Automatic Metadata Extraction from Museum Specimen Labels

P. Bryan Heidorn
University of Illinois
Champaign, IL, USA
pheidorn@uiuc.edu

Qin Wei
University of Illinois
Champaign, IL, USA
qinwei2@uiuc.edu

Abstract

This paper describes the information properties of museum specimen labels and machine learning tools to automatically extract Darwin Core (DwC) and other metadata from these labels processed through Optical Character Recognition (OCR). The DwC is a metadata profile describing the core set of access points for search and retrieval of natural history collections and observation databases. Using the HERBIS Learning System (HLS) we extract 74 independent elements from these labels. The automated text extraction tools are provided as a web service so that users can reference digital images of specimens and receive back an extended Darwin Core XML representation of the content of the label. This automated extraction task is made more difficult by the high variability of museum label formats, OCR errors and the open class nature of some elements. In this paper we introduce our overall system architecture, and variability robust solutions including, the application of Hidden Markov and Naïve Bayes machine learning models, data cleaning, use of field element identifiers, and specialist learning models. The techniques developed here could be adapted to any metadata extraction situation with noisy text and weakly ordered elements.

Keywords: automatic metadata extraction; machine learning; Hidden Markov Model; Naïve Bayes; Darwin Core.

1. Introduction

"Metadata can significantly improve resource discovery by helping search engines and people to discriminate relevant from non-relevant documents during an information retrieval operation" (Greenberg, 2006). Metadata extraction is especially important in huge and variable biodiversity collections and literature. Unlike many other sciences, in biology researchers routinely use literature and specimens going back several hundred years but finding the information resources is a major challenge. Metadata and data extracted from natural history museum specimens can be used to address some of the most important questions facing humanity in the 21st century including the largest mass extinction since the end of the age of the dinosaurs. What is the distribution of (the) species on earth? How has this distribution changed over time? What environmental conditions are needed by a species to survive?

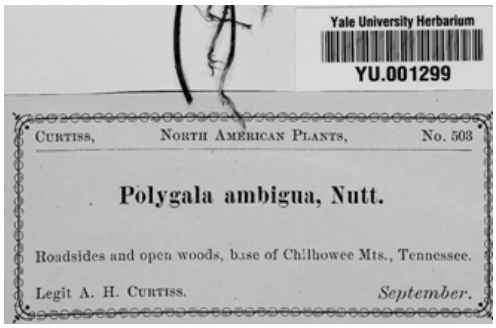

FIG. 1 Example Museum Specimen Label

There are over 1 billion specimens in museums worldwide collected over the past several hundred years. These specimens have labels (see Figure 1 for an example label) and catalog entries containing critical information including, the name of the species, the location and date of collection, revised nomenclature when the taxonomic name was changed, the habitat where it was found such as marsh or meadow as well as many other pieces of information. This knowledge will allow us to better predict the impact of global climate change on species distribution (Beaman, 2006). However, only a small fraction of this specimen data is available online. Consequently, digitization has become a high priority globally. Recent advances in digital imaging make it possible to quickly create images of specimen labels. However, the usefulness of the scanned images is limited since images cannot be easily manipulated and transformed into useful information in databases and full-text information systems. Optical Character Recognition (OCR) has proven to be useful but also challenging because of the age and variety of museum specimens. As is the situation with biomedical literature (Subramaniam, 2003), because of the volume and heterogeneity of the data it is difficult and expensive for humans to type in and extract critical information by hand. Automated and semi-automated procedures are required. "Results indicate that metadata experts are in favor of using automatic metadata generation, particularly for metadata that can be created accurately and efficiently. … metadata functionalities which participants strongly favored is running automatic algorithm(s) initially to acquire metadata that a human can evaluate and edit," (Greenberg, 2006).

Research on museum labels is also important to other digitalization projects, eg. collection digitization in libraries. In general, the techniques developed here could be adapted to any information extraction situation of noisy text and with weakly ordered elements. In this paper, we discuss noisy-text extraction in more complex documents than in most prior works (e.g. Kahan,1987; Takasu, 2002; Takasu, 2003). Most noisy-text classification research is focused on how to automatically detect and correct the OCR errors, text segmentation, text categorization and text modeling (e.g. Takasu, 2002; Takasu, 2003; Foster, 2007). Some techniques that are used to reduce the effect of OCR introduced imperfections include: combining prior knowledge, N-grams, morphological analysis, and spatial information. Our research is focused on how to automatically extract metadata from noisy text using machine learning with limited training data. Since the output of handwriting OCR is still extremely poor, we limit our analysis below to labels that are primarily type written. Our experimental results demonstrate the effectiveness of exploiting tags within labels, and collection segmentation to improve performance.

