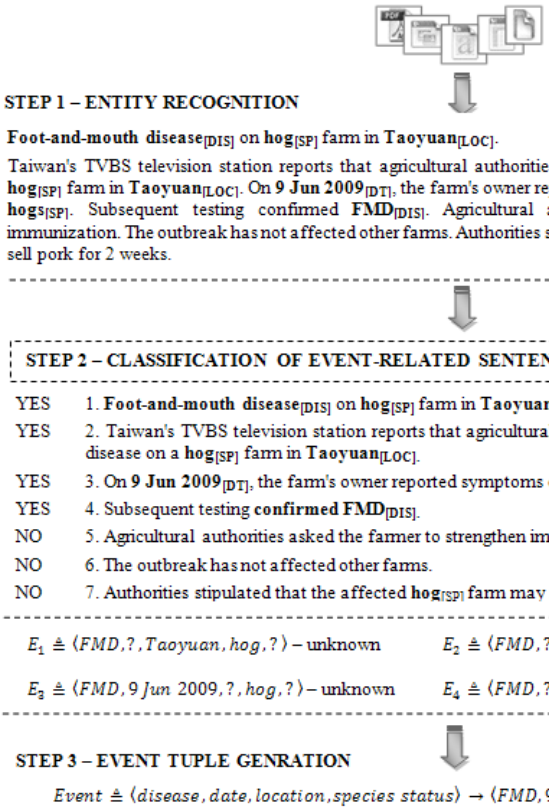


Fig. 1: Description of the system workflow through an example: first, entities are recognized using several extractors; second, the true event sentences are identified and classified as suspected or confirmed; next, instances from true event sentences are grouped together into potential event tuples; finally, instances of the same event are consolidated into one comprehensive tuple.

2.3 Event Tuple Generation

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

$$Event_i =< disease, date, location, species, status >, \quad (1)$$

where each attribute in the tuple obtained with one of the extractors described in Section 2.1. The following tuple <*FMD, 9 Jun 2009, Taoyuan, hog, confirmed*> is an example of an event. Given the incomplete and the uncertain nature of the information available online, it is possible for events to have missing values, *e.g.,* $< disease, ?, location, species, ? >=< FMD, ?, Taoyuan, hog, ? >$, $< disease, date, ?, species, ? >=< FMD$, *09 Jun 2009* $, ?, hog, ? >$. For instance, news reports can contain information about disease-related events that happened in some location without a specific date or species being provided.

Furthermore, several sentences in a document can contain information about the same event and we aggregate the corresponding event tuples into a unique tuple based on the attributes available, as shown in the last step in Figure 1.

Algorithm 1 Entity Recognition, Sentence Classification and Tuple Generation

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

Algorithm 1 summarizes the steps for entity recognition, event-related sentence classification and tuple generation.

3 Experimental Design and Results

We used the existing *DUCView Pyramid* scoring tool [6] to score automatically generated event tuples and evaluate our approach. Pyramid scoring is a technique for evaluating summarization results, which was introduced in [7] and relies on multiple summaries to assign the significance weights to summarization content units (*i.e.*, entities) [8].

To perform the evaluation, we used Google to retrieve 100 documents related to two animal diseases: rift valley fever (RVF) and foot-and-mouth disease (FMD). We manually created two sets of summaries for each of the 100 documents and extracted entities corresponding to event tuples from each summary and each document as described in Section 2.1. Then, we used the *DUCView* tool to compare automatically generated event tuples with entities from human summaries. As a result, the entities from event tuples are assigned weights in the range [0, 1] where 1 represents the best recognition score and it means that entity from automatically-generated tuple is present in all summaries. The entity weights are used to calculate an aggregated score for event tuples. Specifically, the score for an event tuple described in Equation 1 is given by:

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

where $disease, \cdots, species$ take 0/1 values (entity present or not in the tuple) and a tuple is valid only if both *disease* and *status* are present. The resulting scores are reported as a measure of the accuracy of the proposed event tuple recognition and classification approach and shown in Table 2.

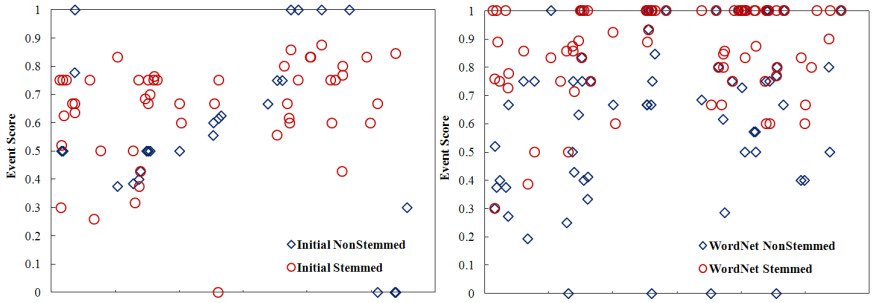
More precisely, we evaluate our event tuple recognition and classification approach by applying three lists of verbs and verb phrases for confirmation status extraction which are introduced in Table 1. Furthermore, we consider stemmed *S vs.* non-stemmed *NS* versions of these lists. The results for the non-stemmed version of the lists are shown in the first three columns of the Table 2 for the initial list, *GoogleSets*[1] augmented list and *WordNet*[2] augmented list, respectively. Similarly, the results for the stemmed version are shown in the last three columns of the Table 2.

