



Machine Learning to Produce Structured Records from Herbarium Label OCR'd Text

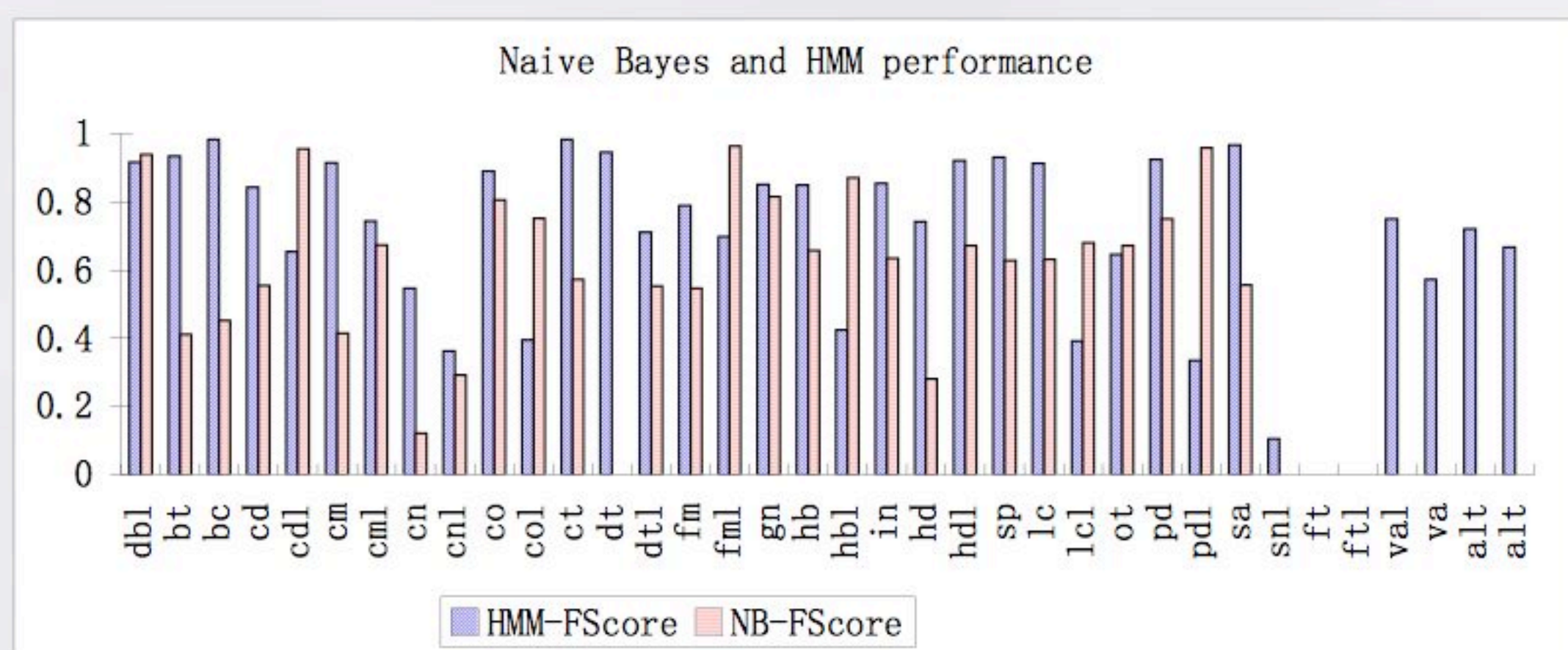
Qin W. Yin and P. Bryan Heidorn

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
501 E. Daniel St., Champaign, IL 61820

Does it work? --Ongoing Research

Self-evaluation and Comparison-evaluation module: Getting the evaluation information instantly is of importance to the users. It could tell the users whether the model they are using is the most suitable one for their data set. Users will get feedback instantly to see how they are doing in the training process and to understand the overall efficiency of the process. Graphs such as Evaluation Chart below would be helpful to users. In addition, individual institutions or projects may care more about different elements of the label such as scientific name and collector rather than overall performance so we will give them some flexibility in selecting the element sets for which they get feedback.