The paper is organized as follows. Section 2 is a discussion of the properties of museum label metadata and information extraction challenges. Section 3, details how this problem has been addressed in other contexts, especially in the "address" and "bibliographic entry" problem. Section 4 details the system architecture, algorithm and the performance of the algorithms. Section 5 presents the conclusion and future work.

2. Metadata Properties

The research objective is to develop methods to extract an extended element set of Darwin Core (DwC) from herbarium records. DwC is an extensible data exchange standard for taxon occurrence data including specimens and observations of (once) living organisms. DwC has been adopted as a standard by the Biodiversity Informatics Standards (formerly the Taxonomic Database Working Group: http://darwincore.calacademy.org/). We extend the DwC to 74 fields that are particularly useful in museum specimen label context. Nearly 100% of the original label content can be assigned to some element. The 74 elements and their meanings are presented in Table 1. Some codes are optionally preceded with an "R" to indicate re-determination or appended with an "L" to indicate a field element label/identifier as discussed in section 4.3.

TABLE 1: 74 Elements and Element Meaning

Code	Element Meaning	Code	Element Meaning	Code	Element Meaning
ALT[L]	Altitude [Label]	HD	Header	PPREP	Possession Transfer Preposition
BC	Barcode	HDLC	Header Location	PTVERB	Possession Transfer Verb
BT	Barcode Text	[R]IN	[Re-determination] Institution	[R]SA	[Re-determination] Species Author
CD[L]	Collect Date [Label]	INLC	Institution Location	SC[L]	Species Code [Label]
CM[L]	Common Name [Label]	LATLON	Latitude and Longitude	[R]SN[L]	[Re-determination] Species Name [Label]
CN[L]	Collection Number [Label]	LC[L]	Location [Label]	SP	Species
CO[L]	Collector [Label]	MC[L]	Micro Citation [Label]	TC[L]	Town Code [Label]
CT	Citation	NS	Noise	TGN	Type Genus
DB[L]	Distributed By [Label]	OIN	Original Owning Institution	THD	Type Label Header
DDT[L]	Determination Date [Label]	OT	Other	TSA	Type Species Author
[R]DT[L]	[Re-]Determiner [Label]	PB[L]	Prepared By [Label]	TSP	Type Species
FM[L]	Family [Label]	PD[L]	Description [Label]	TY	Type Specimen
FT[L]	Footnote [Label]	PDT	Possession Transfer Date	[R]VAA	[Re-determination] Variety Author
[R]GN	[Re-determination] Genus	PIN	Possessing Institution	[R]VA[L]	[Re-determination] Variety [Label]
HB[L]	Habitat [Label]	PPERSON	Person Doing Possession Transfer		

The key problems with extracting information in this domain are heterogeneity of the label formats, open-ended vocabularies, OCR errors, and multiple languages. Collectors and museums have created label formats for hundreds of years so label elements can occur in almost any position and the any typography and script: hand written, typed and computer generated. In addition to typographic OCR errors, in these labels OCR error are also artifacts of format and misalignment (e. g. See ns(Noise) elements for OCR errors in the following xml example). These errors have several causes including: the later addition of data values to preprinted labels, label formats often included elements that are not horizontally aligned or because new labels were added to the original, making it difficult for OCR software to properly align the output. Following is the OCR output of the label in Figure 1 and the hand markup xml document. This markup is

the target output for HLS and the format of the training and validation datasets. The tags indicate the semantic roles of the enclosed text.

OCR output of Figure 1:

```
^
¶£,&&¶
I ]   CUKTISS,
(}    -----------------
Poly gala ambigua, Nutt.
{¶>  Roadsides and open woods, b.ise of Chllhowec Mts., Tennessee.  5
Q
O  Legit, A. H. Cubtiss.
September.  9
```

XML markup of the OCRed text:

```
<?xml version="1.0" encoding="UTF-8"?>
<?oxygen RNGSchema="http://www3.isrl.uiuc.edu/~TeleNature/HERBIS/semanticrelax.rng" type="xml"?>
<labeldata><ns>^
 ¶£,&amp;amp;amp;¶
 I ]   </ns><co cc="Curtiss">CUKTISS,</co>
<ns> (}    -----------------
</ns><gn cc="Polygala"> Poly gala</gn><sp> ambigua,</sp><sa> Nutt.
</sa><ns> {¶&lt;</ns><hb>  Roadsides and open woods,</hb><lc cc="Base of Chilhowee Mts., Tennessee"> b.ise of
Chllhowec Mts., Tennessee.</lc><ns>  5
 Q
 O</ns><col>  Legit,</col><co cc="A. H. Curtiss"> A. H. Cubtiss.
</co><cd> September.  9</cd>
 </labeldata>
```

