Learning User-Defined, Domain-Specific Relations: A Situated Case Study and Evaluation in Plant Science

Ana Lucic and Catherine Blake
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street, Champaign, IL 61801
alucic2@illinois.edu, clblake@illinois.edu

ABSTRACT
Although methods exist to identify well-defined relations, such as is_a or part_of, existing tools rarely support a user who wants to define new, domain-specific relations. We conducted a situated case study in plant science and introduce four new domain-specific relations that are of interest to domain scientists but have not been explored in information science. Results show that precision varies between relations and ranges from 0.73 to 0.91 for the manufacturer location category, 0.89 and 0.93 for the seed donor-bank relation, 0.29 and 0.67 for the seed origin location, and 0.32 and 0.77 for the field experiment location. The manufacturer location category recall varies from 0.91 to 0.94, the seed bank-donor location recall ranges between 0.93 and 1, the seed origin relation from 0.33 to 0.82 while the field experiment location from 0.67 to 0.83 depending on the classifier and using a combination of lexical and syntactic features in the background.

Keywords
Semantic relationship extraction and disambiguation, domain-specific relations, text mining, exploratory search

INTRODUCTION
Relations that bind concepts that co-occur in a sentence can broadly be divided into those that are more general in nature and can be found across domains and those that are more specific in nature and involve entities of special kind (Smith et al., 2005) that only appear in certain domains. More work has been done in the area of identifying and defining more general types of relations that appear across domains whereas little support exists for users to define new, domain-specific relations. Domain-specific relations are tightly connected to domain knowledge and a search for domain-specific relations across a collection of scholarly articles would generally not be considered a “known item search.” Domain-specific types of relations emerge or, better to say, reveal themselves in the aftermath of spending time with the retrieved set of articles, or more specifically, after the user has sufficiently engaged with the retrieved set of articles. Thus, rather than a “known item search,” interest in domain-specific relations would more likely fall under the category of an “exploratory search”—the phrase that has been used to emphasize learning, which differs from item search, navigation, question answering, fact retrieval, and verification (Marchionini, 2006). From a learning perspective, however, current digital libraries support only the first level of Bloom’s taxonomy—knowledge, which focuses on fact retrieval (Bloom, 1956). We envision that digital libraries of the future will support users who are at a higher level in the learning taxonomy, which includes (in ascending order of complexity) knowledge, comprehension, application, analysis, synthesis and evaluation. (Bloom, 1956). To achieve this goal will require that we rethink how a user interacts with a digital collection, and how information from within a collection is represented.

Consider a user that is in the analysis stage of learning which “emphasizes the breakdown of the material into its constituent parts and detection of the relationships of the parts and the way they are organized” (Bloom, ibid, p.144). A user who was analyzing the following two sentences from the Wikipedia article on Albert Einstein would see that both sentences contain a reference to a person (Einstein) and a location (Ulm and Princeton, New Jersey) related to the famous physicist.

Albert Einstein was born in Ulm, in the Kingdom of Württemberg in the German Empire on 14 March 1879.

Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

During the analysis stage, a learner would establish that these sentences actually reflect two different relations, born_in and works_in. In this example, the argument type to each relation is the same (a person and a location), but the
nature of the relations differs. Current digital libraries that employ named entity recognizers provide search support for each of relation arguments but do not allow a user to search for similar relations (i.e. born_in or works_in). Few systems enable a user to constrain both the relation and arguments. Although precoordinated indexing (an example is the Library of Congress Subject Headings) has supported associations between index terms as far back as 1967, Farradae pointed out that such implied relations expressed through precoordinated indexing are unambiguous only in a narrow domain (Khoo et al., 2006).

Automated relation extraction is an active area of research (Moldovan, 2000; Khoo et al., 2006; Frunza et al., 2011; Arighi et al., 2013, Konstantinova, 2014) and approaches tend to cluster around two different strategies. At one end of the spectrum are the domain specific relations that operate over particular entities, as is typified in biomedicine where gene-protein or protein-locations (Kim et al., 2003) have been well explored. At the other end of the spectrum lie tools such as TextRunner (Yates et al., 2007), which identify predicates that are frequently used with the same arguments, where the nature of the relation is not made explicit. Kozareva & Hovy point out that TextRunner does not seek “specific” semantic relations and that to support the patterns should “be able to specify both the actual semantic relation sought and use its textual expression(s) in a controlled manner for maximal benefit.” (2010).

