

# 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 a review of the literature 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 to extract semantic relationships between concepts. Such relationships include asserted synonymy, hyponymy and causality. Specifically, these relationships are extracted by using a set of syntactic patterns and part-of-speech tagging. The resulting ontology contains richer semantics compared to the manually-constructed ontology. We compare our approach for extracting synonyms, hyponyms and other disease related concepts, with an approach where the ontology is expanded using *GoogleSets*<sup>1</sup>, on the veterinary medicine entity extraction task. Experimental results show that our semantic relationship extraction approach produces a significant increase in precision and recall as 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 the World Animal Health Information Database (WAHID)<sup>2</sup>, Food and Agricultural Organization of United Nations Emergency Prevention System (EMPRES)<sup>3</sup>, and unstructured free text *e.g.*, news and official reports from the Department for Environment Food and Rural Affairs (DEFRA)<sup>4</sup>, World Organization

for Animal Health (OIE)<sup>5</sup>, Centers for Disease Control and Prevention(CDC)<sup>6</sup>; medical literature - PubMed<sup>7</sup>, e-mails - ProMED-Mail<sup>8</sup> *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 accuracy of the extraction procedure is crucial, as it has influence on the accuracy of 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 the recognition of domain-specific events such as “*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 direct influence on the event recognition accuracy. The entity extraction accuracy should be maximized in order to produce 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 an ontology as explained in Section III-A and identify semantic relationships (synonymic, hyponymic and causal) using syntactic patterns and part of speech tagging, as described in Section III-B. We then show how to use these semantic relationships to expand the manually constructed ontology and automatically construct a new ontology in Section III-C. In Section III-D we present an overview of the biomedical entity extraction task for the domain of veterinary medicine using an animal disease example. In Section IV, we discuss the results of biomedical entity extraction using manually vs. automatically-constructed ontologies. Furthermore,

<sup>1</sup>GoogleSets API - <http://labs.google.com/sets>

<sup>2</sup>WAHID - <http://www.oie.int/wahis/public.php?page=home>

<sup>3</sup>EMPRES - <http://www.fao.org/ag/againfo/programmes/en/empres/>

<sup>4</sup>DEFRA - <http://www.defra.gov.uk>

<sup>5</sup>OIE - [http://www.oie.int/eng/en\\_index.htm](http://www.oie.int/eng/en_index.htm)

<sup>6</sup>CDC - <http://www.cdc.gov>

<sup>7</sup>PubMed - <http://www.ncbi.nlm.nih.gov/pubmed/>

<sup>8</sup>ProMED-Mail - [www.promedmail.org](http://www.promedmail.org)

we compare the automatically-constructed ontology obtained using our relationship extraction approach with an ontology constructed using a *GoogleSets* expansion approach, which expands a given partial set of objects into a more complete set of similar objects. We compare the entities extracted using all abovementioned ontologies in terms of precision and recall. We also report the F-measure values as a function of the ontology size. The results show that our semantic relationship extraction approach brings new useful 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*<sup>9</sup>;
- structured domain-dependent *e.g.*, *Unified Medical Language System (UMLS)*<sup>10</sup>, *World Health Organization International Classification of Diseases (WHO ICD)*<sup>11</sup>, *Systematized Nomenclature of Medicine - Clinical Terms (SNOMED)*<sup>12</sup>;
- semi-structured domain-independent *e.g.*, *Wikipedia*<sup>13</sup>.

While *WordNet* is a manually constructed lexical database with structured knowledge and *Wikipedia*, by contrast, is an unstructured source of knowledge, neither of them is domain-specific, and thus, they do not contain enough information about infectious animal diseases, their synonyms, viruses and serotypes. Furthermore, domain-specific resources such as *UMLS*, *WHO ICD* and *SNOMED*, cannot be applied for biomedical entity extraction in the domain of veterinary medicine, because they focus on human-related diseases and are poor in terms of animal diseases. Therefore, a unified ontology for the veterinary medicine domain is needed.

The process of 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], among many others systems discussed in [5]. These systems are generally based upon shallow natural language processing techniques and mainly extract concepts with 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 the extraction of 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] described a system that is very effective at 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. Other related works include extraction of semantic relations from web [9] and from bioscience text [10], [11].

By contrast with many ontology learning systems that use shallow parsing, *Concept-Relation-Concept Tuple-based Ontology Learning (CRCTOL)* [12] 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.

