Label Annotation through Biodiversity Enhanced Learning

P. Bryan Heidorn
University of Arizona
heidorn@email.arizona.edu

Qianjin Zhang
University of Arizona
zhqjnj@gmail.com

Abstract

The LABELX (Label Annotation through Biodiversity Enhanced Learning) is an extension of the HERBIS NLP system reported previously (Heidorn & Wei, 2008). The objective of the system is to formaly structure output from Optical Character Recognition (OCR) of the highly variable labels of natural history museum specimens. OCR errors are common in the OCR output. Genus and species names are particularly prone to errors. Records are preprocessed using a fuzzy-match algorithm to find and replace genus and species names, including those with OCR errors, and replace those with a constant token. Integers and strings that begin with Alphabetic characters and end with numbers are also replaced with tokens. LABELX generates structured XML data and RDF and makes corrections to OCR errors in some fields. The main algorithm is a Hidden Markov Model (HMM). This poster reports an enhancement to the previous system with a larger data set.

Keywords: OCR, parsing, semantic markup, machine learning

Introduction

LABELX is a set of programs designed to process herbarium label data into a format that can be easily ingested into museum databases and distributed in standard formats over the Internet. “T(t)wo thirds of the collections have less than 75% of their collection data online. Images are less accessible than other data: more than 90% of collections have less than 10% of their holdings represented online in digital images.” (Skog et al., 2009). Collections of plants have existed since the beginning of civilization. In the 1540s (Pavord, 2005) the collection of plants took on what we would call scientific organization. According to the Index Herbariorum there are about 3,400 herbaria in the world (Thiers, continuous). The labels on specimens are a rich source of information about not only the name of plants but their historical distribution and environmental requirements. Early labels were of course handwritten. In the 20th century the labels begin to be typed but even then scientific names are handwritten. Even within typewritten labels there is a great variety of formats. Formats are created by individual collectors or by curators of particular collections or by the project managers for biodiversity surveys or other projects.

For type-written labels Optical Character Recognition (OCR) is a potentially valuable method to convert label images into machine readable UTF-8 format. In the resulting text the elements of the label will still be in varying orders. Labeling the elements of the label is the focus of the current study. Unfortunately, OCR on labels are prone to errors. Our objective in this study is to properly label the elements of these labels in spite of the OCR error rate. These OCR errors have a number of sources. The fonts being used in labels are often non-standard and vary within and across labels. Typewriter ribbons were not always of high quality so there are partial letters. Typewriters sometimes had loose pinions so letters do not line up with one another. Many of the words that appear in labels are not in standard OCR dictionaries so the algorithms built into many OCR engines are not able to successfully guess full words based on the context of remaining characters.

Acknowledgement: Funding provided by National Science Foundation, #0956271 to Bryan Heidorn. We thank Steven Chong for assistance in gathering data and RDF output.

Heidorn, P. B., & Zhang, Q. (2013). Label annotation through biodiversity enhanced learning. *iConference 2013 Proceedings* (pp. 882-884). doi:10.9776/13450

Sample Labels

The labels in the current sample come from two sources. The first set of 130 labels was from the Yale Herbarium similar to those reported in 2008 (Heidorn & Wei, 2008). The original Yale OCR collection was relatively small with just a couple of hundred items and the two-thirds with the best OCR were selected for analysis. 200 new labels were added from a research collection from the University of Alaska, Museum of the North digitization project. The new labels from the Museum of the North are part of 55,000 labels we have on file. The selected records are the first from that file excluding records with poor OCR. In the current project we identified 55 different element types in the labels. The current analysis includes 40 elements because the remaining items appeared less than 20 times in the collection making machine learning methods ineffective.

The Algorithms

If the museum labels were highly regular it would be possible to write regular expressions to extract information for the labels. The labels are not that regular but there are some common patterns that can be exploited. Some label elements include human readable tags that were used to tell humans where to type content on a label. Some labels contain the string "S/N:" to signal scientific name. Others contain the string "Name:" and others contain no indication the scientific name follows except the general location on the label. Sometimes this is right after the collector number. A scientific name is composed of predictable parts so Genus is followed by species, which is usually followed by authority. To exploit these regularities two main algorithms were used in this study. The first is Hidden Markov Models (HMM). The performance of HMM in this domain has been reported elsewhere (Heidorn & Wei, 2008). This algorithm which is used in genetics and other domains identifies sets of orderings of elements that best matches previously seen orders of elements (Frasconi, Soda, & Vullo, 2003).

