Interactive Machine Learning (IML) Markup of OCR Generated Text by Exploiting Domain Knowledge: A Biodiversity Case Study

Several digitization projects such as Google books are involved in scanning millions of books. The Biodiversity Heritage Digital Library (BHL http://www.bhl.si.edu/) plans to scan 1 million volumes of biodiversity literature over the next five years. However, the usefulness of the scanned images is limited because they can only be accessed through existing catalog information. Images can not be easily manipulated and transformed to useful information in full-text information systems. "Because of the very large amounts of data being generated, it is difficult to have human curators extract all these information and present them in a form useful to researchers. Information Extraction (IE) from such sources is becoming crucial for the timely dissemination of information." (Subramaniam, 2003). Consequently, simple approaches that transform the text to structured format such as XML or relational databases will not be successful.

Machine Learning (ML) techniques, especially Supervised ML (SML) have been used widely in information extraction (IE) and automatic markup. "ML has proven to be of great practical value… They are especially useful in (a) data mining problems where large databases may contain valuable implicit regularities that can be discovered automatically…" (Mitchell, 1997). IE and automatic markup of the biodiversity documents is this kind of domain. Substantial research has been conducted on the usefulness of ML in IE and automatic markup (e.g. Borker 2005; Cui 2005). Borker demonstrated 87% F-score in automatically extract address elements (eg. house number, street name, city) from addresses and bibliographic entries from bibliography resources. Cui's dissertation (2005) demonstrated that domain knowledge gained from machine learning models in one publication is very useful for improving the performance of automatic markup in another publication in the same field.

One of the least tapped sources of biodiversity knowledge is the collection locations, dates, species identification and other information on over a billion natural history specimen labels worldwide. Only a very small fraction of these have been digitized and the information added to databases (Beaman et al., 2006). The HERBIS (http://www.herbis.org) project has build tools to allow researchers to submit images of these specimens to a web service and receive an extended Darwin core document[1] in return. Using the Herbis Learning System (HLS), we extract 36 independent elements of information 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. The classification of the sub-elements is accomplished using SML. A training set was constructed using a collection of 145 examples which contains 4183 element classifications. The dataset comes from digitized OCR records from the Yale Peabody Herbarium with multiple label formats randomly selected from the type written labels and OCRed by ABBYY software. We coded the data as a Relax NG Schema allowing all elements occurrence to be

[1] See http://wiki.tdwg.org/twiki/bin/view/DarwinCore/DarwinCoreDraftStandard

optional, potentially occurring multiple times and in any order as is required by the variability in the input data. The relaxNG schema could be found online[2].

Many text classification ML algorithms are available such as: Naive Bayes, Hidden Markov Model, Decision Trees, Support Vector Machines. Each algorithm has its own advantages and shortcomings. The properties of our data helped us select particular algorithms. Museum labels have a relatively loose sequential structure, a high level of OCR errors, some fields have restricted sets of possible fillers while others are "open world" and may contain almost any text. For our tasks, a few of the fields are more important than others such as: family, genus, species, collector, and date. Several experiments need to be carried out to test several promising candidate algorithms and analysis their potential benefits and limitations of using them. For static evaluation, f-scores and ten-fold validation can be used. Because of the structure of the data we implemented a modified Hidden Markov Model and Naïve Bayes Model. A Hidden Markov Model (HMM) consists of states (in our case, they are the different kinds of elements), observations, start probability, transition probability and emission probability. Each state emits one or more symbols in the dictionary from a probability distribution for that state. Beginning from the start state, a HMM generates an output sequence by making transitions from one state to the next up to the end state. So the HMM model is an order preserving algorithm which is the primary feature of this model. It is currently widely used in web-content mining and speech recognition. A Naïve Bayes (NB) model is a probability model based on conditional probabilities. NB model make predictions based on the probability distribution of features from the training set. NB then uses the distribution information to calculate the probabilities of a new instance belong to the classes. The example would then be classified to the highest probability class. NB model has been proven good in some problems both in data mining and text mining. The performance of both models in our dataset could be found in figure 1.