Several domain-specific biomedical entity extraction approaches more similar to our approach deal with human disease, gene and protein name extraction. Some examples of such approaches include: dictionary-based bio-entity name recognition in biomedical literature [13], protein name recognition using gazetteer [14], and gene-disease relation extraction [15]. All these methods are based on static dictionaries for entity extraction, which limits the recall of the system by the size of the dictionary. A more effective method based on conditional random fields has been applied for identifying gene and protein mentions in text [16]. However, this method requires annotated training corpora for learning, which is not available for the veterinary medicine domain yet.

At last, several emergency surveillance systems that perform automated extraction of animal disease names from web documents exist, including: *BioCaster*<sup>14</sup> (Japan, 2007), *MedISys*<sup>15</sup> and *PULS*<sup>16</sup> (European Union, 2007), *HealthMap*<sup>17</sup> (USA, 2007). *BioCaster* is limited to 50 animal diseases and uses a manually constructed multilingual ontology [17]. It uses support vector machines to extract entities including animal diseases, synonyms [18], viruses and agents [19]. *Pattern-based Understanding and Learning System (PULS)* [20] and *HealthMap* [21] extract as high as 2400 and 1100 disease names, respectively (both human and animal diseases). They are based on a dictionary lookup 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], [12]. We use this ontology to extract veterinary medicine entities. Similar to other systems [3], [4], we use a semantic relationship extraction approach for automated ontology expansion, but by applying a comprehensive set of syntactic

<sup>9</sup>WordNet - <http://wordnetweb.princeton.edu/perl/webwn>

<sup>10</sup>UMLS - <http://www.nlm.nih.gov/research/umls/>

<sup>11</sup>WHO ICD - <http://www.who.int/classifications/icd/en/>

<sup>12</sup>SNOMED - <http://www.nlm.nih.gov/research/umls/Snomed/>

<sup>13</sup>Wikipedia - <http://www.wikipedia.org/>

<sup>14</sup>BioCaster - <http://biocaster.nii.ac.jp/>

<sup>15</sup>MedISys - <http://medusa.jrc.it/medisys/homeedition/all/home.html>

<sup>16</sup>PULS - <http://sysdb.cs.helsinki.fi/puls/jrc/all>

<sup>17</sup>HealthMap - <http://healthmap.org/en>

patterns and part of speech tagging, we capture non-taxonomic relations between concepts in addition to taxonomic relations.

### III. METHODOLOGY

#### A. Manual Ontology Construction

We manually construct an initial ontology  $O_{INIT}$  using several lists of diseases retrieved from publicly available domain-specific dictionaries such as: CFSPH<sup>18</sup>, DEFRA<sup>19</sup>, OIE<sup>20</sup>, Wikipedia<sup>21</sup>. 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 ontology manually-updated 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 equivalent to  $E_2$ ”, e.g.,  $E_1 =$  “swine influenza” is equivalent to  $E_2 =$  “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  $E_2$  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 - in our example, their hypernym *disease*, but are 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” concepts;
- $C_{INIT}$  corresponds to concepts from the initial ontology;
- $C_L$  denotes learned concepts added to the new ontology (add a  $C_L$  learned concept to the new ontology  $O_R$  if the concept is not present in the initial ontology  $O_{INIT}$ );
- “/” represents a flexible substring within a pattern;
- $C_i, C_j$  correspond to any two concepts;

TABLE I: Subset of syntactic patterns for semantic relationship extraction between concepts.

Relationship Type	$C_{INIT}$	Relationship Pattern	$C_L$
Synonymic	$C_i$	“is a” “is equivalent to” “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 $_{G \rightarrow S}$	$C_{GEN}$	“and/or other” “/, and/or $C_j$ are”	$C_i$
Causal	$C_i$	“is caused by” “causes”	$C_j$

- hyponymic $_{G \rightarrow S}$  represents a relationship from a general concept to a specific concept;
- hyponymic $_{G \leftarrow S}$  denotes a relationship from a specific concept to a general concept (read from right to left);

The following are examples of patterns considered in our approach:

- synonymic relationship - “*foot and mouth disease is also called FMD*”,
- hyponymic $_{G \rightarrow S}$  relationship - “*diseases, for instance baylisascariasis and typeworm*”,
- hyponymic $_{G \leftarrow S}$  relationship - “*west nile virus is an animal infectious disease*”,
- causal relationship - “*lyme disease is caused by borrelia burgdorferi senu lato, 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<sup>22</sup> 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*”  $\xrightarrow{\text{is equiv. to}}$  “*hoof-and-mouth disease*”  $\xrightarrow{\text{is equiv. to}}$  “*aphtae epizooticae*”  $\xrightarrow{\text{abbrev.}}$  “*FMD*”  $\xrightarrow{\text{is a}}$  *disease*, where  $\xrightarrow{\text{is equiv. to}}$ ,  $\xrightarrow{\text{abbrev.}}$  denote synonymic relationships between concepts,  $\xrightarrow{\text{is a}}$  denotes hyponymic $_{G \rightarrow S}$  relationships.

