Boosting Biomedical Entity Extraction by using Syntactic Patterns for Semantic Relation Discovery

Svitlana Volkova, Doina Caragea, William H. Hsu, John Drouhard, Landon Fowles
Department of Computing and Information Sciences
Kansas State University
Manhattan, Kansas 66506
Email: {svitlana, dcaragea, bhsu, drouhard, eclipsor}@ksu.edu

***Abstract*—Biomedical entity extraction from unstructured web documents is an important task that needs to be performed in order to discover knowledge in the veterinary medicine domain. In general, this task can be approached by applying domain-specific ontologies, but the literature review shows that there is no universal dictionary, or ontology for this domain. To address this issue, we manually construct an ontology for extracting entities such as: animal disease names, viruses and serotypes. We then use an automated ontology expansion approach that relies on extracting semantic relationship between concepts. Such relationships include asserted synonymy, hyponymy and causality. To extract semantic relationships, we expand the manually-constructed ontology by using a set of syntactic patterns and part-of-speech tagging. As a result, we obtain an ontology which contains richer semantics compared to the manually-constructed ontology. We compare our approach of learning synonyms, hyponyms and other disease related concepts with an approach where the ontology is expanded using *GoogleSets*[1], on the veterinary medicine entity extraction task. Our experiments show that the semantic relationship learning approach results in a significant increase in precision and recall compared to the *GoogleSets* approach.**

I. Introduction

In epidemiology, a disease outbreak is defined as a disease occurrence that is greater than expected in a particular time and place. Outbreaks, which result in large-scale spread of infectious diseases, have great negative impact on society. They can influence relationships between bordering countries in terms of trade restrictions, which in turn can cause economical and political instability in the region. Thus, detecting, managing, preventing and responding to disease outbreaks are very important tasks [1].

The success of such tasks relies on the ability to extract information from large amounts of domain-specific data available online. This data includes both structured formats *e.g.,* official emergency surveillance databases from World Animal Health Information Database (WAHID)[2], Food and Agricultural Organization of United Nations Emergency Prevention System (EMPRES)[3] and unstructured free text *e.g.,* news and official reports from Department for Environment Food and Rural Affairs (DEFRA)[4], World Organization for Animal Health (OIE)[5], Centers for Disease Control and Prevention(CDC)[6]; medical literature - PubMed[7], e-mails - ProMED-Mail[8] *etc*. The goal of this paper is to automate the process of animal disease extraction from unstructured web sources using text mining and natural language understanding techniques.

The extraction accuracy is crucial, because it influences on several related tasks including disease pattern classification, disease-related event recognition, and domain-specific information retrieval. For example, animal disease extraction is a required prerequisite step for domain-specific event recognition *e.g., "The US saw its latest FMD outbreak in Montebello, CA in 1929 where 3,600 pigs were slaughtered"*, where animal disease names are the major structural components of the event descriptors [2]. Therefore, the precision and recall of extracted entities has a direct influence on the event recognition accuracy. Thus, the biomedical entity extraction accuracy should be maximized in order to bring more accurate results for all abovementioned tasks.

In order to identify potential animal disease outbreaks within domain-specific unstructured web documents (*e.g.,* news, reports, papers, e-mails, *etc.*), we need to extract veterinary medicine entities including animal diseases, viruses and serotypes. However, there is no comprehensive dictionary or ontology for this domain and previous approaches to similar problems in information extraction were based upon human diseases and medical dictionaries.

To address the lack of a veterinary medicine ontology, we first manually build a set ontologies as explained in Section III-A and expand the initial ontology with semantic relationships (synonymic, hyponymic and causal) identified using syntactic patterns and part of speech tagging, as described in Section III-B. We then show how to use these semantic relationships for expansion of the manually constructed ontology and automatically construct new ontology in Section III-C. In Section III-D we present an overview for the biomedical entity extraction task for the domain of the veterinary medicine using an animal disease example. In Section IV, we discuss the results of biomedical entity extraction using manually vs. automatically-constructed ontologies. Fur-

