Name Matters: Taxonomic Name Recognition (TNR) in Biodiversity Heritage Library (BHL)

Qin Wei*
Graduate School of Library and Information Science
University of Illinois
501 E Daniel St
Champaign, IL, USA
qinwei2@illinois.edu

P. Bryan Heidorn
School of Information Resources and Library Science
University of Arizona
1515 East First St
Tucson AZ, USA
heidorn@email.arizona.edu

Chris Freeland
Missouri Botanical Garden
4311 Shaw Blvd
St. Louis, MO, USA
chris.freeland@mobot.org

ABSTRACT

Taxonomic Name Recognition is prerequisite for more advanced processing and mining of full-text taxonomic literatures. This paper investigates three issues of current TNR tools in detail: (1) The difficulties and methods used in TNRs. (2) The performance of Optical Character Recognition (OCR) and TNR tools by samples from Biodiversity Heritage Library (BHL). (3) The methods for potential improvement. We found that the performances of current TNR techniques need to be improved. A detailed error analysis reveals that sublanguage characteristics account for much of the error. A preliminary experiment using NaiveBayes (NB) models shows the potential of using machine learning (ML) in TNR.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Systems Issues, User Issues; I.2.7 [**Systems Issues, User Issues**]: Natural Language Processing

General Terms

Algorithms, Design, Performance, Experimentation, Languages

Keywords

Taxonomic Name Recognition, TNR, biodiversity informatics, Machine Learning, Digital Libraries, Information Retrieval

1. BACKGROUND

Digitization of library materials has become a global trend especially for biodiversity informatics such as the BHL (`http://www.biodiversitylibrary.org/`) project. BHL has been funded through a sub-award from the Encyclopedia of Life to digitize more than 60 million pages of legacy scientific literature within 5 years. 25,995,854 pages are available to date via the BHL Portal (updated 11/15/2009). An important aspect of BHL is the incorporation of "taxonomic intelligence" provided by the Universal Biological Indexer and Organizer (`uBio:http://www.ubio.org`) to automatically identify taxonomic names. The image files created by high-resolution scanners are processed through ABBY FineReader or PrimeReader (OCR softwares) to create text files. Those text files are then submitted to uBio's TaxonFinder web service to identify the candidate names. All candidate name strings are compared to NameBank, uBio's repository of about 10.7 million scientific names. When a match is made, the verified name string is then made available for search and display in the BHL portal.

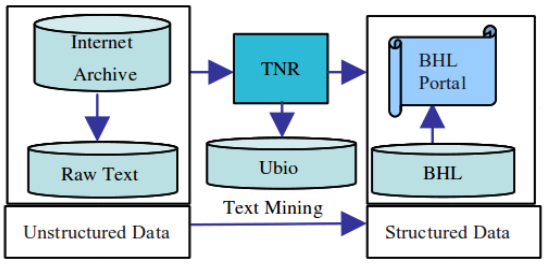

Figure 1: BHL architecture

Figure 1 presents current BHL architecture. The ultimate goal of the BHL is to build an intelligent user-driven digital library that provides the most authoritative information on all species and the means to navigate and analyze the information.

The paper is organized as follows. Section 2, introduces the methods used, and details why TNR is difficult. Section 3 presents the experimental design and characteristics of the BHL collection. Section 4 details the performances of OCR. Section 5, shows the performances of TNRs, an in-depth error analysis and methods for potential improvement. Section 6 presents the discussions and future work.

*Corresponding author.

Table 1: The methods used in TNR

Method	How	Example
Dictionary Lookup	Compare the target string to the strings in a dictionary	FAT [7]
Rule-Based	Using domain knowledge to construct rules	TaxonFinder [3]
Machine-Learning	Using corpus information to make decisions	MARTT [1] and Herbis [2]