Relation extraction is particularly important to exploratory search, summarization tasks, faceted navigation—an established manner of navigating digital collections through established facets (Hearst & Stoica, 2009) as well as automated subject metadata creation. Although faceted search allows bringing two concepts into connection and finding items that are associated with two or more concepts, the nature or type of these relations typically remains out of reach of faceted search. If identified in an automated way semantic relations have the potential to enrich subject metadata assigned to articles and datasets. Similarly, the rich information contained in scholarly articles, if identified and categorized in an automated way, can be very useful to researchers in different domains as it would allow targeted and exploratory searches. It would also allow linking of the articles through supplemental information contained in articles.

Simply adding relation support is not sufficient for a digital library to support the analysis stage of learning. Rather, the system should enable a user to add and alter relations as they engage in deep reading of materials in a collection. Users naturally annotate physical books and the idea that a digital library should also support annotation is not new (Marshall, 1997). A digital library that supports high level learning simply extends this idea and employs machine learning to scale a small set of user annotations to the remaining collection.

In addition to how a user interacts with a digital library, support for high level learning may require that the underlying representation of content within a digital library evolve from the current word representation, to consider concepts (a.k.a. entities) and relations. Providing relations can improve retrieval performance by enabling a user to fully define their information need more clearly (such as in the Einstein example) and summarization by providing the raw materials necessary to aggregate information from multiple articles. The trade-off, however, is that an increase in semantics typically corresponds to increase in the level of domain specificity. Interestingly, the transition from word to concept to relation is typified in the GENIA experimental collection where the first task involved entity recognition (Kim et al., 2004), the second included event, speculation, and negation (Kim et al., 2009) and the third task involved co-reference and entity relations (Pyssalo et al., 2011).

Our goal in this study is two-fold: 1) we aim to infer the nature of the relation that holds between two entities in an automated way and also 2) using this case study in the background, we aim to provide a model for how a digital library can support high level learning. With reference to the first goal, the system employs a lightly-supervised machine learning approach to predict the nature of the relation that holds between two entities (plants and locations) that occur in the same sentence in a plant science collection. A common assumption with relation extraction tasks is that the type of argument determines the type of relation. The relations in this study are similar to the Einstein example where arguments are the same, but the relations differ. Our study describes a method that was used to infer the nature of the relation in an automated way.

With reference to the second goal we anticipate that scholars and scientists would be early adopters of the proposed system. The system is not yet fully operational, but the example relations were created in collaboration with a domain expert. The target relations are domain specific and have not been explored in information science and so our goal in this paper is to evaluate the overall process and system performance before user testing will be viable.

RELATED WORK
More than a decade ago, Moldovan et al. (2000) made a distinction between general purpose knowledge bases such as Freebase and WordNet and domain-specific knowledge bases such as Medline and Genomes Online Database. Both more general knowledge databases and domain-specific databases are needed for knowledge intensive applications. At the present moment, however, general purpose knowledge databases are both more common and more accessible than domain-specific ones. The development of domain-specific databases, however, according to Moldovan et al. (2000), can be updated and developed in a similar fashion that general purpose databases are developed. The authors proposed a method of establishing seed concepts, identifying noun phrases in which the concepts appear, recording the environment of the noun phrases while paying
special attention to lexico-syntactic patterns that occur between nouns and eventually using the patterns to extract patterns of the same kind in which at least one of the concepts appears as a constituent.

The gap that exists between the relationships encoded in general purpose and domain-specific knowledge bases is filled by domain knowledge. Domain-specific knowledge is specialized and the extent and the accessibility of material in domain specific disciplines differs between domain-specific and more general applications. Freebase contains approximately a half million instances of a more general type of relationship such as book_author while the situation is quite different with more specialized relationships: neither they are so omnipresent (if we take into account journal subscription walls that provide barriers to information) nor so frequent. Thus, data sparsity issues as well as the lack of domain-specific knowledge bases presents itself as one of the barriers to a more efficient way of identifying and extracting relationships from a scientific corpus. However, this is not the only issue. In their review of semantic relations in information science Khoo et al. (2006) concluded that “no systematic analysis of the type of semantic relations used in ontologies has been reported in the literature.” One implication of this conclusion is that there is no easy way of verifying whether the type of ontology or knowledge base that would contain the relations that are of interest to researchers in specialized domains already exists. Li et al. (2008) observe that the patterns that occur in general text may not be applicable in a domain specific area which only further separates general purpose from domain-specific knowledge bases. Similarly, Smith et al. (2005) in their review of biomedical ontological relations come up with the concept primitive relations that identifies domain neutral relations—those that can be found across domains rather than in only one domain. In fact, the located_in relation which is of interest in this study is identified as a primitive instance relation that is found across domains.