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

[2] WAHID - http://www.oie.int/wahis/public.php?page=home

[3] EMPSES - http://www.fao.org/ag/againfo/programmes/en/empres/

[4] DEFRA - http://www.defra.gov.uk

[5] OIE - http://www.oie.int/eng/en_index.htm

[6] CDC - http://www.cdc.gov

[7] PubMed - http://www.ncbi.nlm.nih.gov/pubmed/

[8] ProMED-Mail - www.promedmail.org

thermore we compare the automatically-constructed ontology obtained using our relationship extraction approach with an ontology constructed using *GoogleSets* expansion approach, which refers to expanding a given partial set of objects into a more complete set. We compare the entities extracted using all abovementioned ontologies in terms of precision and recall, build ROC curves and report F-measure values as a function of the ontology size. The results show that the our semantic relationship extraction approach brings new knowledge to the initial ontology and, therefore, boosts the domain-specific biomedical entity extraction results.

II. Related Work

Resources that can be used for boosting biomedical entity extraction results by discovering semantic relationships between entities, can be divided into several categories:

- structured domain-independent *e.g., WordNet*[9];
- structured domain-dependent *e.g., Unified Medical Language System (UMLS)*[10], *Word Health Organization International Classification of Diseases (ICD)*[11], *Systematized Nomenclature of Medicine - Clinical Terms (SNOMED)*[12];
- semistructured domain-independent *e.g., Wikipedia*[13].

Although, *WordNet* is a manually constructed lexical database with structured knowledge and *Wikipedia*, by contrast, is an unstructured source of knowledge, they both are not domain-specific, therefore, they do not include enough information about infectious animal diseases, their synonyms and viruses. Also, the other domain-specific resources mentioned above *UMLS*, *ICD* and *SNOMED* can not be effectively used for biomedical entity extraction in the domain of veterinary medicine, because they include structured information related to both human and animal diseases. Therefore, a unified ontology in a veterinary medicine domain is needed.

The process of the ontology construction is very difficult, labor-intensive and time consuming. In order to reduce the cost of building ontologies, there are several ontology learning systems which allow to extract concepts and relations between concepts from text *e.g., OntoLearn* [3], *OntoMiner* [4] and many others that are discussed in [5]. However, such systems are generally based upon shallow natural language processing techniques and, therefore, extract concepts with only taxonomic (*e.g.,* synonymic *"is-a"*) relations between them. The taxonomic relation discovery approaches have been addressed primarily within the biomedical field as there are very large text collections readily available *e.g.* PubMed.

Other systems for automated ontology construction, such as *Text-To-Onto* [6] and its successor *Text2Onto* [7], allow extracting also non-taxonomic (*e.g.,* hyponymic) relations between concepts using association rule-mining and predefined regular expressions. Their main drawback is that they cannot effectively extract domain-specific concepts, because they identify semantic relations based on part-of-speech tags only. However, Cimiano and Staab [8] demonstrated the effectiveness of their system for extracting general concepts including person and location named entities. They use taxonomic and non-taxonomic patterns for semantic relation discovery between concepts, as a preliminary step for entity classification.

By contrast with many ontology learning systems that use shallow parsing, *Concept Tuple-based Ontology Learning (CRCTOL)* performs full-text parsing using statistical and rule-based syntactic analysis of documents. It, thus, allows constructing richer ontologies in terms of the range and number of semantic relationships present in the ontology [9].

Regarding other similar to ours domain-specific biomedical entity extraction works, there are some that deal with human disease, gene and protein name extraction: dictionary-based bio-entity name recognition in biomedical literature [10], protein name recognition using gazetteer [11], and gene-disease relation extraction [12]. All these methods are based on static dictionaries for entity extraction, that limits the recall of the system by the size of the dictionary. There is a more effective method based on conditional random fields that has been applied for identifying gene and protein mentions in text [13]. This approach requires annotated training corpora for learning, which is not available for the veterinary medicine domain yet.

Furthermore, there are several emergency surveillance systems that perform automated extraction of animal disease names from web documents such as:

- *BioCaster*[14] (Japan, 2007),
- *MedISys*[15] and *PULS*[16] (European Union, 2007),
- *HealthMap*[17] (USA, 2007).