2. INTRODUCTION

2.1 TNR Methods

Taxonomic Name Recognition could be regarded as a subtask of Named Entity Recognition (NER). TNR "has been developed to exploit the linguistic and contextual nature of taxonomic names, as dictated by Linnaean rules used for most organism scientific names since 1754" [6]. Many methods used in NER are adopted in TNR. The most common ones are dictionary lookup, grammars and rules matching, and machine learning, as summarized in table 1. Currently most real life applications are some combinations of the three methods while they might take advantage of one method more than the others. The two TNR tools evaluated here are TaxonFinder (`http://www.ubio.org/index.php?pagename=soap_methods/taxonFinder`) and FAT (Find All Taxonomic names. `http://idaho.ipd.uni-karlsruhe.de/GoldenGATE/`), both of which adopted the combination approach. TaxonFinder relies more on rules while FAT focuses more on dictionary lookup. The two are selected because they are the most widely used tools within the biodiversity community.

2.2 Difficulties of TNR

The main challenge of the TNR task and the requirement for an effective algorithm to perform this task depend on the type and degree of name variations in the collection to which the matching algorithm is applied. The name variations can be divided into three types: different kinds of literatures (language, genre, age), taxonomic naming variation and OCR errors. The variations in this research are listed as following:

1. BHL collection is a typical biodiversity collection that contains a huge volume of diverse literatures. The varieties include multi-language, a long time span (from 1500 to present), multi-discipline, multi-genre (journals and books) and so on.

2. Naming variation is ubiquitous in all kinds of taxonomic literatures. The variations increase the difficulties for any kind of automatic text processing. An informal analysis was conducted by the author and several main categories of naming variations were identified: with/without Species Author, genus abbreviation, Species author and genus abbreviation, invalid strings following correct name and the combinations of them. But there are also many exceptions, which are not listed here.

   (a) Variation because of author string: e.g. Cytisus supinus and Cytisus supinus Pimpinella, Cetraria aculeata and Cetraria aculeata (Ehrl.), Smelophyllum capense and Smelophyllum capense Rdlkf;

   (b) Variation because of Genus abbreviation: e.g. Amoora speciosa and A. speciosa;

   (c) Variation because of Genus abbreviation and author: e.g. Baeomyces intermedia and intermedia (Del.), Cladonia fimbriata and Cl. fimbriata Hffm (Note the Genus abbreviation has 2 letters), Cladonia pungens and Cl. pungens (Ach.), Durio carinatus and D. carinatus Mart;

   (d) Variation because of invalid characters following correct name: e.g. Orobus albus and Carduus mollis ("Orobus albus" is right, but "Carduus mollis" is not a valid author or name);

   (e) Variation because of c & d: e. g. Parmelia conspersa and P. conspersa Ach - Usque; "Parmelia conspersa" and "P. conspersa" shoud match. "Parmelia conspersa" and "P. conserpersa Ach." should match. However, "Parmelia conspersa" & "P. conserpersa Ach. - Usque" should not match. "-Usque" is not a valid name.

3. The most important factor is introduced by OCR errors. Since we are automatically transforming the image files into text files, errors are introduced at the same time. Although we are able to identify the most frequent patterns of OCR errors as presented in Section 4, generally speaking, the errors are unpredictable in a sense that the error patterns in real texts are irregular.

However, it is those difficulties in TNR make it different from other NER tasks and interesting to information researchers.

3. EXPERIMENTAL DESIGN

Three related questions are going to be answered by the following analysis: (1) Performances of OCR and TNRs in BHL. (2) Error analysis. (3) Candidate methods for improvements and the expected performances.

3.1 Experiment Procedures

For answering the first two analyses, first we need to construct the ground truth from the sample pages. We sent the pages to 14 volunteer biologists recruited at the beginning of the project along with a excel spreadsheet. The procedure began with their manually identification of all valid names in each page. The spreadsheet includes three columns: pageid (BHL unique identifier for a page), name_as_printed, names_as_OCRed. Name_as_printed and name_as_OCRed record the characters represent the names as printed and in OCRed text. The name_as_printed is served as the ground fact for the following discussions. The OCRed texts are used to evaluate the OCR performance. The names identified by TaxonFinder were retrieved from the BHL portal. The results include pageid and names identified. Software used for testing FAT is called GoldenGate (`http://idaho.ipd.uni-karlsruhe.de/GoldenGATE`). The version is 2008.03.25. 20.30. The results from FAT include the same fields: pageid and names identified.