Frunza et al. (2011) refer to three major approaches to extracting relationships in science: co-occurrence analysis, rule-based approaches, and statistical methods. The cooccurrence methods assume that a relationship exists between two entities if they occur in the same sentence. These methods are known to provide a good recall but poor precision (Frunza et al., 2011). The reason why the use of this method frequently ends up in a poor precision is the nature of the relationship that holds between entities that can be multifarious. Relationship disambiguation, is an additional layer of semantic relation identification and extraction that requires more attention and scrutiny. Rule-based approaches usually rely on syntactic or semantic information extracted from the text and they require a large amount of human effort in devising rules. The rule-based approaches can be divided into syntactic rule-based approaches and semantic rule-based approaches. Statistical approaches use a variety of algorithms and data representation techniques such as bag-of-words, part-of-speech information, syntactic dependencies, as well as semantic labels to learn the environment in which the concepts appear.

The scholarly articles used in this case study come from the plant science field where a number of text mining and information retrieval applications already exist. Many of these applications are highly specialized and tailored to the particular needs of researchers and biocurators. For example, the web-based interface, PLAN2 (zope.bioinfo.cnio.es/plan2l/overview) (Arighi et al., 2013) classifies sentences in the result set based on relevant developmental processes in higher plants such as flowering, leaf development, root and seed development. The sentence is given a score based on whether it contains a developmental process. Such an entry into literature allows the researchers to extract relevant information and focus on the articles that report or describe these processes faster and more efficiently. Another example is the text mining framework EVEX that was developed for identification of high precise plant data form PubMed articles. It was reported that the tool has the potential to infer new knowledge that still has not been included in knowledge databases by extracting relevant biological events from articles. Based on this experiment, a wider use of text mining and social network analysis methods in the field of plant sciences is predicted (Van Landeghem, 2013). Rather than proposing a text mining framework, this case study presents a method that can either collect information that can be used in a faceted search, provide the basis for a new type of user facilitated indexing, or assist data curators or metadata librarians in their work as they assign metadata to the dataset or scholarly article. The method presented in this paper reports the performance achieved with a limited number of instances in the training set taking into account that a large already annotated corpus, to the best of our knowledge, does not exist and considering that annotating a large set of instances would take a considerable time and effort. This paper is thus interested in the limits and potentials of a smaller training corpus and the limits of a supervised learning method used for domain-specific relation disambiguation task.

**METHOD**

The process comprises the following steps:

1. Define domain-specific relations of interest.
2. Annotate a small set of training examples.
3. Identify features
4. Learn new relation model.
5. Evaluate model
6. Apply to remaining collection.

**Define domain-specific relations**

For this case study, we worked with a domain expert and the members of his research group and used a corpus of plant science articles to identify the candidate relations that hold between a plant and a location. We are not aware of any knowledge base that would contain these particular types of relations. When talking about ontological relations
Smith et al. (2005) imply relations “that obtain between entities in reality, independently of our ways of gaining knowledge about such entities (and thus of our experimental methods) and independently of our ways of representing or processing such knowledge in computers.” They also distinguish between general purpose relations that can be employed in any biological ontologies and specific relations that only apply to biological entities of certain kind. Primitive instance is another term that reflects the domain-neutral and self-explanatory relations. Although location_of is one example of a primitive instance this relation would have been too general for the purpose of identifying very specific relations that hold between a plant and a location that co-occur in the same sentence. Similarly, the Unified Medical Language System Semantic Network includes relations such as spatially_related_to, location_of, traverses, adjacent_to, and surrounds. Although these relations resemble and come close to the types of relations we are interested in predicting in an automated way none of them rises to the level of specificity as the relations that will be described in the following paragraphs. Thus, the relations that are explored in this paper are neither primitive nor domain-independent nor general. They can be considered to be the subtype of a more general location_of or spatially_related_to relations.

The first relationship identified was between a plant and the manufacturer of the tool or instrument that was used with relation to the plant. The following is an example of such relation:

**Arabidopsis thaliana plants** were grown in an environmentally controlled growth room (Korea Instruments, **Seoul, Korea**) at 22degC with a 16-hr-light/8-hr-dark cycle. **139127**

The location in this sentence (Seoul, Korea) refers to the location of the manufacturer (Korea instruments) of the controlled growth room that was used in the experiment.

Another candidate relation identified between a plant and a location can broadly be described as the seed_donor_location relation. This category identifies the donor who donated the seed, typically an individual or seed bank. The nature of these relationships is similar and the amount of semantic overlap between these two categories is high so these two categories were merged. A user may decide to disambiguate between these two types of seed donors (an individual versus a seed bank or an institute) at a later stage. The following are the sentences that represent the seed_donor_bank_location relation:

**The pea (Pisum sativum)** mutants brz and dgl and their parent genotypes cv Sparkle and cv Dippers Gelbe Viktoria were kindly provided by Michael A. Grusak (Children's Nutrition Research Center, **Houston**). **88825** (seed donor relation)

**Seeds of 37 different Antirrhinum cultivars** were kindly provided by Ball Seed Co. (West Chicago, **IL**). **149095** (seed bank relation)

Seed origin represents the third candidate category. This relation indicates the origin of the seed which can be different from the seed donor or seed bank category. This information is important for documenting the experiment process or the result of the experiment. The following is an example of such relation:

**Ulva compressa** was collected in **Cachagua** (32deg 34minutesS), a nonimpacted site of central **Chile**, during spring 2010 and transported to the laboratory in sealed plastic bags in a cooler at 4degC. **3291273**

Although this category is related to the seed_donor_bank_location category, to an expert, the nature of the relationship is different.