BioCaster is limited to 50 animal diseases and uses a manually constructed multilingual ontology [14]. It uses support vector machines to extract entity including animal diseases, synonyms [15], viruses and agents [16]. *Pattern-based Understanding and Learning System (PULS)* [17] and *HealthMap* [18] extract as high as 2400 and 1100 disease names respectively (both human and animal diseases). They both are based on dictionary look-up approach and do not recognize any other disease related concepts such as viruses or serotypes.

In this paper we propose an approach for automated construction of a domain-specific ontology, in contrast to other systems that construct general concept ontologies [7], [6], [5], [9], and use this ontology to extract veterinary medicine entities. Similar to other systems [3], [4], we use a semantic relation extraction approach for automated ontology expansion, but by applying a comprehensive set of syntactic patterns and part of speech tagging, we capture non-taxonomic relations between concepts in addition to taxonomic relations.

[9] WordNet - http://wordnetweb.princeton.edu/perl/webwn

[10] UMLS - http://www.nlm.nih.gov/research/umls/

[11] WHO ICD - http://www.who.int/classifications/icd/en/

[12] SNOMED - http://www.nlm.nih.gov/research/umls/Snomed/

[13] Wikipedia - http://www.wikipedia.org/

[14] BioCaster - http://biocaster.nii.ac.jp/

[15] MedISys - http://medusa.jrc.it/medisys/homeedition/all/home.html

[16] PULS - http://sysdb.cs.helsinki.fi/puls/jrc/all

[17] HealthMap - http://healthmap.org/en

III. Methodology

A. Manual Ontology Construction

We manually construct an initial ontology O_{INIT} using lists of diseases retrieved from publicly available domain-specific dictionaries such as: CFSPH[18], DEFRA[19], OIE[20], Wikipedia[21]. After manual merging and deduplication of the abovementioned disease lists, we have 429 concepts in the initial ontology O_{INIT}. Next, we manually discover and update this ontology with sets of synonyms and abbreviations. The size of the manually-updated ontology with synonyms is $|O_S| = 581$ concepts, with abbreviations is $|O_A| = 453$ concepts and with both is $|O_{S+A}| = 605$ concepts. The initial manually-constructed ontology O_{INIT} is expanded with semantic relationships extracted as described in the next section.

B. Automated Relationship Extraction

Our relationship extraction approach is based on discovering semantic relationships between concepts in the collection by using rule-based syntactic pattern matching and part-of-speech (POS) tagging. We look for taxonomic and non-taxonomic linguistic relationships between entities using the initial ontology and raw data from the veterinary medicine domain. There are several relationships that we are interested in, such as:

1) Synonymic relationships of the form "E_1 *is a kind of* E_2", *e.g.,* E_1 = *"swine influenza" is a kind of* E_2 = *"swine fever"*, where E_1 and E_2 are synonyms - different words with identical or very similar meanings.
2) Hyponymic relationships of the form "E_1 *and* E_1 *are diseases*", *e.g.,* E_1 = *"anthrax"*, E_2 = *"yellow fever"* are diseases, where E_1 and E_2 are hyponyms (words that are conceptually included within the definition of another word - their hypernym *disease*, but not synonyms).
3) Causal relationships that capture causative dependencies between diseases and viruses such as "E_1 *is caused by* E_2", *e.g.,* E_1 = *"Ovine epididymitis" is caused by* E_2 = *"Brucella ovis"*.

We present syntactic patterns in Table I for synonymic, hyponymic and causal relationship discovery from text in the domain of veterinary medicine. We use the following notation:

- C_{GEN} corresponds to general "*disease*" concept,
- C_{INIT} represents the concept from the initial ontology,
- C_L represents the learned concept added to new ontology (add C_L learned concept to new ontology O_R if it is not present in the initial ontology O_{INIT}),
- "/" represents a flexible substring within a pattern,
- C_i, C_j correspond to the concepts,
- hyponymic$_{G \to S}$ represents the relationship with learning from general concept to specific,
- hyponymic$_{G \leftarrow S}$ denotes the relationship which is read from right to left using the set of rules.

TABLE I: Syntactic patterns for semantic relationship extraction between concepts: synonymic, causal and hyponymic.