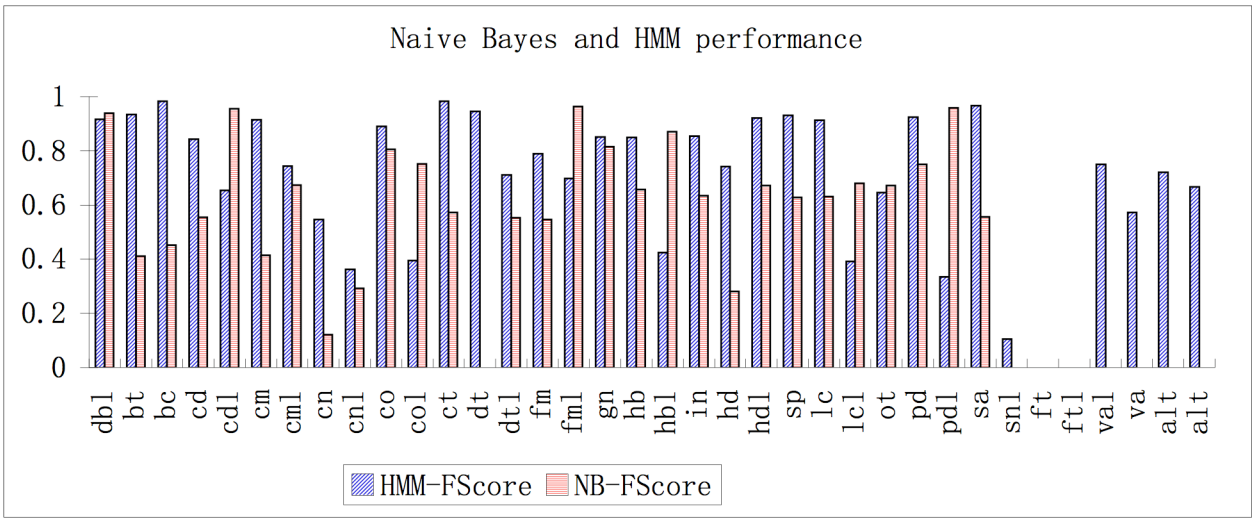

Fig. 1 Performance of NB and HMM

[2] http://www3.isrl.uiuc.edu/~TeleNature/Herbis/semanticrelax.rng

The results from the two unintegrated algorithms are encouraging. Generally, NB performs better than HMM on elements that are "labels" or markers for other labels. All these codes end with a letter "l". For example taxonomic family is coded as "fm" and family labels are "fml." Performance could be improved by integrating the algorithms, using the best algorithm for individual labels. MorphBank[3] and some other projects are coordinating with the HERBIS development teams to provide an automatic markup module for museum specimen digitization projects. We are expanding our research to use a more active architecture, Interactive Machine Learning (IML) as introduced by Ware, et al. 2002. Currently most ML systems are built by computer scientists (programmers) using expert generated data, not the domain experts. In the standard (non interactive) ML procedure, building a learner/classifier is a fully automated process. As Cui demonstrated most ML systems do not fully take advantage of the domain knowledge which could be very beneficial if used properly. IML "offers a natural way of integrating background knowledge into the modeling state. " (Ware, 2002).

Future Work:

System Design and Implementation: Unlike traditional ML, IML is a "human-in-the-loop" system. The system would be initialized with one of previously constructed models for one or more ML algorithms such as NB and HMM. A person using the system for the first time would feed raw museum label images through these models, which in turn would return the labels marked up in XML. Using a graphical user interface which represents the XML with more user friendly color coding, the user corrects any errors in the machine classification. The system can use these new label instances as a new training set to create new ML models tailored specifically to this users data. Given a sufficient number of examples, the performance of these new models should exceed the performance of the generic models that came with the system. With each batch of new records that the user submits, the system gives the user feedback on the relative performance of the available ML models. As the system performance increases the number of corrections that the user needs to make decreases. In essence each user can tailor a personalized IML system. By sharing the resulting models with other users, we have a social network of IML modules. The machine learning components of the system will be provided as a web service so that other people can build other interfaces over the IML web service modules.