The final candidate category identified is the field_experiment_location category which indicates the location of the field experiment. The following sentence is an example of this type of relation:

**Rice plants** were grown in the paddy field of the Hokuriku Research Center, **Joetsu, Japan** (37degrees 06' N, 138degrees 16' E) or in plastic pots filled with soil from the paddy field under outdoor conditions. **2921200**

This sentence conveys the location of the field experiment which is different from the location of the seed origin or the location of the seed donor or seed bank. We see from these examples and sentences that come from the Method section of the articles that although the nature of the entities is the same throughout the examples (plants and locations), the relations that hold between them vary and identifying the right type of relation is a necessary task if the user is interested in only one type of relation.

**Annotate training data**

In this case study, the first author undertook the annotation of 662 candidate sentences that contained a plant and location, which took approximately 40 hours. Of the candidate sentences 110 were set aside for testing, 518 sentences were used for training and 34 were removed because the emphasis was on target relations (rather than the instances that were marked as “Named Entity Recognizer mistake: or “Not Clear” or “Candidate for a New Relation” in the annotated set). A user facilitated indexing that we envision in this paper requires that the user provides the sentences that contain the relation of interest that have been extracted from the article. Given that a user has a limited amount of time and resources to allocate for this activity, we envision a small sample.

Certainly, the annotation task conducted by an individual is always associated with the potential for introducing bias into annotation task and this is what prevents us from reporting
the inter-rater agreement. Future work will explore the possibility of adding the collaborative functionality that would allow reporting the inter-rater reliability as well as the difficulty of identifying these relations.

Identify features
The system is provided with 120 total features comprising entities, lexical and syntactic features that we established would be helpful using the training sentences.

Entity features
Plant names were identified using the Unified Medical Language System (UMLS) semantic type Plant and its 120,199 concept unique identifiers whereas locations were identified using the Stanford Named Entity Recognizer (version 3.2).

Lexical features
Lexical features included the most discriminatory words for each of the four categories. The focus was on the words that appear between a plant and a location. All other words were not considered. Authors in science primarily use passive voice so minimal pre-processing was needed. This means that rather than converting originated or purchased to their base forms originate and purchase, we kept the original appearance of the term and did not use any type of stemming. For the sentences in each category, the words with the highest term-frequency, inverse document frequency weight that occur between a plant and location were extracted and then ranked to obtain approximately 25 words with the highest rank in each category. The intuition behind term-frequency-inverse document frequency measure (tf*idf) is that the words that are relatively rare across documents but more frequent within a document have a higher discriminatory power than the words that are frequent across documents. The final feature vector contained 104 words. The categories that were initially separate—seed_bank_location and seed_donor_location—had 7 terms that overlapped and this was an additional reason to collapse these categories into one. The terms with the highest tf*idf scores that overlapped between these two categories included cv (meaning cultivar), obtained, provided, collected, grown, plants, and seeds).

Syntactic features
Syntactic features were drawn from the Stanford syntactic dependency parser (version 3.2). Figure 1 shows a syntactic parse of the sentence: “Fresh flowers of Crocus sativus, grown under conditions, in field, France, Porcheres, were used throughout the experiment.” (143450). The root of the syntactic tree of each sentence was recorded (feature 1). Also, of interest was the syntactic path from the root of the sentence to the plant and from root of the sentence to the location. In case the sentence contained two plants and/or two locations, the nearest plant or location was taken into account. For example, in this sentence, the root of the tree is the verb used (a syntactic root is usually the main verb of the sentence). In the sentence represented in Figure 1, the syntactic path from the root to the nearest plant includes the following paths: nsubjpass (the relation between used and flowers)/prep (the relation between flowers and of) /pobj (the relation between of and sativus, the head noun of Crocus sativus). The syntactic path from the root of the sentence to the nearest location involves the following path: nsubjpass/prep/pobj/partmod/prep/pobj/prep/pobj. Since the path from the root to the plant or location can sometimes be long as in the last example, the subsets of these paths were also used, in particular the first two and first three dependencies from the root to the plant and location (features 2-7).