After processing the next sentence, “*FMD is caused by foot-and-mouth disease virus (FMDV)*”, we extract a causal

<sup>18</sup>CFSPH - urlhttp://www.cfsph.iastate.edu/diseaseinfo/animaldiseaseindex.htm

<sup>19</sup>DEFRA - http://www.defra.gov.uk/foodfarm/farmanimal/diseases/atoz/

<sup>20</sup>OIE - http://www.oie.int/eng/maladies/en\_alpha.htm

<sup>21</sup>Wikipedia - http://en.wikipedia.org/wiki/Animal\_diseases

<sup>22</sup>NLTK POS Tagger - http://www.nltk.org/

relationship between concepts and update the ontology  $O_R$  with “*foot-and-mouth disease*”  $\xrightarrow{\text{is caused by}}$  “*foot-and-mouth disease virus*” by associating “*FMD*” with its canonical disease name from the initial ontology  $O_{INIT}$  and relating “*foot-and-mouth disease virus*” with its synonym “*foot-and-mouth disease virus*”  $\xrightarrow{\text{is equiv. to}}$  *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*”  $\xrightarrow{\text{has serotype}}$  *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<sup>23</sup> 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*<sub>DS</sub> (goats and sheep, see *Brucella melitensis*<sub>DS</sub>), *B. suis*<sub>DS</sub> (pigs, see *Swine brucellosis*<sub>DS</sub>), *B. abortus*<sub>DS</sub> (cattle and bison), *B. ovis*<sub>DS</sub> (sheep), and *B. canis*<sub>DS</sub> (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].

As can be seen from the example above, there are several subtasks of the entity extraction task [22]. The first is *terminology extraction*, which identifies specific relevant concepts named in documents based on the ontology (e.g., diseases, viruses, serotypes). For example, we extract one disease term from the sentence: “Epidemics of *foot-and-mouth disease*<sub>DS</sub> 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*<sub>VR, 1–25</sub> (*ASFV*<sub>VR, 28–31</sub>) is the causative agent of *African swine fever*<sub>DS, 60–78</sub> (*ASF*<sub>DS, 60–78</sub>)”.

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 equivalent to “*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*<sub>DS</sub> 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, which is an example of set expansion technique also applied to named entity recognition task [23]. The limitation of using *GoogleSets* expansion approach 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). We denote the ontology automatically-constructed using *GoogleSets* by  $O_G$ .