3. Related Work

3.1. Evaluation Measures and Cross Validation

Before we introduce the related work, two important evaluation concepts are needed: F-score and K-fold cross-validation. The F-score is widely used in information retrieval and extraction evaluation and is calculated based on precision and recall (see following F equation). Generally speaking, the higher the F-score, the better the results. Precision is defined as the ratio of tokens correctly assigned to a class divided by the total number assigned to the class. The recall is the ratio of correctly classified tokens for a class divided by the total number of tokens of that class.

F = 2*precision*recall/(precision + recall) ***(F equation)***

Since trainings and validation data is expensive and time consuming to gather, K-fold cross-validation is frequently used to evaluate machine learning effectiveness in order to get more reliable results. Cross-validation is the statistical practice of partitioning a sample of training data into subsets so that machine learning is performed on one subset, while the other subset(s) are used to confirm the validity of the machine learning output (Witten, 2005). 5-fold, 10-fold and leave-one-out cross validation are very popular. In 5-fold validation the system randomly partitions the training set into five equal subsets. On each of 5 iterations, the machine learner will use one of the subsets for testing and use the other 4 sets as the training set.

3.2. Machine Learning Driven Information Extraction

Several automatic metadata extraction methods have been studied, e.g. hand-coded rule-based parsers (e.g. Han H. et al., 2005) and machine learning (e.g. Han, 2003; Borker, 2001). For highly structured tasks rule-based methods are easy to implement. The resulting rule system is usually domain-specific and can not be easily translated for use in other domains. Machine learning, on the other hand, is more robust and efficient (Han, 2003). Several learning models are available.

Among the most popular are the Naïve Bayes model (NB), the Hidden Markov Model (HMM), Support Vector Machines and Expectation Maximization. Supervised machine learning (SML) algorithms include training data and machine self-correction based on errors in machine performance against the training set. HMM and NB are discussed with more details in Section 4.

Substantial research has been conducted on the usefulness of ML in Information Extraction (IE). The most relevant prior research has been conducted on U.S. Postal address and bibliographic data. (e.g. Lewis, 1994; Frasconi, 2002; Borkar, 2001; Han, 2003; Hu, 2005, Cui, 2005). Borkar et al. developed a HMM system, similar to an algorithm we use, to handle the information extraction task (Borkar, 2001). The methods for "segmenting unformatted text records into structured elements" they proposed are successful in solving a simple U.S. postal address problem. They reported F-scores of 99%, 88.9% and 83.7% respectively on datasets of USPS addresses, Student addresses and Company addresses. For bibliographic data, they achieved an F-score of 87.3% for 205 records from Citeseer by using 100 training records. Han et al. implement a Support Vector Machine as the classifier to extract mainly Dublin Core metadata from Citeseer and EbizSearch, using 10-fold cross-validation on 500 training headers and 435 test headers. Their method achieves an overall accuracy of 92.9%. Cui's dissertation (2005) demonstrated that domain knowledge gained from machine learning models in one publication is very useful for improving the performance of IE in another publication in the same field. This is a necessary property of some machine learning algorithms we need to move the HERBIS Leaning System across herbarium collections.

Table 2 documents some of the differences between the address and the museum label information extraction problem and demonstrates the need for the new algorithms discussed below. This is an analysis of 200 U.S. addresses and 200 HERBIS (http://www.herbis.org/) "printed" label instances. The work cited above demonstrated that 200 records are sufficient for this type of analysis. The US address data and museum labels are randomly selected from regular USPS mail envelops and the HERBIS label database which includes more than 20,000 records from the Yale Peabody Herbarium. This herbarium was founded in 1864 and containing 350,000 specimens. We would expect similar results in similar Herbaria collections. Address labels and Museum labels were processed in exactly the same manner: image scanning followed by OCR and then markup. The museum label data is substantially more complex than the postal data (see Table 2).