Figure 1 – Stanford syntactic dependency parse of the sentence: “Fresh flowers of Crocus sativus, grown under conditions, in field, France, Porcheres, were used throughout the experiment.” 143450

Another feature captured whether or not the location appeared under the same branch as the plant. If it did then the path as well as the subsets of the path (first two and three syntactic paths) that connected the plant and the location were recorded (features 8-10). The syntactic parse of the sentence shown in Figure 1 indicates that the location occurs under the same branch as the plant and so the following syntactic path that connects the plant and the location became the value of this feature account: partmod/prep/pobj/prep/pobj. Additionally, the number of dependencies between the plant and location, if they occurred under the same branch, was taken into account (feature 11). Only in approximately 10% of the sentences in
the training sample, the location appeared under the same branch as the plant.

Other features included the immediate syntactic environment before the location: two and three syntactic dependencies that immediately preceded the reference to the location were also recorded (feature 12-13). As shown in Figure 1, the syntactic dependencies prep/pobj immediately precede the reference to France and thus prep/pobj was used as the feature. The number of syntactic paths that connected the reference to a plant and the reference to a location was also calculated. If a plant and a location did not occur under the same branch then the root of the syntactic tree was used as the connecting point and the length of the path that led from the plant to the root and from the root to the location was taken into account (feature 14). Although not strictly a syntactic feature, we took into account whether a location was immediately preceded by an organization which is a pattern very frequently found with the manufacturer and also seed_bank-donor_location relation. (feature 15). The final feature included the combination of the syntactic path immediately preceding the location and the word associated with that dependency. For example, in the example shown in Figure 1, the syntactic dependency that immediately precedes location was pobj and the word that participates in this dependency was in and so pobj/in was used as a feature (feature 16).

Learn domain-specific relations
The annotated data and the features were used to create a series of binary classifiers (one versus the rest) were run on 518 sentences in the training set in which the sentences that belonged to a specific category were marked as 1 and all other sentences were marked as 0. All models were built using the Oracle Data Miner (ODM) 3.2 toolkit with the default settings: Decision Tree (NB), Generalized Linear Model (GLM), Support Vector Machines (SVM), and Naïve Bayes (NB). The models were first applied to the 110 sentences that had been set aside for testing, and then applied to 1933 sentences that contained one or two plant references and one or two location references, and which were not part of the training or test sets.

Evaluate model
Experiments described in this paper use a collection of plant science articles that are available through PubMed Central collection. The total of 11,601 articles from three journals have been used for exploring the distribution of plant and location entities throughout articles: Plant Physiology (7,352 articles), Plant Cell (3,341 articles) and Journal of Experimental Botany (908 articles).

Three experiments were conducted using lexical features, syntactic features and the combination of lexical and syntactic features. Standard measures of precision, recall, F1 and accuracy were used to compare the result of each set of features for each of the four different relations.

RESULTS
Plant names occurred throughout the different sections of plant science articles, but locations are more likely to appear in the Methods section. Figure 2 shows where in an article locations appear in each of journals and resulted in subsequent annotation sentences to be drawn exclusively from the Method section.

![Figure 2 – Distribution of location references across different sections in the plant science collection](image)

The Method section contains some of the key elements of the experiment process and understanding the relationship between a plant and location referred to in the same sentence can assist in the process of retrieving some of the key information about the study, for example, the origin of the plant or the location of the field experiment.

There was a total of 2,595 sentences that contained one or two plant references and one or two location references in the Method section of the article. Of these, 662 sentences were randomly selected and annotated by the first author to identify the nature of the relationship between a plant and a location. Features were informed by only the 518 sentences in the training set.

Distribution of domain-specific relations
Figure 3 indicates the distribution of relations that hold between a plant and a location in 662 sentences from the training set. The most frequent type of relation (41%) is the one that indicates the location of the manufacturer of the tool or instrument that was used in an experiment. At this point the user may decide to draw additional sentences for the smaller relations (seed origin or seed bank), or they may have exhausted the number of sentences that satisfy this relation and decide to move forward with building the model.
Predictive performance - training set

Figure 4 indicates the Support Vector Machines classifier results for four types of domain-specific relations that hold between a plant and a location in the Method section of the plant science corpus of articles. As these results indicate, the combination of lexical and syntactic features typically helps the performance especially with regard to the F₁ measure. More particularly, precision, F₁ measure, and accuracy with manufacturer_location, seed_bank-donor_location, and field_experiment_location categories resulted in an improvement after syntactic features have been added. The recall measure with the manufacturer_location and field_experiment_location categories do not show an improvement after adding syntactic features. These two categories record a somewhat better recall result when only lexical features are used. Interestingly, with the seed_origin_location category, the use of syntactic features only results in a better recall and F₁ measure while the hybrid approach produces the best precision (67% compared to 53% obtained when using lexical features only and 50% when using syntactic features only).