The contribution in this study is the use of fuzzy matching to identify genus and species using an authority file. Scientific names and collector names are difficult for HMM because they are relatively rare. A training set may have several hundred entries but each name may appear only once. We could preload the HMM with many thousands of names from an authority file but this leads to performance problems. Instead, we replace Genus and Species with fixed tokens. For scientific names we use the International Plant Name Index (IPNI). IPNI is a database of the names and associated basic bibliographical details of seed plants, ferns and lycophytes. IPNI is a dynamic resource, depending on direct contributions by all members of the botanical community (IPNI, 2012). Collector names, Genus names and species names are not always unique and may sometimes be other words in English or other languages such as "rose". Simple substitution leads to false positives. OCR errors leads to false negatives. This fuzzy algorithm takes these conditions into account in most cases. If both the genus name and adjacent species name match the entries in IPNI they are replaced with a constant "BiSciColGenus" and "BiSciColSpecies" respectively. If only the genus name is found in IPNI the algorithm creates a list of just species of that genera from IPNI. It then measures the Levenshtein distance between the unknown word and the species names. If a threshold of .8 is met the string is assumed to be a misspelling of the species and is substituted with the string "BiSciColSpecies". The match from IPNI is recorded as the correct spelling. In the following example the genus has an OCR error. The algorithm finds the species as an exact match and then recognizes the genus.

1) The original misspelled full genus names (H1126562):
   AGRCSTIS EXARATA Trin.
2) The substituted result: BiscicolGenus BiscicolSpecies Trin.
   The spelling correction: AGROSTIS EXARATA Trin.

Conversely, if an exact match is found for a species name but the string before that exact species match is not a genus, the algorithm collects all genus names that are valid for that species name from IPNI. The Levenshtein distance is calculated between the unknown prior string and the members of this genus list. If any meet the threshold the unknown string is replaced with the constant "BiSciColGenus." The match from IPNI is recorded as the correct spelling. The following table (Table 1) lists the fields with performance changes of greater than +/- 3%. Species author names increase because of ordering effects detected by the HMM because of improvement of Genus and species performance. Labels of fields are excluded since they are not placed in databases. Deleterious effect of number substitution are seen in altitude and date fields. These will be countered by adding explicit matching for common date formats.

Habitat accuracy decreases because scientific names sometimes appear in habitat descriptions as associated species. The loss of common name accuracy is attributable to confusion with collector names.

Table 1
Change in F-Score

Head	Control	Treatment	%Diff
Location	.709	.747	+3.84
Genus	.771	.812	+4.07
Species	.768	.806	+3.81
Species author	.699	.758	+5.93
Collector name	.760	.796	+3.60
Collection number	.589	.670	+8.16
Barcode	.463	.919	+45.56
Latitude/Longitude	.232	.343	+11.07
Habitat	.699	.665	-3.41
Determiner name	.746	.683	-6.32
Common name	.762	.672	-8.99
Altitude	.577	.498	-7.84
Determination Date	.739	.656	-8.28

Conclusion

Fuzzy Matching techniques such as Levenshtein distance cannot be applied directly to all strings in errorful OCR of museum labels because of performance limitations and the false positive matches. However, OCR corrections are possible by evaluating context of the words before or after a potential fuzzy match. This can result in improved semantic labeling of neighboring items when using HMM. Substitution of numbers and alphanumeric stings with high variability can also improve semantic classification of items on labels at the cost of classification of some dates. Minor modifications of the substitution algorithm for recognition of standard date patterns should help overcome this limitation.

References

Frasconi, P., Soda, G., & Vullo, A. (2002). Hidden markov models for text categorization in multipage documents. *Journal of Intelligent Information Systems*, 18(2-3), 195-217.

Heidorn B., & Wei Q. Automatic Metadata Extraction from Museum Specimen Labels. (2008). In J. Greenberg and W. Klas (Eds.), *Proceedings of the International Conference on Dublin Core and Metadata Applications.* (pp. 57-68). Berlin, 22-26 September 2008 DC 2008: Berlin, Germany. http://hdl.handle.net/2142/9138

The International Plant Names Index (2012). Published on the Internet http://www.ipni.org [accessed 24 September 2012].

Pavord, A. (2005). *The naming of names: The search for order in the world of plants*. New York, NY: Bloomsbury.

Skog, J., McCourt, R.M., & Corman, J. (2009). The NSF Scientific Collections Survey: A Brief Overview of Findings. Retrieved from http://www.nsf.gov/pubs/2009/nsf09044/nsf09044.pdf

Thiers, B. (continuously updated). Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. Retrieved with permission from http://sweetgum.nybg.org/ih/