TABLE 2. Statistics about experimental collections

Collection Statistics	**HERBIS**	**USPS Address**
Record count	200	200
Number of elements to recognize	74	10
Average number of words per instance	50	6.5
Approximate OCR error rate (error words/ total words)	15%	1%
Total number of element transitions	4736	969
Average fan-out factor	7.76	2.78
Average number of elements per instance	23.6	4.85
HMM F-score	76.9%	95.2%

The museum labels differ from prior datasets along a number of dimensions:

(1) Structure and order: Museum label data has a much looser structure than the address and bibliographic data. In spite of the fact that some of the museum labels are pre-printed and have a specific structure, there are still thousands of different formats. Some of the elements may appear anywhere of the original label (e.g. barcode, common name). Some elements are intertwined in natural language sentences. A particularly troublesome example is the mixing of habitat and location information e.g. "In boggy soil, 3.5 miles northeast of Deer Mountain."

The orderliness of these labels is reflected in the transitional probabilities. The transitional probabilities are the non-zero probability of one element follows another. This can be summarized in several ways, including the "Total Number of element transitions." This is a count of arcs connecting one element to another. This number is somewhat biased by the number of elements. The "Average Fan-out factor" is the average number of elements that can follow another. The value of 7.76 for museum labels means that on average any element can be followed by any 7 different elements.

(2) Variability within elements: Dictionary aided classification is usually unavailable in herbaria label data set. In Address problems, the proper name is an open class but the other elements are much more finite. The number of states, cities within a state and roads within a city are finite. In contrast, there are on the order of 1.5 million scientific names. The International Plant Name Index (IPNI) (http://www.ipni.org/) contains many thousands of entries but is far from complete particularly for older names that have been replaced yet appeared on museum labels. There are also variations in spelling because of human error or changes in nomenclatural rules. The list of all Collectors is also exceedingly long and labels do not follow any single authority. The location where a specimen was found is also an open class. It includes descriptions of locations, e.g. "300 meters NNW of the last rapids on Stanley Falls, Belgium Congo".

4. HERBIS Architecture

The museum domain is much more complex than the address problem as showed above and information extraction accuracy using the previously developed methods are inadequate. In this section we discuss methods we have used to enhance performance by extending the methods used for previous data sets. The goal of the learning phase of machine learning is to use representative examples to develop models that can, when presented with novel input, create proper classification of the input. Our first training data consists 200 digitized OCR records from the Yale Peabody Herbarium with multiple label formats randomly selected from the typed labels which requiring 10,095 element classifications.

4.1. Deployment

The HERBIS Learning System (HLS) is part of the overall HERBIS system. Museums anywhere in the world can create digital images of their specimens on their site. These images can be passed to the Yale Peabody Museum OCR processing unit where the label is detected and converted to a string sequence. This text packet is sent to HLS at UIUC though a web services connection. The text is converted to an XML document with appropriate information labeled and returns them to the end user. Other image handling services such as MorphBank (http://www.morphbank.net) can call the classification programs directly.

4.2. Learning Phrase: Application of HMM and Naïve Bayes

HLS uses a modified Hidden Markov Model(HMM). The HMM algorithm is discussed elsewhere (Borkar, 2001). The HMM induces a probability distribution on sequences of symbols. The HMM model is an order-preserving algorithm. There are three canonical problems associated with HMM could be solved by different algorithms. One of them is useful in information extraction context. Given the output sequence (O_1 O_2 O_3. . . O_t), find the most likely sequence of hidden states (S_1 S_2 S_3. . . S_t) that could have generated a given output sequence. In other words, given the word sequence ("Polygala ambigua, Nutt."), find the most likely sequences of element (i.e. gn(genus),sp(species),sa(species author)). This problem is solved by the Viterbi algorithm.

A Naïve Bayes (NB) model is a probability model based on conditional probabilities. The NB model makes predictions based on the probability distribution of features from the training set. The NB algorithm uses the distribution information to calculate the probabilities that a new instance belonging to the classes. The example would then be classified to the highest probability class. For computational efficiency NB assumes that each feature is conditionally independent of

every other feature (Mitchell, 1997). This "independent" assumption greatly simplifies the model but the assumption is far from accurate in many cases. However, the overall classifier works surprisingly well in practice (Witten, 2005). The NB calculations are imbedded in part of the HMM algorithm.

In order to show the HMM performance comparing to others, we also implemented a non-ordered algorithm NB as the baseline and then present of series of extension to HMM below. The following example used the training data that was enhanced by including both the original OCR errors in the training set plus examples where the OCR errors were hand corrected. The difference from this correction is small so we only present the difference between HMM and NB on 41 elements that occur more than 20 times in the training set (Figure 2).