Table 2: Pyramid Event Score Distribution by Range

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

As can be seen from the Table 2, the initial list of verbs results in many low score events which means that not many tuples can be extracted with high confidence using only these verbs. While the augmented lists, without stemming, give better results, only approximately one third of the events are scored with a

high confidence for both *GoogleSets*[1] and *WordNet*[2]. However, the scores increase significantly for all lists when stemming is used. The best results are obtained for the *WordNet*[2] augmented list where the average score is as high as 0.75.



(a) Initial List (b) WordNet

Fig. 2: The scatter plot of the event scores using Pyramid method

Figure 2 shows the scatter plot of the event score distribution using the initial and *WordNet*[2] lists, for both stemmed and non-stemmed versions of lists. As can be seen, more events are identified using the *WordNet*[2] list and they have higher scores (many of them have the max score 1).

4 Related Work

There are several systems for disease-related event detection that extract diseases and locations from text. BioCaster[7] is an online ontology-based system for detecting and mapping infectious disease outbreaks from news [9]. Their approach for event detection is based on searching for disease-location pairs and calculating their frequency in the document and in the collection [2]. The methodology for deriving synonyms for disease-related verbs that are part of events (*e.g., disease, verb, location*) is similar to our approach. However, *BioCaster* does not provide assistance with classification of extracted events as confirmed or suspected. As opposed to *BioCaster, HealthMap*[8] is a manually supported web system, which crawls data from Google News and the ProMED-Mail[9] portal and provides reports about disease outbreaks to the public [10]. *Pattern-based Understanding and Learning System (PULS)*[10], which is part of the *MedISys*[11], allows extracting meta-data and structured facts related to the disease outbreaks [3]. Similar to other systems, it does not classify extracted events and does not report anything about past outbreaks. Our approach addresses the abovementioned limitations. It supports automated extraction of disease-related event tuples, which include disease, date, location, species entities and confirmation status. It also classifies them into two categories such as: suspected or confirmed.

[7] BioCaster Global Health Monitor - `http://biocaster.nii.ac.jp/`
[8] HealthMap System - `http://healthmap.org/en`
[9] ProMED-Mail - `www.promedmail.org`
[10] PULS - `http://sysdb.cs.helsinki.fi/puls/jrc/all`
[11] MedISys - `http://medusa.jrc.it/medisys/homeedition/all/home.htm`

5 Conclusions

In this paper, we presented an approach for animal disease event recognition and classification. Entity and confirmation status extraction methods are used to automatically generate structured summaries about domain-specific events in the form of tuples. Furthermore, we apply several lists of verbs for confirmation status extraction including *WordNet*[2] and *GoogleSets*[1]. We used the *Pyramid* method and *DUCView* tool [6] to calculate scores for automatically generated event tuples, which can be seen as a measure of accuracy of our approach. The highest accuracy was obtained using a *WordNet*[2] augmented list of verbs. As part of future work we intend to apply a deeper syntactic analysis of the sentence and part-of-speech tagging in addition to the list of verbs that we used.

Acknowledgments. This work was supported through a grant from the U.S. Department of Defense. We would like to acknowledge the Knowledge Discovery in Databases Laboratory assistants: John Drouhard, Landon Fowles (disease/species extractor), Wesam Elshamy, Andrew Berggren (location extractor), Danny Jones, Srinivas Reddy (date/time extractor).

References

1. Chen, H., Fuller, S.S., Friedman, C.P.: Medical Informatics: Knowledge Management and Data Mining in Biomedicine (Integrated Series in Information Systems). Springer (June 2005)
2. Kawazoe, A., Chanlekha, H., Shigematsu, M., Collier, N.: Structuring an event ontology for disease outbreak detection. BMC Bioinformatics **9 Suppl 3** (2008)
3. Steinberger, R., Fuart, F., Groot, E., Best, C., Etter, P., Yangarber, R.: Text mining from the web for medical intelligence. Mining Massive Data Sets for Security (2008)
4. Volkova, S., Hsu, W., Caragea, D.: Named entity recognition and tagging in the domain of epizootics (2009) Women in Machine Learning Workshop, `http://wimlworkshop.org/`.
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998) http://wordnet.princeton.edu/.
6. Nenkova, A.: Pyramid Annotation Guide - DUC 2006 `http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html_ducview`.
7. Nenkova, A.: Understanding the process of multi-document summarization: content selection, rewriting and evaluation. PhD thesis, New York, NY, USA (2006)
8. Nenkova, A., Passonneau, R., McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Trans. Speech Lang. Process. **4**(2) (2007)
9. Doan, S., QuocHung-Ngo, Kawazoe, A., Collier, N.: 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) 951–956
10. Freifeld, C., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. J Am Med Inform Assoc (December 2007)