[3] http://www.morphbank.net/

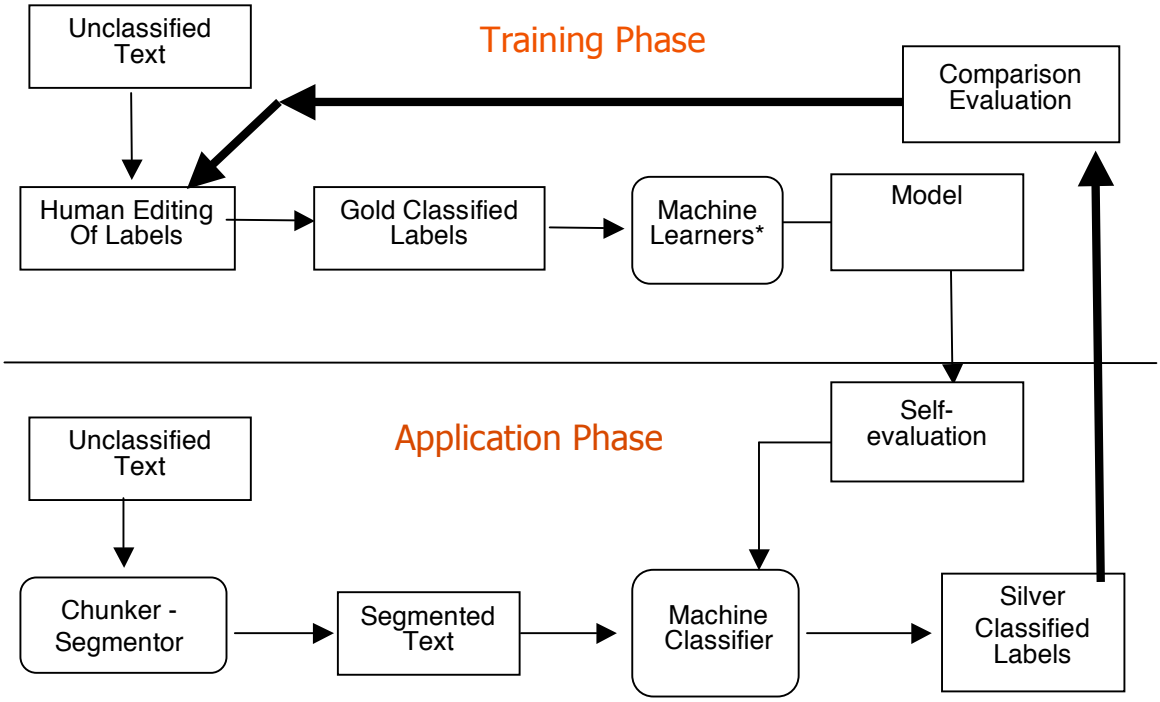

Fig 2. Interactive Machine Learning Architecture (*Machine Learners" in the diagram should be a stack of overlapping learners depending on which one the user selected.)

User centered experiments and data analysis. Standard precision, recall and F-Scores are not sufficient for evaluating interactive systems. While IML is fairly new, both Interactive Information Retrieval (IIR) and Interactive Data Mining (IDM) have been studied extensively. The importance of IIR could be found in TREC tasks. It appeared since the first TREC interactive query mode (TREC-1,2), interactive track (TREC 3-8), Manual query mode (TREC 1-7), and high precision (TREC-6,7). Our vision of the user-end experiments would be similar as the experiments done in IIR in TREC. The focus would be studying user's behavioral details, the process, and interim results as well as the summary of final results and the effects of the system, searcher and their interactions. Important variables are the number of human corrections required per some number of records, 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. But we would investigate several more measures that would be more suitable for Machine Learning and Automatic Markup. We will identify and discuss why we chose the measure and what's the advantages and limitation of each measure.

Reference:

David Robins. Interactive Information Retrieval: context and basic notions. *Information Science (Special issue on information science research)*, 3(2):57-61, 2002.

Hong Cui. Automating semantic markup of semi-structured text via an induced knowledge base: a case-study using floras. Ph.D Dissertation. University of Illinois at Urbana-Champaign. 2005.

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*. 8:39-44, 2003.

Jurgen Koeneman, Nicholas Belkin. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *Proceedings of CHI'1996*, 205-212, 1996.

L. Venkata Subramaniam. et al. Information extraction from biomedical literature: methodology, evaluation and an application. In *the Proceedings of the Twelfth International Conference on Information and Knowledge Management*. New Orleans, LA. Pages:410-417, 2003.

Malcolm Ware, et al. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*. 56(3): 281-292, 2002.

Reed S Beaman, at el., HERBIS: Integrating digital imaging and label data capture for herbaria. *Botany 2006*, Chico, CA. July 28-August 2, 2006.

Rupesh R. Mehta. et al. 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. Pages: 928-929, 2005.

Special issue on interactivity at the Text REtrieval Conference (TREC). *Information Processing and Management*, 37(3), 2001.

Tom M. Mitchell. Machine learning. McGraw Hill, 1997.

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