Table 2: Characteristics of the sample

Number of Pages	392
Average Number of Words per Page	446.8
Average Number of Names per Page	7.7
Total Number of Names	3003

For answering the third question, a NaiveBayse (NB) classifier is implemented since NB is usually used as the baseline classifier in machine learning (ML). The toolkit used in this experiment is called WEKA version 3.4 (`http://www.cs.waikato.ac.nz/ml/weka/`). A 5-fold cross validation method is then used to evaluate the performance of NB.

Evaluation measures used in this study are standard Information Retrieval (IR) evaluation measures: precision, recall and F-score (detailed information about the measures could be found in Salton, 1971 [5]).

3.2 Sample Characteristics

We randomly selected 392 pages from the BHL database that contained 4,843,619 pages at the beginning of our project. Table 2 shows some characteristics of the sample. We denote a word to be any sequence of one or more letters that begin and end with a punctuation or space. We can see that the literatures are rich in names.

Meanwhile, we categorize the pages into three types: index pages, sublanguage pages and regular pages. Index pages are those pages that do not have grammars. Generally speaking, they include a list taxonomic names with/without page number. Sublanguage pages contain the most important taxonomic information. Here is an example: "Plants terrestrial, on rock, or rarely epiphytic. Stems erect or nearly erect, rarely long-creeping, scaly."[1] Sublanguage is different from natural language (or complete language, such as English or Chinese) from the perspective of vocabulary, grammar and more importantly, how it carries knowledge. In sublanguage, not only words but also grammars carry meanings. Regular pages are the pages include complete sentences that could be processed by regular NLP techniques.

These three different page types contain very different information. For index pages, taxonomic names appear intensively. Taxonomic descriptions are usually in Sublanguage pages while they contain fewer names than index pages but more than Regular pages. And those descriptions contain morphological information of species that are of importance to biologist. Regular pages may contain any kinds of information but fewer names.

Within our sample of 392 pages, 25 are index pages, 110 are sublanguage pages and 257 are regular pages.

4. OCR

4.1 OCR Performances

OCR, transforming the images to text files, is very important since the results of OCR are where the TNRs are going to be applied. Since the two TNR tools use morphological features to identify the name, we consider the OCR a failure if one or more letters of the generated word are wrong including those in wrong case. For example, one of the rules for matching names could be: Genus name is capitalized and is probably followed by lowercase species and subspecies names. Thus, if a capitalized word were not correctly recognized, the matching process would fail.

[1]Flora of North America online: `http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=10072`

Table 3: Overall OCR performance

Total	Wrong OCR	Error Rate
3003	1056	35.16%

Table 4: Frequent OCR error patterns in BHL

1	Insert Space	8	n->v
2	Omit Space	9	l->i
3	e->c	10	r->i
4	u->i	1	u->ii
5	u->n	12	h->l
6	i->l	13	h->ii
7	c->e	14	e->o

The table shows among the OCRed text of the 3003 valid names, 1056 of them contains at least one wrong character. It's worth mentioning that the performances of the OCR might be very different comparing to other types of text. Our evaluation collection is multi-language and older compared to other collections used in similar studies (e.g. [4] Rice, Kanai, and Narker, 1993). And the target for evaluation is limited to the name string that also makes a difference. We are also able to identify the top OCR error patterns in our sample as listed in table 4.

4.2 Performances On Different Languages

Our sample includes 242 English pages and 150 non-English pages. The precision for English and non-English pages are 64.78% and 64.04% respectively. A student t-test gets the p-value of 0.3333 which is not significant which means there is no significant different between the OCR performance on different languages. The result is reasonable since our target is only limited to taxonomic name strings. And name strings in taxonomic literatures tend to be Latinized in most circumstances where the language of the rest text in the page might be German, Italian or even Chinese.

4.3 Performances On Different Page Types

The OCR performances over different page types are 62.41%, 62.77% and 68.29% respectively for Index, Sublanguage, Regular pages. Several t-tests at 5% level show that the OCR performance on Regular pages is significantly better than the other two types of pages. But the performance difference is not significant between Index pages and Sublanguage pages.