Relationship Type	C_{INIT}	Relationship Pattern	C_L
Synonymic	C_i	"is a" "is a kind of" "and/, | , " "/, /also known as " "/, /is also called "	C_j
Hyponymic$_{G \leftarrow S}$	C_{GEN}	"such as/: | :" "*e.g.,* | for example" "/, for instance /," "including" "/, especially /,"	C_i and/or/, C_j
Hyponymic$_{G \to S}$	C_{GEN}	"and|or other" "/, and|or C_j are" "is a" | ", a"	C_i
Causal	C_i	"is caused by" "causes"	C_j

Let us consider several examples of patterns for:

- synonymic relationship - *"foot and mouth disease is also called FMD"*,
- hyponymic$_{G \to S}$ relationship - *"diseases, for instance baylisascariasis and typeworm"*,
- hyponymic$_{G \leftarrow S}$ relationship - *"west nile virus is a disease"*,
- causal relationship - *"lyme disease is caused by borrelia burgdorferi sencu lato, borrelia afzelii and borrelia garinii"*.

As can be seen through these examples, the relationship extraction phase can be used to improve the descriptiveness of the ontology by including domain-specific semantic relationships between concepts.

C. Automated Ontology Construction

We construct a new ontology O_R using the initial ontology O_{INIT} and semantic relationships extracted by applying syntactic patterns described in Table I. In addition, we use POS tagging[22] to extract n-gram concepts (*e.g., "swine vesicular disease"*). The resulting ontology O_R will contain automatically extracted disease synonyms, abbreviations and viruses.

More precisely, we start with the canonical disease name *"foot-and-mouth disease"* taken from the initial ontology and after processing the sentence "*Foot-and-mouth disease*, *FMD* or *hoof-and-mouth disease (Aphtae epizooticae)* is a highly contagious and sometimes fatal viral disease", we update the ontology O_R with *"foot-and-mouth disease"* $\xrightarrow{is\ a\ kind\ of}$ *"hoof-and-mouth disease"* $\xrightarrow{is\ a\ kind\ of}$ *"aphtae epizooticae"* $\xrightarrow{abbrev.}$ *"FMD"* $\xrightarrow{is\ a}$ *disease*, where $\xrightarrow{is\ a\ kind\ of}$, $\xrightarrow{abbrev.}$ denote synonymic relationships between concepts, $\xrightarrow{is\ a}$ denotes hyponymic$_{G \to S}$ relationships.

After processing the next sentence "*FMD* is caused by *foot-and-mouth disease virus (FMDV)*", we extract a causal relationship between concepts and update the ontology O_R with *"foot-and-mouth disease"* $\xrightarrow{is\ caused\ by}$ *"foot-and-mouth disease virus"* by associating *"FMD"* with its canonical

[18] CFSPH - urlhttp://www.cfsph.iastate.edu/diseaseinfo/animaldiseaseindex.htm

[19] DEFRA - http://www.defra.gov.uk/foodfarm/farmanimal/diseases/atoz/

[20] OIE - http://www.oie.int/eng/maladies/en_alpha.htm

[21] Wikipedia - http://en.wikipedia.org/wiki/Animal_diseases

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

disease name from the initial ontology O_{INIT} and relating *"foot-and-mouth disease virus"* with its synonym *"foot-and-mouth disease virus"* $\xrightarrow{is\ a\ kind\ of}$ *FMDV*.

From the sentence "Pandemic Strain of *Foot-and-Mouth Disease Virus Serotype O*" we extracted serotype of the disease and updated the ontology O_R with *"foot-and-mouth disease virus"* $\xrightarrow{has\ serotype}$ *serotype O*.

D. Entity Extraction

We define the biomedical entity extraction task as the automated extraction of structured information related to animal diseases from unstructured web documents. This task requires the development of an extractor for tagging entities such as: animal disease names (*e.g., "brucellosis"*), their synonyms (*e.g., "Malta fever", "Undulant fever", "Bang's disease", "Gibraltar fever"*), viruses or other causative agents (*e.g., "Brucella abortus", "Brucella canis"*) and serotypes (*e.g., "A+M-", "A-M+", "A+M+"*).

We used an ontology-based pattern matching approach to design a biomedical entity extractor DSEx[23] that takes raw web documents as input and returns a set of attributes for the matching concepts as output.