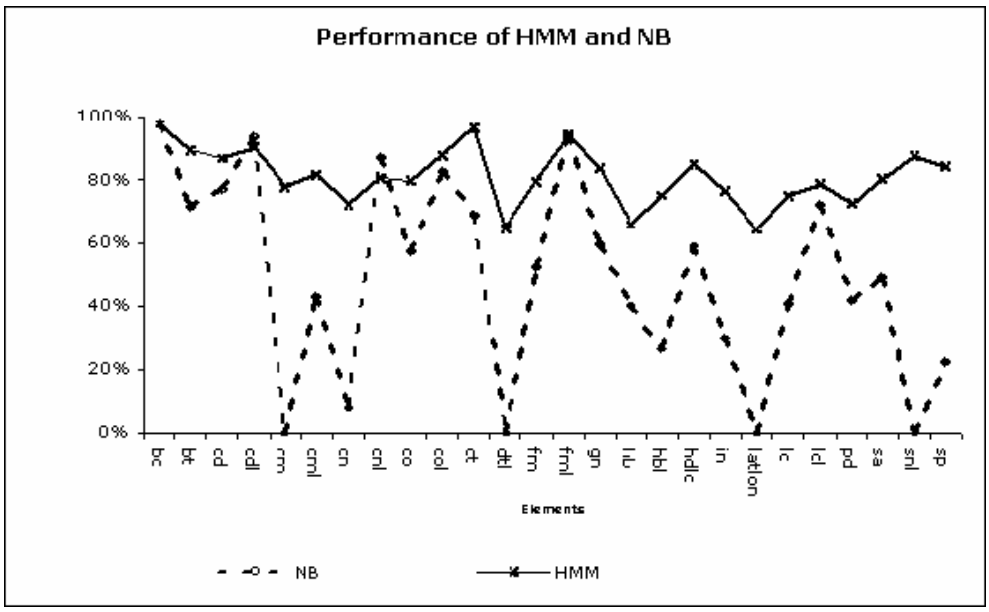

FIG. 2. Performance of HMM and NB

4.3. Field Element Identifiers

There is a set of elements in our dataset which we call "field element identifiers" (FEI). Some elements of some data labels are preceded by a string to identify the information that follows. For example, the term "Legit" in the string "Legit A. H. Curtiss" or "No." in "No. 503" in Figure 1. In the museum label training data and machine learning output, we mark these with a terminal "L", e.g. COL(collector label), LOL (location label), HBL (habitat label). Those label elements usually indicate that there is respectively a CO(collector), LC(location), HB(habitat) element following it, except in cases of missing data and alignment errors.

Rather than training the HMM algorithm to extract the Darwin Core elements and treat these other elements as NS(noise), we train the algorithms to recognize the field element identifiers as well. Our result shows that those label elements improve the ML overall 4%. Figure 3 presents the detailed performance differences between with label encoding in the schema and without those field element identifiers.

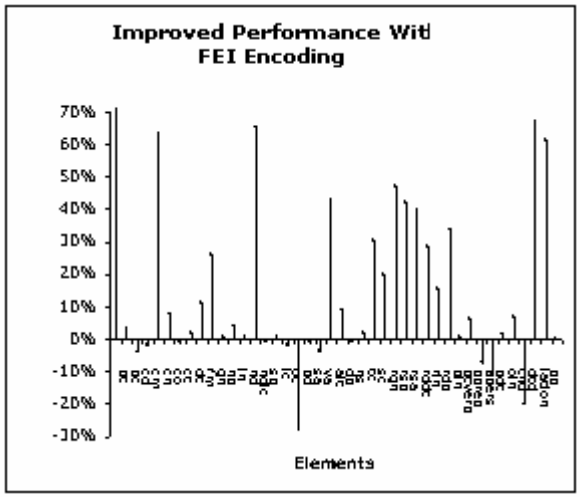

FIG. 3 Improved Performance With FEI Encoding

4.4. Dataset Segmentation and Social Computing (multiple User Feedback)

It is very difficult to improve the performance the ML without large numbers of training examples (Witten, 2005). Unfortunately, it is very expensive to get botanists to create these examples because creating the training examples from the raw OCR output is very time consuming.

The analysis above indicated that the performance difference between the USPS address and HERBIS collection are mainly attributable to the relative homogeneity in the format of the USPS addresses. There may be thousands of different formats of labels that have evolved over the last couple hundred years and now reside in museum collections. However, each collector has their own preferred format of label. This means that a particular museum will tend to have a relatively finite number of collectors supplying the museum at any one time and therefore will have a finite number of label formats represented in the collection. Further, if many museums are digitizing labels, then eventually, there will be corrected sets of labels for many collectors. It may be possible to develop multiple training modules each of which specializes in a particular collector and therefore label format. This observation leads to the hypotheses that the specialist model will perform better for records by the same author than for a generalized model trained on a random data collection and That fewer training examples will be required to reach a given level of performance using all labels from the same collector than would be required for a mixed collection of collectors. These hypotheses are supported in the results of the experiment below.