The results of each of the four classifiers are indicated in Tables 1–4 (best results for each classifier are bolded). Table 1 indicates combining both lexical and syntactic features typically boosts the classifier performance when predicting the manufacturer_location category. Interestingly, though, we don’t see this consistent improvement with the addition of the syntactic features over each of the four relations. Lexical features alone used in combination with Generalized Linear Model and Decision tree classifier to predict the seed_bank-donor_location relation had a better performance.
than syntactic and hybrid method. Support Vector Machines and Naïve Bayes classifiers, however, produced best results with the hybrid feature model for the same category. In general, the model that used syntactic features only consistently performed worse than lexical features alone with the exception of the seed origin category when they resulted in a better performance than both the lexical and hybrid methods. This finding requires further experimentation and verification.

<table>
<thead>
<tr>
<th>GLM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex Syn Hyb</td>
<td>Lex Syn Hyb</td>
</tr>
<tr>
<td>Precision</td>
<td>0.86 0.50 0.91</td>
</tr>
<tr>
<td>Recall</td>
<td>0.94 0.84 0.94</td>
</tr>
<tr>
<td>F1</td>
<td>0.90 0.63 0.92</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.94 0.71 0.95</td>
</tr>
</tbody>
</table>

Table 1 – Manufacturer_location relation disambiguation with four classifiers

<table>
<thead>
<tr>
<th>GLM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex Syn Hyb</td>
<td>Lex Syn Hyb</td>
</tr>
<tr>
<td>Precision</td>
<td>0.92 0.64 0.91</td>
</tr>
<tr>
<td>Recall</td>
<td>1.00 0.69 0.98</td>
</tr>
<tr>
<td>F1</td>
<td>0.96 0.66 0.95</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.95 0.65 0.95</td>
</tr>
</tbody>
</table>

Table 2 – Seed_bank-donor_location relation disambiguation with four classifiers

<table>
<thead>
<tr>
<th>GLM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex Syn Hyb</td>
<td>Lex Syn Hyb</td>
</tr>
<tr>
<td>Precision</td>
<td>0.38 0.31 0.36</td>
</tr>
<tr>
<td>Recall</td>
<td>0.42 0.83 0.33</td>
</tr>
<tr>
<td>F1</td>
<td>0.40 0.45 0.35</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.86 0.78 0.86</td>
</tr>
</tbody>
</table>

Table 3 – Seed_origin_location relation disambiguation with four classifiers

<table>
<thead>
<tr>
<th>GLM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex Syn Hyb</td>
<td>Lex Syn Hyb</td>
</tr>
<tr>
<td>Precision</td>
<td>0.71 0.32 0.71</td>
</tr>
<tr>
<td>Recall</td>
<td>0.83 0.75 0.83</td>
</tr>
</tbody>
</table>

F1 | 0.77 0.45 0.77 0.73 0.44 0.80
Accuracy | 0.95 0.80 0.95 0.92 0.79 0.95

Table 4 – Field_experiment_location relation disambiguation with four classifiers

Figure 5 indicates the average performance across all four categories using Support Vector Machines classifier and shows that the hybrid method results in the best average precision of 82%. The best average recall is achieved with lexical method (92%). The best F1 measure (81%) and accuracy (95%) are achieved with a hybrid method (81%). Although these results seem quite promising, the average performance for automatically separating the four categories of semantic relation clearly received a boost from a rather good performance achieved with the prediction of the manufacturer location category. This is the most frequent category in the training set—approximately 40% of the instances belonged to this category—and apparently easiest to predict. Seed origin location and field experiment location categories were least frequent categories (approximately 8% of the instances in each category) and also hardest to predict.

Figure 5 – Average performance with disambiguation of 4 domain-specific relations

Predictive performance - test set
To evaluate generalizable performance, the SVM model that was created from the training set was applied to the remaining 1,967 sentences that contained one or two plant
and location terms. A random sample 200 sentences, 50 instances for each category, was then evaluated manually.

Figure 6 indicates the precision with which the classifier predicted each of the relation categories. The SVM identified manufacturer_location relations with the highest precision of 94% which is consistent with the training set results (91%. Table 1). The seed_bank-donor_location category had precision of 58% which represents a significant drop from the 91% precision achieved in the training set (Table 2). The seed_origin_location and field_experiment_location categories were harder to predict (36% and 42% precision respectively) and the precision achieved also represents a drop from the training set evaluation.