In Figure 1, we show the attributes that the entity extractor outputs. Let us consider the sentence: "Species infecting domestic livestock are *B. melitensis*$_{DS}$ (goats and sheep, see *Brucella melitensis*$_{DS}$), *B. suis*$_{DS}$ (pigs, see *Swine brucellosis*$_{DS}$), *B. abortus*$_{DS}$ (cattle and bison), *B. ovis*$_{DS}$ (sheep), and *B. canis*$_{DS}$ (dogs)", where tag $_{DS}$ corresponds to animal disease names. The attributes extracted for the first entity in this sentence are: *[41 - 54, B. melitensis, 13, Brucellosis, {Malta fever, Undulant fever, Brucella}, 1]*.

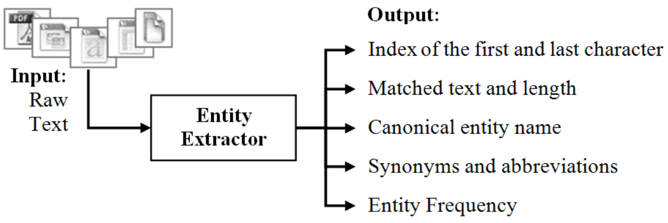

Fig. 1: The output from the entity extractor.

As can be seen from the example above, there are several subtasks of the entity extraction task [19]. The first subtask is *terminology extraction*, which identifies specific relevant concepts named in documents based on the ontology (*e.g.,* disease names, viruses and serotypes). For example, we extract one disease term from the sentence: "Epidemics of *foot-and-mouth disease*$_{DS}$ have resulted in the slaughter of millions of animals".

The second subtask is the *segmentation task*, which means finding the starting and ending character positions of the named entities, for example: "*African swine fever virus*$_{VR,\ 1-25}$ *(ASFV*$_{VR,\ 28-31}$*)* is the causative agent of *African swine fever*$_{DS,\ 60-78}$".

The next subtask is the *association extraction task*, which we consider as a separate prerequisite task for the automated ontology construction in Section III-B. It looks for phrases indicating relationships between entities and matches them against the set of patterns from Table I for inferring associations between diseases, their synonyms and abbreviations (*e.g., "avian influenza" is a kind of "bird flu" is a "H5N1"*) or disease and the causative virus (*e.g., "Brucellosis"* is caused by *"Bacillus abortus"*).

The *normalization subtask* matches all disease names to their canonical versions based on the constructed ontology. For example in the sentence: "*Tick fever*$_{DS}$ is a significant disease of cattle in Australia with up to 7 million animals potentially at risk", the extractor relates *"Tick fever"* with its canonical disease name *"Babesiosis"*.

Algorithm 1 shows the overview of the whole biomedical ontology-based entity extraction process. In the first *"for"* loop the initial ontology O_{INIT} is expanded using semantic relationships. We denote the resulting ontology as O_R. Alternatively, in the second *"for"* loop the initial ontology O_{INIT} is expanded using the *GoogleSets* approach. The resuting ontology is denoted by O_G. The main drawback of using *GoogleSets* is the absence of any explicitly defined relationships between newly-discovered concepts and concepts from the initial set (*e.g., foot-and-mouth disease* and *FMDV* are not related).

After expanding the initial ontology using the two approaches described above, we perform entity extraction at third *"for"* loop using manually-constructed ontologies - O_{INIT}, O_A, O_S, O_{S+A} and automatically built ontologies O_R and O_G. To summarize, the objective of the entity extraction task is to resolve domain-specific terminology extraction, segmentation and normalization subtasks as described above.

Algorithm 1 Biomedical ontology-based entity extraction and semantic relationship discovery using syntactic patterns

```
Input: Two document collections $D_1$ and $D_2$, initial ontology $O_{INIT}$ and other manually-constructed ontologies $O_S$, $O_A$, $O_{S+A}$, sets of patterns from Table I
Output: Automatically-constructed ontologies $O_R$, $O_G$, sets of entities obtained using $\{E_{INIT}\}$, $\{E_S\}$, $\{E_A\}$, $\{E_{S+A}\}$, $\{E_R\}$ and $\{E_G\}$
  for all $d_j \in D_1$ do
    $R_i \Leftarrow ExtractRelation(O_{INIT}, D_1)$;
    $O_R \Leftarrow ConstructOntology(O_{INIT}, R_i)$;
  end for
  for all $\{C_i\} \in O_{INIT}$ do
    $O_G \Leftarrow ConstructOntology(\{C_i\}, GoogleSets)$;
  end for
  for all $d_j \in D_2$ do
    for all $O_i \in \{O_{INIT}, O_S, O_A, O_{S+A}, O_R, O_G\}$ do
      $\{E_i\} \Leftarrow ExtractEntity(\{O_i\})$;
    end for
  end for
```