After expanding the initial ontology using the two approaches described above, we perform entity extraction within the third “*for*” loop using the manually-constructed ontologies  $O_{INIT}$  and  $O_A$ ,  $O_S$ ,  $O_{S+A}$  (obtained by extending  $O_{INIT}$  with synonyms and abbreviations), and the automatically built ontologies  $O_R$  and  $O_G$ . To summarize, the objective of the entity extraction task is to perform domain-specific terminology extraction, segmentation and normalization, 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 ontologies  $O_S$ ,  $O_A$ ,  $O_{S+A}$  obtained by extending  $O_{INIT}$  manually with sets of synonyms and abbreviations; and the set 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
     $R_i \leftarrow \text{ExtractRelation}(O_{INIT}, D_1)$ ;
     $O_R \leftarrow \text{ConstructOntology}(O_{INIT}, R_i)$ ;
end for
for all  $\{C_i\} \in O_{INIT}$  do
     $O_G \leftarrow \text{ConstructOntology}(\{C_i\}, \text{GoogleSets})$ ;
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
         $\{E_i\} \leftarrow \text{ExtractEntity}(\{O_i\})$ ;
    end for
end for

```

---

<sup>23</sup>KDD DSEx - <http://fingolfin.user.cis.ksu.edu:8080/diseaseextractor/>

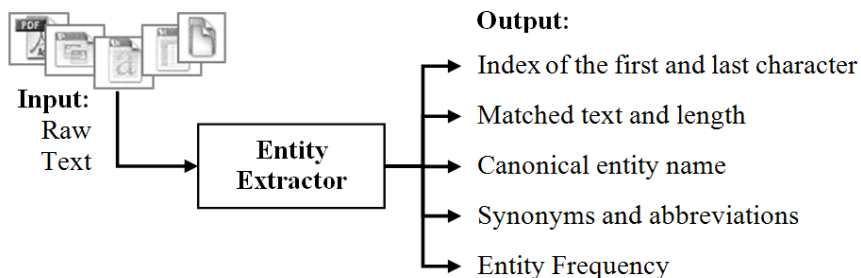


Fig. 1: The output from the entity extractor.

#### IV. EXPERIMENTAL DESIGN AND RESULTS

To evaluate our approach on the task of ontology-based biomedical entity extraction in the domain of veterinary medicine, we have extracted entities that match at least one concept in the ontology, where the concept can be a disease name or one of its synonyms, abbreviations, causative viruses or disease serotypes. We have compared the results for domain-specific biomedical entity extraction from different ontologies as summarized in Figure 2:

- first, we have used the manually-constructed ontologies  $O_{INIT}$ ,  $O_S$ ,  $O_A$ ,  $O_{S+A}$ , where the ontologies  $O_S$ ,  $O_A$ ,  $O_{S+A}$  are obtained from the initial ontology  $O_{INIT}$  by adding synonyms and abbreviations;
- second, we have used the ontology  $O_R$  obtained based on our semantic relationship extraction approach;
- third, we have used the new ontology  $O_G$  based on the *GoogleSets* expansion approach .

To compare and evaluate the ontologies that we have designed, we used *Google* to retrieve 2000 domain-specific web documents (including pdfs) that report animal disease outbreaks. From the 2000 documents, we sampled 200 documents where the distribution of the domain-specific entities was appropriate for the evaluation of the proposed approach (e.g., number of the disease names is more than 5 for each document). We used 100 documents to construct the ontology  $O_R$  and the other 100 documents to evaluate the entity extraction results obtained with all ontologies considered in our study. The size of the collection used to evaluate the extraction results is constrained by the effort required for manual annotation of the domain-specific entities. However, the number of documents that are used for new concept extraction and automated-ontology construction could be potentially increased.

We used both manually and automatically-constructed ontologies to perform domain-specific entity extraction and obtained sets of entities  $\{E_1, E_2 \dots E_n\}$  and their attributes for each document  $D_i \in C$  in the collection, as described in Figure 1.

In Figure 3, we report results obtained using different ontologies, 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)

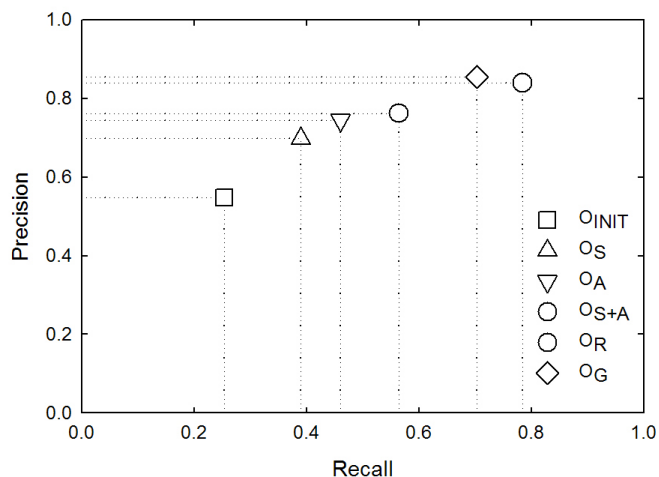


Fig. 3: Entity extraction results using different ontologies. Points from left to right represent the values obtained using: the manually constructed ontology  $O_{INIT}$  (429 concepts), ontology with manually-collected synonyms and abbreviations  $O_{S+A}$  (605 concepts), ontology  $O_G$  extracted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