5. TNR

5.1 TNR Performences

In digitization projects such as BHL the algorithms must also be able to find names even if they have OCR errors. So our evaluation included name strings that were identifiable

Table 5: Performances of TaxonFinder and FAT

With_OCR_Error	TaxonFinder	FAT
No. of Names (identified by biologist)	1696	1937
No. of Names Found by algorithms	1540	1603 %
Correct	621 %	452 %
Precision	40.32%	28.20%
Recall	36.62%	23.34%
F-score	38.47%	25.77%
With_OCR_Error	TaxonFinder	FAT
No. of Names (identified by biologist)	2610	3003
No. of Names Found by algorithms	1540	1603
Correct	674	517
Precision	43.77%	32.25%
Recall	25.82%	17.21%
F-score	34.80%	24.73%

Table 6: TNR performances on different page types

TaxonFinder	Precision	Recall	F-score
Index	32.82 %	17.89 %	25.36 %
Sublanguage	39.71 %	24.86 %	32.28 %
Regular	60.67 %	36.11 %	48.39 %
FAT	Precision	Recall	F-score
Index	58.01 %	20.94 %	39.47 %
Sublanguage	17.42 %	12.72 %	15.07 %
Regular	35.35 %	18.24 %	26.80 %

by humans even when they had OCR errors. Both TaxonFinder and FAT employ some forms of fuzzy matching that tried to addresses this problem. For example, Cardnus mollis is a valid name that could be found in page (`http://www.biodiversitylibrary.org/page/22001`). The OCR output Carduus mollis where 'n' was changed to 'u'. But TaxonFinder is able to correctly find Carduus mollis while the string is not confirmed by NameBank. Therefore, it is not shown in the portal. Here we present the performances of both algorithms under the situations with or without OCR errors.

Three t-tests at 5% level show that TaxonFinder is significantly better than FAT in precision, recall and F-score in both scenarios. But even TaxonFinder only achieved an F-score of 34.90%, which is still relatively low for an efficient retrieval. It means among the all names in the literatures, the TNR is only able to identify a quarter of them while leaving out the majority of the names. And among those found names, at least half of them are invalid names.

5.2 Performances on Different Page Types

Table 6 presents the performances of TaxonFinder and FAT in different page types. Both algorithms have significantly different performances over different page types while performing better in different types. TaxonFinder has its best performance in Regular pages and worst performance in Index pages. However, FAT has its best performance in Index pages and worst performance in Sublanguage pages.

Table 7: Page type classification confusion matrix

Confusion Matrix	Index	Sublanguage	Regular
Index	21	3	1
Sublanguage	1	81	28
Regular	1	22	234

5.3 Machine Learning (ML) Approach?

How to improve OCR softwares performance is not the focus of this research. Improving TNR algorithms effectiveness is the main focus. Compared to parsing by dictionary lookup and rules, ML has its own advantages that makes it much more suitable for TNR in OCRed text parsing. Here, we hypothesize that there are two approaches that might lead to the improved performance of TNR.

(1) We can see from table 6, a possibly better tool could combine the result of TaxonFinder and FAT by different page types (use TaxonFinder for Sublanguage and Regular pages, and FAT for Index pages) if page types could be efficiently and effectively identified.

(2) Also, we can see that sublanguage pages have the worst performance for FAT and modest performance for TaxonFinder. We propose that using machine learning to parse sublanguage pages would improve the performance.

5.3.1 Page Type Classification

By using the same sample data, a small-scale page type classification experiment was conducted in order to show the feasibility of combining the results from different page types. NaiveBayes (NB) is selected for this experiment since it is commonly used as the baseline model in various machine-learning tasks. The procedures of selecting features for NB model are explained as follows. Since we are dealing with multi-language pages, the features from a single language might not work well. Instead, we look into the generic features (e.g. morphological features) that exist in a broader range of languages. The features include 1-gram character, 2-gram characters, 3-gram characters, number of words per sentence, and number of sentences per page. We used a NaiveBayes classifier and evaluated it with 5-fold cross validation. The precision is 85.71% and the confusion matrix is shown in table 7.