[23] KDD DSEx - http://fingolfin.user.cis.ksu.edu:8080/diseaseextractor/

IV. Experimental Design and Results

For ontology-based biomedical entity extraction in the domain of veterinary medicine, we aim to extract entities that match at least one concept in the ontology such as a disease or one of its synonyms, abbreviations, causative viruses or disease serotypes. We compared results for domain-specific biomedical entity extraction from different ontologies as summarized in Figure 2:

- first, we used the manually-constructed ontologies O_{INIT}, O_S, O_A, O_{S+A};
- second, we used the ontology O_R obtained based on semantic relationship extraction approach;
- third, we used the new ontology O_G based on *GoogleSets* expansion approach .

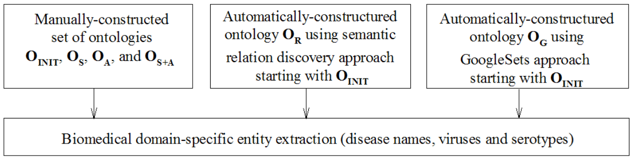
Manually-constructed set of ontologies O_{INIT}, O_S, O_A, and O_{S+A}
Automatically-constructured ontology O_R using semantic relation discovery approach starting with O_{INIT}
Automatically-constructured ontology O_G using GoogleSets approach starting with O_{INIT}
Biomedical domain-specific entity extraction (disease names, viruses and serotypes)

Fig. 2: Summary of the ontologies used for entity extraction.

To compare and evaluate the ontologies that we designed, we retrieved 200 domain-specific web documents using *Google*, including pdfs that report animal disease outbreaks. To avoid any bias, we used first 100 documents to construct the ontology O_R. The other 100 documents were used to evaluate the entity extraction results obtained with all ontologies. The size of the collection that we used for evaluation is constrained by the effort required for manual annotation of the entities. As a result of the entity extraction task, we obtained sets of entities $\{E_1, E_2 \ldots E_n\}$ and their attributes for each document $D_i \in C$ in the collection, as described in Figure 1.

In Figure 3, we report results for the different ontologies we used in terms of precision and recall, where precision represents the number of correctly extracted entities divided by the total number of extracted entities and recall (sensitivity) represents the number of correctly extracted entities divided by total number of existing correct entities in the collection.

As expected, an increase in precision and recall is achieved when switching from the manually-constructed initial ontology O_{INIT} to an ontology which is also manually built, but enriched with synonyms and abbreviations O_{S+A}. Furthermore, the precision and recall values obtained using the automatically-constructed ontologies O_R and O_G are higher compared to the values obtained using the manually-constructed ontologies. As can be seen, the ontology O_R that is built using the semantic relationship extraction approach achieves the highest recall value of 0.83.

Figure 4 presents the ROC curves corresponding to the entity extraction results obtained using different ontologies. As can be seen, the results obtained using manually-constructed ontologies O_{INIT}, O_S, O_A, O_{S+A} are inferior compared to the results obtained using automatically-constructed ontologies O_R and O_G.