HLS includes the following Specialist Bootstrapping Architecture (SBA) (see Figure 4). Rather than following the standard machine learning model of creating training data >> generate model >> deploy model, we design a model where multiple museums could use available models to classify their data but as part of their workflow when they correct the machine learning data to put into their own database those examples are added to a new training pool. This pool can be subdivided into sub-collections to construct new specialist models (for particular collectors or collections).

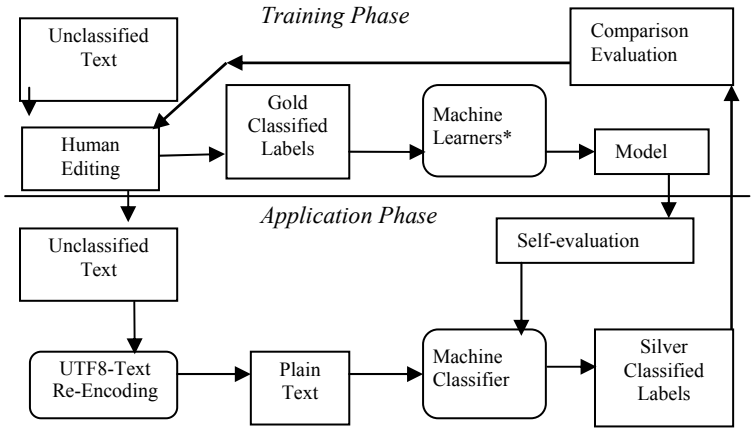

FIG. 4. Specialist Bootstrapping Architecture (SBA) for HERBIS
(*Machine Learners" in the diagram is one of many specialist learners.)

When the end-user sends a museum image to the server, the server would perform OCR, classify based on collector and then process the document with the appropriate collector or collection model. If a specialized collector module is not contained in the server, the information will be extracted from the label using the generic model based on a random sample of labels (see specialist learning algorithm below). For this strategy to work, it is necessary to be able to categorize labels into subsets prior to the information extraction step so that the highest performance model could be used for extraction. A Naïve Bayes pre-classifier can successfully perform this task. The 200 generic Yale training set includes 15 records from the collector "A. H. Curtiss". The 5-fold evaluation of NB classifier trained to differentiate "Curtiss" from "non-Curtiss records" preformed well, F-Score of 97.5%.

Bootstrapping is a process where a small number of examples are used to create a weak learning model. This learning model, while weak is used to process a new set of examples. When a museum staff member corrects the output, it can be added to their database. The new result can help to form a stronger model. There are fewer errors generated by this new model making it easier for the users to correct the model's errors. Museum staff who digitize records need to perform this step for key fields in any case in order to import the records to their database. These corrected examples are fed back into the process again to create an even stronger model. Successive generations of examples improve performance making it easier for the users to generate more examples.

A user wishing to create their own specialized model could begin by processing a set of labels from one collector through the generic Yale model. With each iteration the performance of the specialist system would improve but initially the generic model would perform better, with fewer errors per record. At some crossover point, the performance of the specialized model would exceed that of the generic model. In the example below the crossover point is at about 80 examples. In this framework the user only needs to correct machine output for 80 records to create a model that performs as well as a random collection of 200 records. This crossover point is what the algorithm is looking for in Phase 2 step 7 below.

Specialist Learning Algorithm --The steps could be described as follows:

Phase 1 (generic model)

1. Developers create a "generic" model alpha, M_0.
2. Developers create an empty training data set for User i (U_i) Training Set I, $\{T_i\}$.
3. Set best model $M_b = M_0$.
4. Go to Phase 2

Phase 2 (specialist model learning)

1. U_i runs a small unlabelled data set through M_b.
2. The system returns the newly labeled data (perhaps imperfect).
3. U_i fixes the errors, returns the fixed-labeled-data back to a learner.
4. The system adds the Records to $\{T_i\}$.
5. The system generates a new model M_i base on the $\{T_i\}$.
6. The system evaluates performance (p) of M_i and saves in performance log (L_i)
7. If $p(M_b) > p(M_i)$ set $M_b = M_i$
8. If U_i is satisfied with $p(M_i)$ got to Phase 3 else repeat Phase 2.