![Figure 6 – Test set precision on 200 sentences (50 in each category)](image)

The results in this experiment show that one of the four relations had a good general performance on the test set, and that the SVM model may have overfitted to the training set, which occurs when the model is too tightly connected to the training samples and does not generalize well. If the analytics were part of a digital library then the user could be provided with feedback on the accuracy and then they could decide if they want to annotate additional sentences to improve performance. These results use the default settings, but a user could also change their preference for precision or recall based on their task.

**Discriminatory features**

Several of the classifiers are white-box so the user can explore which features played the greater role in differentiating the relations. One interesting result is that the syntactic dependencies immediately preceding location reference had the highest coefficient and thus the highest discriminatory power with relation to the type of semantic category. For example, the following syntactic dependencies path preceding the location reference were most indicative of the field_experiment_location category:

- /root/prep/pobj
- /pobj/nn
- /pobj/prep/nn

Some of the most discriminatory lexical terms included: field, grown, growing, and planted.

For the seed_origin_location category, the following syntactic dependencies occurring immediately before the reference location were also some of the most discriminatory features in the hybrid model:

- /prep/pobj
- /pobj/appos/nn
- /nsubjpass/appos
- /pobj/prep/pobj

Also, the combination of the syntactic dependency that immediately precedes a reference to location with the term /pobj/to was amongst the top 10 discriminatory features in the seed origin location model. Not surprisingly, collected and harvested were amongst the most discriminatory terms for this category.

For the seed_bank-donor_location category, the most discriminatory lexical terms included the terms such as: obtained, provided, purchased, and cv (meaning cultivar).

The syntactic dependency pobj in combination with the preposition of immediately preceding a reference to a location was in the top ten Support Vector Machines classifier features and so was the syntactic path /dep/prep/pobj immediately preceding the reference to a location.

The most discriminatory terms for the manufacturer_location category include some of the instrument names such as camera and microscope, also the verb measured, and the words such as model, medium, solution. The thread that connects these words is their association with instruments and tools used in an experiment which in turn can signal the nature of the semantic relation that holds between a plant and a location (manufacture location category).

Of the syntactic dependencies that characterize immediate environment of the location in the manufacturer location category apposition (more specifically, /appos/dep, and /appos/dep/nn) was one of the most promising indicators of this category.

**Error analysis**

Several wrongly classified instances were analyzed. The following sentence was wrongly identified as the field_experiment_location category. One problem with this sentence was that it should not have been identified as a candidate sentence as it does not contain a location:

To observe the morphology of plants grown on soil, the seedlings were grown in a growth chamber (Biotron; LH-200-RDS, NK system) under a 16-h-light (100–130 µE m⁻² s⁻¹)/8-h-dark photoperiod at 23°C. 3440233
Biotron, an organization entity, is wrongly identified as the location. The sentence also uses the word grown twice which was one of the most discriminatory words for the field experiment category and yet, as this example shows, grown is a frequently used word in the collection of plant science articles that can but may not always be associated with the field experiment location category. In this instance, it seems as though the combination of the words that can be associated with more than one category (grown) and the mistake of the Named Entity Recognizer led to a false positive result and eventually a lower precision result.

Consider the following sentence:

Sanuki Gold’ kiwifruit, Actinidia chinensis Planch, were obtained from a commercial orchard in Kagawa, Japan. 3254691

This sentence was classified as the seed_bank-donor_location category although it actually expresses the seed_origin_location. However, when using the seed_origin_location classifier, the same sentence was correctly classified as the seed_origin_location category sentence. The sentence contains the verb obtained which is one of the most discriminatory words for the seed bank-donor. We see from this example that the verb obtained is by no means restricted to only one category and also that a voting scheme amongst the classifiers might be able to produce better results. For example, choosing the category of the sentence based on the output of all four classifiers and selecting the category associated with the highest probability may lead to a better performance. We leave this experiment for later work.

The method described in this paper first identifies the entities of interest and then focuses on their interrelations and tries to disambiguate them using a supervised learning method. Given that our lexical and syntactic patterns depend on correct identification of the entities of interest and establishing their boundaries, this model depends and relies on the existence of named entity recognizers that are able to identify the entities with sufficient precision and recall. The method used the UMLS semantic category Plant to recognize plant names and Stanford Named Entity Recognizer to recognize locations. While these two methods of recognizing entities performed rather well they also introduced certain mistakes and ambiguities. For example, one frequent ambiguity related to this particular task/problem was the location reference that was part of the plant name. For example, Valencia orange is a type of a plant. And, using the UMLS semantic category Plant Valencia Orange was recognized as the plant in the text, however, the problem was that Stanford Named Entity Recognizer also recognized Valencia in Valencia orange as the location reference. While Valencia in Valencia Orange is a nod to Valencia in Spain which has a reputation for sweet oranges, we ideally wanted this particular location reference to be identified as a plant name rather than a location reference. In fact, many plant names contain a location reference within the plant name as, for example, Paris in Paris daisy, Jerusalem in Jerusalem artichoke and Lisbon in Lisbon lemon. An additional amount of preprocessing was needed to separate locations that were truly references to locations from those that were part of plant names.