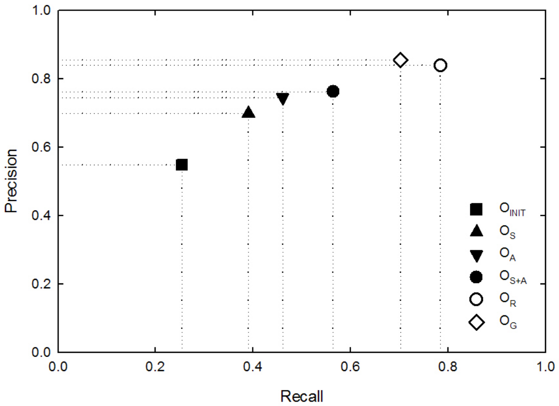

Fig. 3: Entity extraction results using different ontologies. Points from left to right represent the values obtained using: manually constructed ontology O_{INIT} - 429 concepts, ontology with manually-collected synonyms and abbreviations O_{S+A} - 605 concepts, ontology O_G learned using *GoogleSets* expansion approach - 754 concepts, ontology O_R constructed using semantic relationship extraction - 772 concepts.

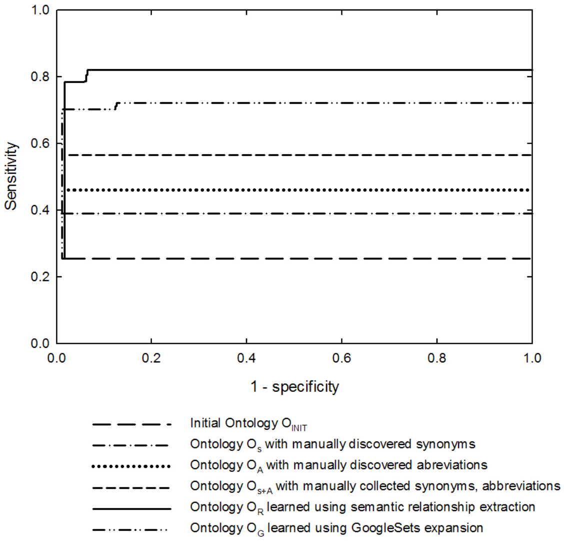

Fig. 4: ROC curves that represent animal disease extraction results for three ontology leaning approaches: baseline - O_{INIT}, O_S, O_A, O_{S+A} and two learned ontologies O_R, O_G.

In Figure 5 we report F-score values obtained by using different ontologies for entity extraction as a function of the ontology size. As we have seen, F-score values increase with transitions from O_{INIT} to O_{S+A} through O_S and O_A. The results obtained using automatically-constructed ontologies O_R and O_G are much higher in comparison to the results obtained using manually-constructed semantic ontologies. However, when the size of the automatically-constructed ontologies O_R and O_G increases, we can see the drop in F-score. It means that we started to add spurious entities and relationships to the ontologies O_R and O_G.

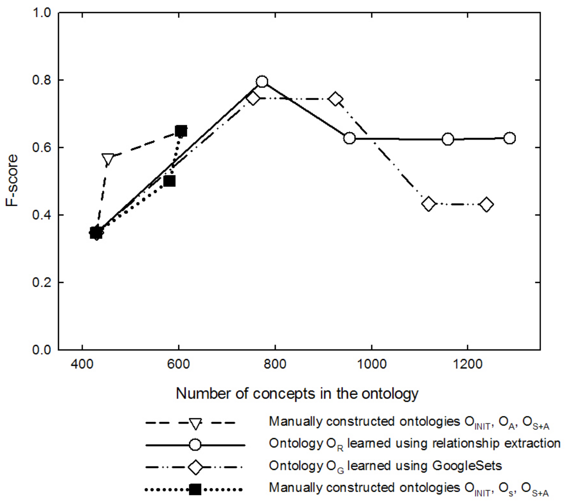

Fig. 5: F-score values as a function of the number of concepts each ontology considered in our experimental design: (1) initial ontology O_{INIT}, (2) O_S with synonyms, (3) O_A with abbreviations, (4) O_{S+A} with synonyms and abbreviations, (5) *GoogleSets* for O_G and (6) and relationship extraction for O_R.

For example, the lowest F-score for the ontology $|O_R| = 1287$ concepts equals 0.63 compared to the highest 0.8 when $|O_R| = 773$ concepts. Similarly, the lowest F-score for the ontology $|O_G| = 1238$ concepts equals 0.43 compared to the highest 0.75 when $|O_R| = 775$ concepts.