The performance level we achieved is not trivial. We can see that automatic classification of page types is feasible and could achieve a high performance by combining the results. Despite the performance of the classifier, there are some important aspects are worth mentioning. First, as we can see from the confusion matrix, the main errors are coming from the confusion between sublanguage and regular pages. Part of reason is that some of the pages include both languages, e.g. (`http://biodiversitylibrary.org/page/2496490`) the last paragraph is sublanguage while the other paragraphs are regular language and similar situation in page (`http://biodiversitylibrary.org/page/3050492`). Second, the performance we gained here is the baseline performance that means the performance gained here is the lower bound of the performance we could get by using machine-learning methods. A better classifier could be gained from more carefully selected features and better

Table 8: NaiveBayes performance on text classification

Class	Precision	Recall	F-Score
Name-String	62.60%	20.60%	41.60%

classification models (e.g. Support Vector Machines) that will be our future work.

5.3.2 Sublanguage Pages

Improvement could also be achieved by improving the TNR performances on sublanguage pages. We could see from Table 6, the performances on sublanguage pages is significantly worse than other type of pages for both TaxonFinder and FAT. Supervised learning has been proposed to gain better performance on information extraction from sublanguage text [1]. A small-scale experiment conducted on name string classification using NaiveBayes also showed the potential of using machine learning in this task. The features used are similar with the features used in the page type classification, that include 1-gram character, 2-gram characters, 3-gram characters, Capitalized word or not. We also used a NaiveBayes classifier and evaluated the performance with 5-fold cross validation. The result we achieved is presented in table 8.

We can see the performance we get from an simple implementation of NaiveBayes is an F-score of 41.60% which is encouraging. Despite the performance of the classifier, it is also worth mentioning the following points. The training size on Name strings and non-names in this experiment is very skewed. The size of non-name strings is 20 times larger than name strings size. Skewed training data would lead to a lower performance of machine learning. One possible improvement would be boosting the training size by using Latin taxonomic name dictionaries or the names from NameBank which again will be our future work. Adapting a better classification model such as SVM and more carefully selected features have a great chance of improved performances.

6. DISCUSSION

The performances of OCR and TNRs are presented in section 4 and 5 and the error analysis leads to two proposed methods. We found the large gap between actual and potential performance of taxonomic recognition suggests a possibly fruitful avenue for the improvement of the taxonomic recognition quality. Given the availability of some start-of-the-art named entity recognition methods and the researches done on noisy information retrieval, it is possible to upgrade exiting methods which would substantially narrowing the gap. The characteristics of sublanguage pages and OCR errors made machine-learning methods very attractive. Two potential improvement methods are presented and evaluated in 5. More advanced techniques will be our future work.

7. REFERENCES

[1] H. Cui and P. B. Heidorn. The reusability of induced knowledge for the automatic semantic markup of taxonomic descriptions. *Journal of the American Society for Information Science and Technology*, 58(1):133–149, 2007.

[2] P. B. Heidorn and Q. Wei. Automatic metadata extraction from museum specimen labels. *Dublin Core and Metadata Applications*, 12(2):291–301, September 2008.

[3] D. Koning, N. Sarkar, and T. Moritz. Taxongrab: extracting taxonomic names from text. *Biodiversity Informatics*, (2):79–82, 2005.

[4] S. Rice, J. Kanai, and T. Nartker. An evaluation of ocr accuracy. Technical Report ISRI TR-93-01, University of Nevada, Las Vegas, April 1993.

[5] G. Salton. *The smart retrieval system: experiments in automatic document processing*. Prentice-Hall, 1971.

[6] N. Sarkar. Biodiversity informatics: organizing and linking information across the spectrum of life. *Briefings in Bioinformatics*, 8(5):347–357, 2007.

[7] G. Sautter, K. Böhm, and D. Agosti. A combining approach to find all taxon names (fat) in legacy biosisystematics literature. *Biodiversity informatics*, (3):41–53, 2006.