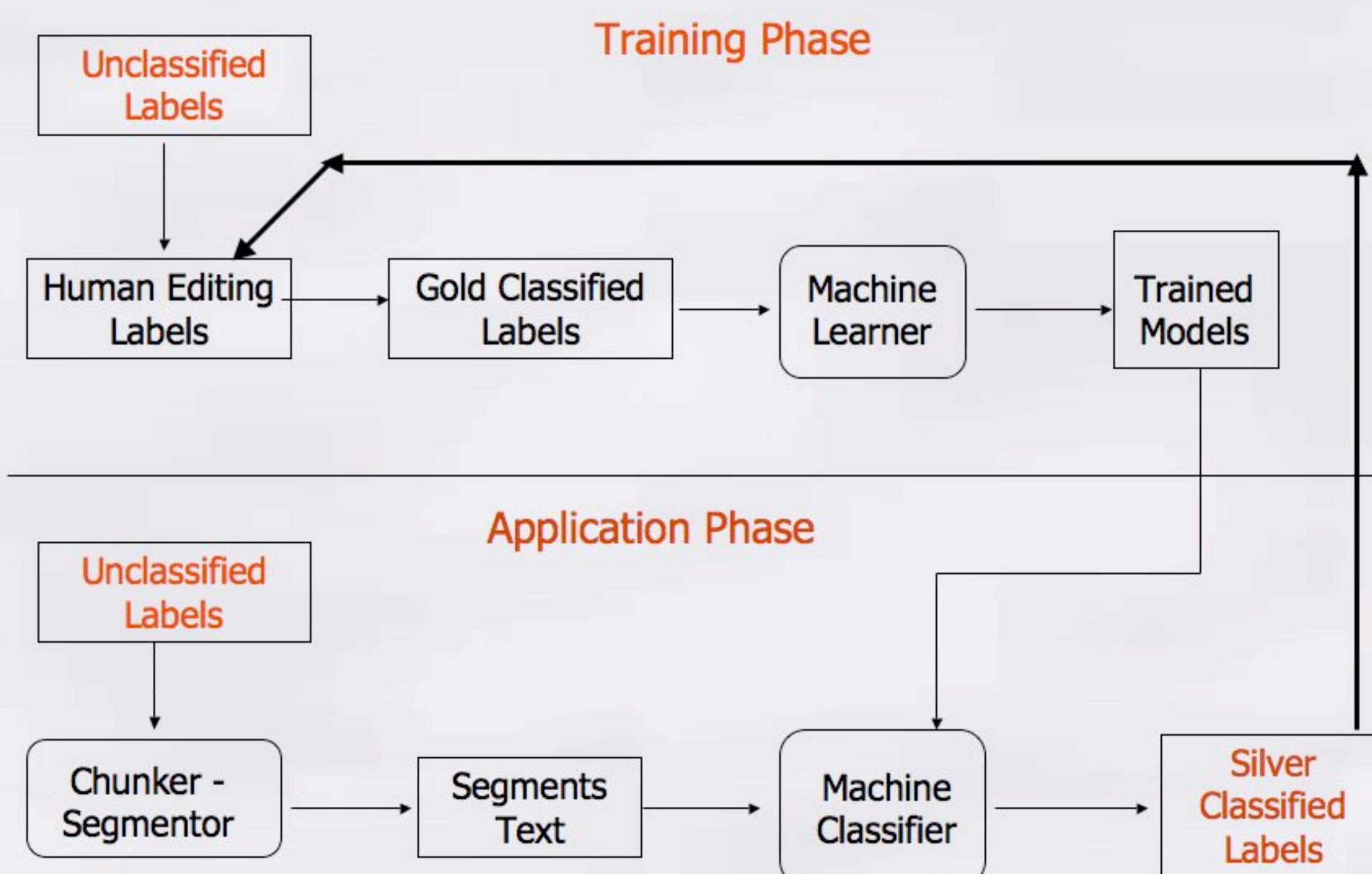
Ten-fold validation will be used as the evaluation method. The performance data would be only relied on the training set. We will conduct research on how the size of the training set influences performance. Testing the effects of the size of training set on the learning algorithms is both theoretically and practically important. If we could demonstrate that there is a "magic" number for training set size, we do not need more than the size of training set which would save the hand-code cost and time at the same time.