Phase 3 (specialized model application)

1 U_i runs any number of unlabelled data set through M_i.

2 The system returns the newly labeled data (perhaps imperfect).

4.5. Experiments and Result Analysis

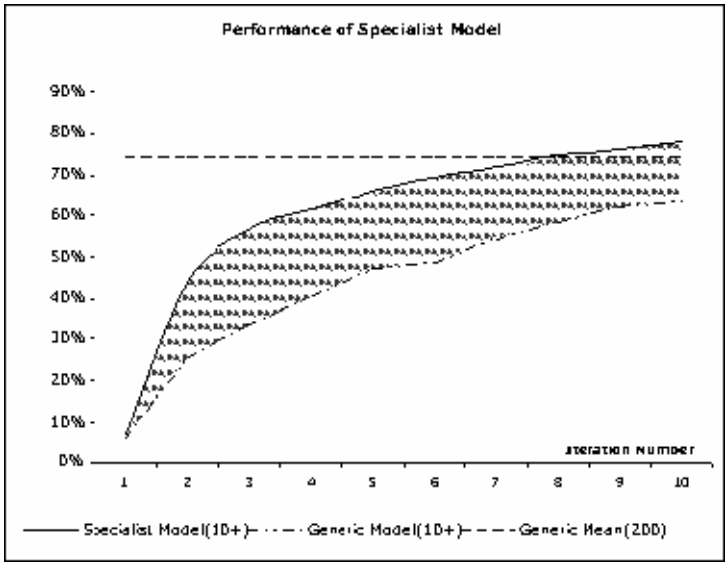

FIG. 5 Improved Performance of Specialist Model

This experiment compares the specialist model and the generic model generated from Yale 200 example collection. The dashed top line in figure 6 is the performance of 200 records independent of iteration. Regular expressions were applied to the 20,000 Yale digitized labels to identify the approximately 100 examples with the collector's name "A. H. Curtiss" who is a well-known collector and botanist. HLS was trained on 10 examples and then 5-fold evaluation used to measure the F-Score. This procedure was repeated 10 times, adding 10 new labels on each iteration producing a training set of 20, 30 and so on until a hundred were used in a training set. The results are presented in the solid curved line, "Specialist Model(10+)." Note that after the specialist model reaches 80 training records it matches the performance of the generic model trained on 200 randomly selected records. The dashed curved line at the bottom, Generic Model(10+), shows the performance of the learning algorithm when given comparable numbers of randomly selected training examples (not necessarily Curtiss) on each iteration. The shaded area is the advantage of using the specialist classification model. If we extended this dashed line out to 200 cases we would see the general model equal to the 200 case general Yale model. This

is not demonstrated here since only 100 Curtiss examples exist in the 20,000 labels digitized at Yale. As predicted, fewer training examples are needed to reach a given level of performance using the Curtiss Specialist collection than a random collection. Given the effectiveness of the NB pre-classifier introduced in the previous section to identify collectors we should be able to create a specialist model for any collector. In fact, we can create a swarm of models for an arbitrary number of collectors and associated label types. The fact that there are only 100 Curtiss labels out of the 20,000 at Yale is a reflection of the fact that there are many labels and many formats.

5. Conclusion and Future work

Hidden Markov and Naïve Bayes models are potentially valuable tools for metadata extraction in herbarium labels but creation of sufficient data sets is a significant barrier to the application of machine learning. The number of required training examples and the associated work can be greatly reduced by establishing collaboration among museums digitizing their collections to support social machine learning. While the current system is a necessary prerequisite for an effective metadata generating system the machine learning swarm has not been implemented or tested with live data. Also, no sufficient user interface exists to deliver a functioning system. In creating such an interface a new set of research questions arise. Standard precision, recall and F-Scores are not sufficient for evaluating interactive systems. A more appropriate measure for botanists would be: How much time this system could save the expert when creating metadata? Important variables are the number of human corrections required per label, the time required to correctly complete a fixed number of labels, number of training examples and number of error corrections needed to meet some performance criteria such as a 90% F-score and other measures.

A number of options exist to improve underlying system performance. For example, label records might be processed in different orders to maximize learning and minimize error rate. OCR correction might be improved using context dependent automatic OCR correction. Dictionary lookup has been used extensively in automatic OCR correction. Context dependent correction means conducting the correct after knowing the word's class. For example, word "Ourtiss" should be corrected as "Curtiss". If the system already identified "Ourtiss" as collector, we can use the smaller collector dictionary instead of using a much larger general dictionary to do the correction. We proposed this method could get a better performance than just dictionary lookup.