Although plants and locations were the primary entities of interest, during the process of feature extraction it became obvious that this method can benefit from correct identification of other named entities in the text, for example, organizations and personal names. Manual analysis of sentences revealed that if a location was immediately preceded by an organization and both of the entities were included in parenthesis this type of pattern was more indicative of the manufacturer location or seed bank location category than the other two categories. Similarly, if an organization was preceded by a personal name this was more indicative of a seed donor category. This example indicates that the method described in this study relies on the use of named entity recognizers that can recognize several types of entities with sufficient precision and recall.

FUTURE WORK

The initial set of sentences were drawn at random from the methods section based on where entities of interest were located. The system could be extended to strategically select sentences for annotation that included a wide array of surface level features, or by employing a method such as in South et al. (2012) where active learning was used to select sentences from the collection that would have the highest impact on the learned model (South et al., 2012). Thus far we have situated this work within an environment where the domain expert and his team of researchers provides the annotations for their project. However, a fully integrated system would increase the opportunities for collaborative annotation by either a community of domain experts or metadata librarians who could be provided with the initial annotations and then asked to manually separate the categories that are less prevalent but still useful, key elements of the study, such as, for example, seed_origin_location and field_experiment_location categories.

A collaborative environment would also enable new features to be shared. In this work we employed standard features, but enriching the features with information about the main verb class may improve classification performance. As new relations are identified by a community, the system may also be able to propose new relations that might be of interest, based on use statistics, and by interacting with users to identify subsets of features that seem to work well for different tasks, i.e. users may move towards annotating at a feature rather than a sentence level.

Perhaps the most important next step is to integrate the proof-of-concept evaluated in this paper into a user interface that can be deployed with actual users. Annotation systems such as general architecture of text engineering (GATE, https://gate.ac.uk/) and the brat rapid annotation tool (BRAT, http://www.nactem.ac.uk/brat-annotation/) have been
created for specific user communities, but have not been tightly coupled with content. Such tools have limited collaborative functionality where users would want to verify their analysis before making the annotations publically available.

CONCLUSION
This case study demonstrated that a common assumption—that the nature of the arguments determines the relation between the arguments—does not always hold. The analysis that focused on plant and location entities that occur in the same sentence revealed up to 6 types of relations between these two entities in the collection of plant science articles. The relations that have been identified, to the best of our knowledge, are new and, more importantly, they are domain-specific and not general in nature. This study developed a method for relation type prediction that represents a crucial step when dealing with new types of domain-specific relations that have not been explored previously.

This study also introduced a new method that supports user-defined domain specific relations that are associated with high level learning—in particular the activities associated with the analysis stage in the Bloom’s taxonomy of learning—by enabling a user to define new relations of interest that the system then scales up. User facilitated indexing of literature that focuses on the identification of interesting relations in literature and on the annotation of a small sample of such relations can enable a new model of interaction with scholarly literature as well as more sophisticated results to researchers’ complex search needs.

The method used in this study relies on the existence of high precision named entity recognizes and syntactic parser and, in addition to facilitating a user indexing of the article, it has the potential to assist biocurators, data curators, as well as metadata librarians with assigning metadata to datasets and articles. While the method described here does not boast a sufficient precision and recall to invoke building of an application or to allow an implementation of a faceted search it nevertheless represents a step in this direction. Further work will involve introduction of a new type of feature, such as for example, semantic features that would focus on the verb class of the main verb in the sentences and the arguments of the verb.

The results of the experiments indicated that lexical features alone have a considerable discriminatory power to disambiguate between four categories. Depending on the category, the addition of syntactic features to the model typically boosted the performance. More particularly, the hybrid model achieved the best performance with three out of four categories. For the seed_origin_location category the syntactic method proved more useful and provided better results than lexical features only or a combination of lexical and syntactic features.

ACKNOWLEDGMENTS
This research is made possible in part by a grant from the U.S. Institute of Museum and Library Services (IMLS), Laura Bush 21st Century Librarian Program Grant Number RE-05-12-0054-12 Socio-technical Data Analytics (SODA). We thank Dr. Andrew Leakey and the members of his research group for their invaluable expertise advice during data collection and relation identification phases of this case study.

REFERENCES
* Systems that do allow relation and argument search, include MEDIE (http://www.nactem.ac.uk/medie/) and text runner (https://code.google.com/p/newstepr/wiki/TextRunner), but such systems do not provide support for user defined relations.


the Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland.