Evaluation Chart

Herbis is the Erudite Recorded Botanical Information Synthesizer (<http://www.herbis.org>)

Machine learning techniques are most valuable for information extraction in domains where there are "learnable" patterns in data but where the regularities in the data are varied enough to make hand programming of regular expressions cost prohibitive and sometimes impossible. Metadata extraction from museum specimen labels is such a domain. We are developing machine learning tools to automatically extract Dublin Core and other metadata, from museum specimen labels processed through Optical Character Recognition (OCR). This poster introduces our overall system architecture, identify challenges posed in natural language processing combining OCR and the application of Naïve Bayes and Hidden Markov Model (HMM).



Herbis Bootstrapping Learning Architecture

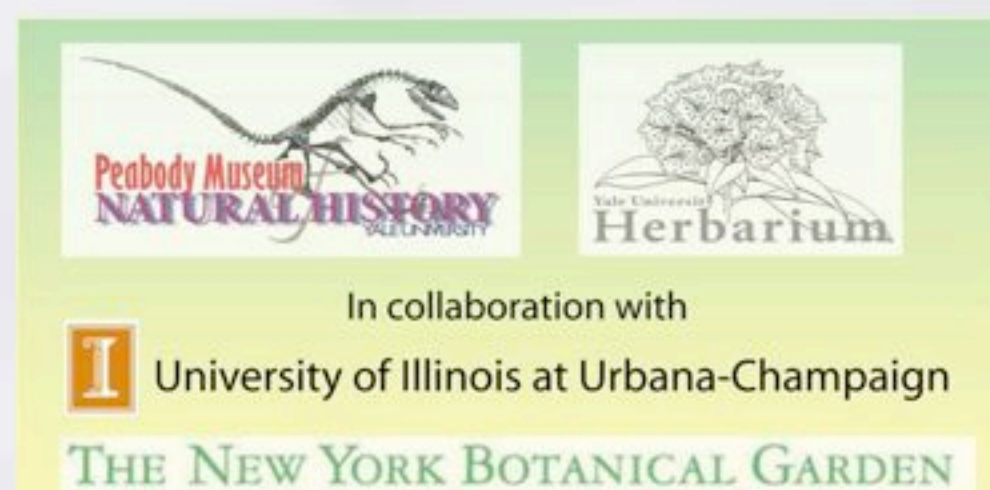
Future work

Bootstrapped Learning and Specialized Learning Models: Efficiency is required both in getting training dataset, training and processing. Specialized models allow individual projects to develop training sets and associated models that can produce more accurate translation of OCR records to XML than would be possible with generalized models. To improve performance we would implement the bootstrapping and specialization training mechanism that would allow users to create their own training sets and customized processing modules for different collections.

Multiple Personalized Schemas: While we have developed a detailed schema to extracting information from labels, in some cases users may wish to extract additional information. In order to improve the applicability of the NLP module and the whole system, the compatibility of multiple schemas would be an important enhancement. The current HERBIS data schema is designed for processing botanical specimen labels scanned and OCR'd texts. The schema includes 36 independent fields and most of them are Darwin Core.

User Editing Interface: A user friendly editing interface would be of important for the whole learning architecture since both the bootstrapping learning and evaluation module would need the input of the biology experts. We would integrate the GoldenGate (<http://idaho.ipd.uka.de/GoldenGATE/>) color-based editing environment. Most biology experts should not need to understand XML. They may not know how to use XML to markup the document and get confused by the tags. But a color based user editing interface would help them by representing the tag semantics in the color coding.

References:
Witten, I. H., and Frank, E. (2005). *Data mining: practical machine learning tools and techniques (2 ed.)*. Boston, MA: Morgan Kaufmann Publishers.
James R. Curran. Blueprint for a high performance NLP Infrastructure. In *Proceedings of the HLT-NAACL 2003 workshop on Software Engineering and Architecture of Language Technology Systems*. Volume 8. Pages: 39-44. 2003



Acknowledgments

NSF DBI Award #03456341