Acknowledgements

This research was funded in part by the National Science Foundation, Grant #DBI-0345341.

References

Abney, Steven. (2002). Bootstrapping. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA* (pp. 360-367).

Beaman, Reed S., Nico Cellinese, P. Bryan Heidorn, Youjun Guo, Ashley M. Green, and Barbara Thiers. (2006). HERBIS: Integrating digital imaging and label data capture for herbaria. *Botany 2006, Chico, CA.*

Borkar, Vinayak, Kaustubh Deshmuk, and Sunita Sarawagi. (2001). Automatic segmentation of text into structured records. *ACM SIGMOD, 30*(2), 175-186.

Cui, Hong. (2005). *Automating semantic markup of semi-structured text via an induced knowledge base: A case-study using floras*. Dissertation. University of Illinois at Urbana-Champaign.

Curran, James R. (2003). Blueprint for a high performance NLP Infrastructure. *Proceedings of the HLT-NAACL 2003 workshop on Software Engineering and Architecture of Language Technology Systems,* (pp. 39-44).

Foster, Jennifer. (2007). Treebanks gone bad: Parser evaluation and retraining using a Treebank of ungrammatical sentences. *IJDAR,* (pp. 129-145).

Frasconi, Paolo, Giovanni Soda, and Alessandro Vullo. (2002). Hidden markov models for text categorization in multi-page documents. *Journal of Intelligent Information Systems*, *18*(2-3), 195-217.

Greenberg, Jene, Kristina Spurgin, and Abe Crystal. (2006). Functionalities for automatic metadata generation applications: A survey of experts' opinions. *Int. J. Metadata, Semantics and Ontologies, 1*(1), 3-20.

Han, Hui, C. Lee Giles, Eren Manavoglu, Hongyuan Zha, Zhengyue Zhang, and Edward A. Fox. (2003). Automatic document metadata extraction using support vector machines. *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries,* (37–48).

Han, Hui, Eren Manavoglu, Hongyuan Zha, Kostas Tsioutsiouliklis, C. Lee Giles, and Xiangmin Zhang. (2005). Rule-based Word Clustering for Document Metadata Extraction. *ACM Symposium on Applied Computing 2005 March 13-17, 2005, Santa Fe, New Mexico, USA,* (pp. 1049-1053).

Hu, Yunhua, Hang Li, Yunbo Cao, Li Teng, Dmitriy Meyerzon, and Qinghua Zheng. (2005). Automatic extraction of titles from general documents using machine learning. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, June 07-11, 2005, Denver, CO, USA.*

Kahan S., Theo Pavlidis and Henry S. Baird. (1987). On the recognition of printed characters of any font and size. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 9*(2), 274-288.

Lewis, David D., and Marc Ringuette. (1994). A comparison of two learning algorithms for text categorization. *Proceedings of SDAIR, 3rd Annual Symposium on document Analysis and Information Retrieval*.

McCallum, Andrew K., and Dayne Freitag. (1999). Information extraction using HMMs and shrinkage. *Papers from the AAAI-99 workshop on Machine Learning for Information Extraction*.

Mehta, Rupesh, R. Pabitra Mitra, and Harish Karnick. (2005). Extracting semantic structure of web documents using content and visual information. *Special interest tracks and posters of the 14th international conference on World Wide Web WWW '05, Chiba, Japan,* (pp. 928-929).

Mitchell, Tom. M. (1997). *Machine learning*. McGraw Hill Higher Education.

Subramaniam, L. Venkata, Sougata Mukherjea, Pankaj Kankar, Biplav Srivastava, Vishal S. Batra, Pasumarti V. Kamesam, et. al. (2003). Information extraction from biomedical literature: Methodology, evaluation and an application. *Proceedings of the Twelfth International Conference on Information and Knowledge Management, New Orleans, LA,* (pp. 410-417).

Takasu, Atsuhiro and Kenro Aihara. (2002). DVHMM: Variable Length Text Recognition Error Model. *In Proceedings of International Conference on Pattern Recognition (ICPR02)*, Vol.3, (pp. 110–114).

Takasu, Atsuhiro. (2003). Bibliographic attribute extraction from erroneous references based on a statistical model. *Proceedings of the 2003 Joint Conference on Digital Libraries.*

Witten, Ian H., and Eibe Frank. (2005). *Data mining: Practical machine learning tools and techniques (Second Edition)*.