All results show that enriching the ontology by discovering additional concepts using relationship extraction or *GoogleSets* expansion approaches, brings new domain-specific knowledge and, therefore, allows boosting domain-specific biomedical entity extraction results. However, the concepts that are newly added to the ontology may add noise if they are based on spurious associations. For instance, results obtained using *GoogleSets* expansion approach for discovering disease synonyms or causative viruses, contain many irrelevant concepts and do not capture any relationship between them explicitly, in comparison to the semantic relationship extraction approach.

V. Conclusions

In this paper, we presented an ontology-based approach for biomedical entity extraction in the domain of the veterinary medicine. We used a semantic relationship extraction approach based on syntactic patterns and POS tagging to construct an ontology (containing animal diseases, their synonyms and viruses). Our experimental results show that the relationship extraction approach boosts the domain-specific biomedical entity extraction results as compared to manually-constructed ontologies enriched with synonyms and abbreviations, and automatically-constructed ontology constructed using the *GoogleSets* expansion approach. Future work plans include automated multilingual ontology construction for the domain of veterinary medicine using other semistructured sources *e.g., Wikipedia*. Furthermore, we plan to enrich the ontology obtained using *GoogleSets* with relationships extracted using the syntactic patterns. At last we will study the effect of the data collection size on the accuracy of the results.

Acknowledgments

This work was supported by a grant from the U.S. Department of Defense. We would like to acknowledge our collaborators from University of Illinois at Urbana-Champaign: ChengXiang Zhai, Dan Roth, Jiawei Han, for their contribution to this research during Data Sciences Summer Institute'09.

References

[1] H. Chen, S. S. Fuller, and C. P. Friedman, *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. Springer, June 2005.

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

[3] M. Missikoff, R. Navigli, and P. Velardi, "The usable ontology: An environment for building and assessing a domain ontology," in *In Proceedings of the International Semantic Web Conference (ISWC*. Springer-Verlag, 2002, pp. 39–53.

[4] H. Davulcu, S. Vadrevu, S. Nagarajan, and I. V. Ramakrishnan, "Ontominer: Bootstrapping and populating ontologies from domain-specific web sites," *IEEE Intelligent Systems*, vol. 18, no. 5, pp. 24–33, 2003.

[5] A. Gómez-Pérez and D. Manzano-Macho, "Deliverable 1.5: A survey of ontology learning methods and techniques," IST Programme of the Commission of the European Communities, Tech. Rep., May 2003.

[6] A. Maedche and S. Staab, "Mining ontologies from text," in *EKAW '00: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*, London, UK, 2000, pp. 189–202.

[7] P. Cimiano and J. Völker, "Text2onto - a framework for ontology learning and datadriven change discovery," 2nd European Semantic Web Conference (ESWC'05), 2005.

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

[9] X. Jiang and A.-H. Tan, "CRCTOL: A semantic-based domain ontology learning system," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 1, pp. 150–168, 2010.

[10] Z. Yang, H. Lin, and Y. Li, "Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature," *Computational Biology and Chemistry*, vol. 32, no. 4, pp. 287 – 291, 2008.

[11] Y. Tsuruoka and J. Tsujii, "Boosting precision and recall of dictionary-based protein name recognition," in *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 41–48.

[12] H. W. Chun, Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii, "Extraction of gene-disease relations from medline using domain dictionaries and machine learning." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 4–15, 2006.

[13] R. Mcdonald and F. Pereira, "Identifying gene and protein mentions in text using conditional random fields," *BMC Bioinformatics*, vol. 6, no. Suppl 1, pp. S6+, 2005.

[14] N. Collier, A. Kawazoe, Y. Tateisi, L. Jin, M. Shigematsu, D. Dien, R. Barrero, K. Takeuchi, and A. Kawtrakul, "A multilingual ontology for infectious disease surveillance: rationale, design and challenges," *Language Resources and Evaluation*, vol. 40, no. 3, pp. 405–413, 2006.

[15] J. Mccrae and N. Collier, "Synonym set extraction from the biomedical literature by lexical pattern discovery," *BMC Bioinformatics*, vol. 9, pp. 159+, March 2008.

[16] 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.

[17] 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.

[18] 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." *Journal Medical Informatics Association*, December 2007.

[19] A. Mccallum, "Information extraction: Distilling structured data from unstructured text," *Queue*, vol. 3, no. 9, pp. 48–57, November 2005.