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 0.84 and precision value 0.77 compared to manually-constructed ontology  $O_{INIT}$  recall 0.25 and precision 0.54.

In Figure 4, we report the F-measure values (harmonic mean of precision and recall) obtained by using different ontologies for entity extraction, as a function of the ontology size. As can be seen, the F-measure values increase with transitions from

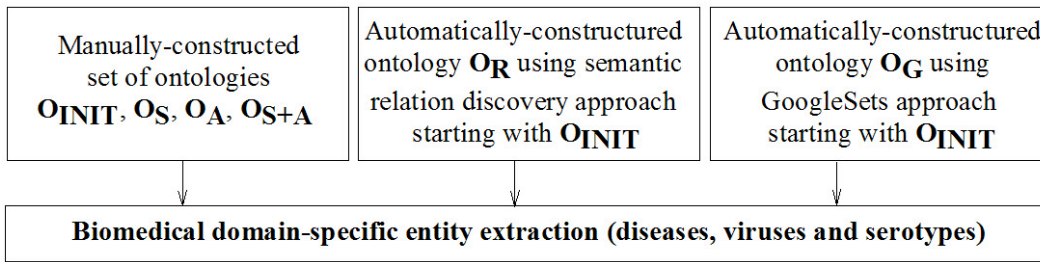


Fig. 2: Summary of the ontologies used for entity extraction.

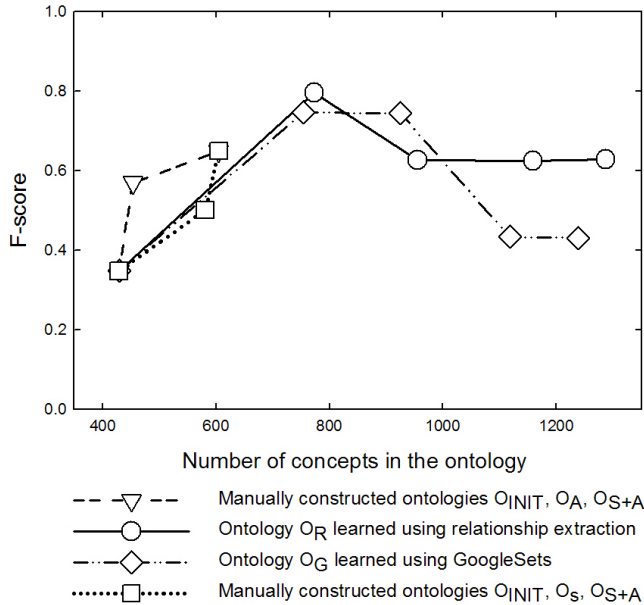


Fig. 4: F-score values as a function of the number of concepts each ontology considered in our experimental design: (1) initial ontology  $O_{INIT}$ , (2)  $O_S$  with synonyms, (3)  $O_A$  with abbreviations, (4)  $O_{S+A}$  with synonyms and abbreviations, (5) *GoogleSets* for  $O_G$  and (6) and relationship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

$O_{INIT}$  to  $O_{S+A}$  through  $O_S$  and  $O_A$ . The F-measure values obtained using the automatically-constructed ontologies  $O_R$  and  $O_G$  are much higher than the values obtained using the manually-constructed semantic ontologies. However, when the size of the automatically-constructed ontologies  $O_R$  and  $O_G$  increases, the F-measure values decrease. This means that the larger the size, the better the chance that spurious entities and relationships have been added to the ontologies  $O_R$  and  $O_G$ .

The lowest F-measure value 0.63 is obtained when the ontology  $O_R$  consists of 1287 concepts, while the highest value 0.8 is obtained when  $O_R$  has 773 concepts. Similarly, the lowest F-measure value 0.43 is obtained when the ontology  $O_G$  consists of 1238 concepts and the highest value 0.75 is obtained when  $O_R$  has 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 relationships. For instance, results obtained using the *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 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 the 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 partially supported by a grant from the U.S. Department of Defense and National Agricultural Biosecurity Center at Kansas State University to William H. Hsu and the NSF grant IIS 0711396 to Doina Caragea. We would like to acknowledge our collaborators from University of Illinois at Urbana-Champaign: ChengXiang Zhai, Dan Roth, Jiawei Han, for their input on this research during Data Sciences Summer Institute 2009.

## 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] R. C. Bunescu and R. J. Mooney, "Learning to extract relations from the web using minimal supervision," in *ACL*, 2007.
- [10] B. Rosario, "Extraction of semantic relations from bioscience text," Ph.D. dissertation, Berkeley, CA, USA, 2005, adviser-Hearst, Marti.
- [11] B. Rosario and M. A. Hearst, "Classifying semantic relations in bioscience texts," in *In Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics*, 2004, pp. 431–438.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] J. Mccrae and N. Collier, "Synonym set extraction from the biomedical literature by lexical pattern discovery," *BMC Bioinformatics*, vol. 9, pp. 159+, March 2008.
- [19] 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.
- [20] R. Steinberger, F. Fuat, E. Groot, C. Best, P. Etter, and R. Yangarber, "Text mining from the web for medical intelligence," *Mining Massive Data Sets for Security*, 2008.
- [21] 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.
- [22] A. Mccallum, "Information extraction: Distilling structured data from unstructured text," *Queue*, vol. 3, no. 9, pp. 48–57, November 2005.
- [23] 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.