Knowledge extraction from archives of natural history collections

Stork, L.

CHAPTER 5

Automating Semantic Annotation

"Perhaps the deepest accomplishment of cognitive development is the construction of larger-scale systems of knowledge: [...] Building these systems takes years, much longer than learning a single new word or concept, but on this scale too the final product of learning far outstrips the data observed."

– Joshua B. Tenebaum, Charles Kemp, Thomas L. Griffiths, Noah D. Goodman, in: *How to Grow a Mind: Statistics, Structure, and Abstraction*

Biological nomenclature and systematics (discussed in Chapter 2), forms the basis of worldwide scientific discourse about the biodiversity of our planet. Employing such prior knowledge about biological structures in machine learning models, enables the process of learning to retrieve these structures accurately from only small data samples that encode them. At the same time, historical field observations, even more than contemporary ones, contain fierce discussions about systematics and nomenclature. Biological taxonomies, once extracted from archives, can be used to search historical records. Systems can exploit extracted taxonomies through query expansion techniques, or allow users to semantically query, or browse through, archival collections.

In this chapter, we aim to answer research question **Q.4**: *How can we use automated methods for knowledge extraction from archives of NHCs?* by aiming to automate part of the pipeline for knowledge extraction from *field books*.

5.1 Introduction

Automatic knowledge extraction from field book manuscripts and illustrations is challenging as content is too heterogeneous to process using common HTR techniques (**Chall.6**). HTR is one of the more challenging tasks within the field of Document Image Analysis and Recognition (DIAR), mainly due to the huge variety in writing styles and languages, paper degradation, overlapping words and historical handwriting. Creating labelled examples

for HTR requires domain expertise (**Chall.7**), and interesting words lie in the long tail of the distribution of words (**Chall.8**). Examples of interesting content that lies in this long tail, are scientific names. In Chapter 4, we saw that domain experts were interested in retrieving such names, see Table 4.1.

Here, we use computer vision and Semantic Web technologies to (i) identify the elements of scientific species names in handwritten document images, and (ii) link and structure the elements, using an ontology for species observations. We use the MONK handwriting recognition system (23) to segment the document images into single word images. Our main contribution is the automatic identification and semantic annotation of word zones in manuscripts that contain species names, and the goal is to integrate such a system with a system for HTR, together tackling the task of named entity transcription and salient named entity recognition and classification (SNERC).

We build on work described in the previous chapter (Chapter 4), where an ontology and software for semantic annotation of species observation records was constructed and tested with domain experts. Here, we advance these methods by automating the process of semantic annotation. We present a a novel approach to identify *scientific names* in historical *handwritten* document images. Rather than first transcribing the text and performing NERC afterwards on the digital text, we exploit characteristics of the document images for identifying the domain specific salient named entities, using terms from the NHC-Ontology[1] to classify and organise them. We argue that the ability to quickly index handwritten document images based on scientific names, ranks and authors, helps users to navigate through large collections of documents in online libraries, such as the Biodiversity Heritage Library (BHL).[2] It opens up possibilities for faceted search, semantic querying and semantic recommendations. Additionally, maintaining a link to the word image and location in the full document image is important to generate ground truth for repetition of image processing experiments as well as to allow researchers to view the original document and therefore the extracted text in context.

5.2 Related Work

Organisations and researchers that dedicate themselves to the preservation of natural history collections, such as *IdigBio*[3] or the BHL (13), continuously develop new methods to digitise specimen collections in a cost-effective and sustainable way, in order to facilitate ongoing species research.

[1] http://www.makingsense.liacs.nl/rdf/nhc/,https://github.com/lisestork/nhc-ontology/
[2] https://www.biodiversitylibrary.org/
[3] https://www.idigbio.org/

The automatic extraction of scientific names from text is essential for the management of archival resources. Therefore, there are several examples of methods for extracting and disambiguating species names from printed texts, but extracting the same information from handwritten texts is much more of a challenge. TaxonGrab (120), for example, automatically extracts species names from printed biological texts. The BHL, which aggregates scans of biodiversity publications and field notes, indexes scientific names extracted from the publications—printed text—in their collection, to improve accessibility for taxonomists. They match the text, extracted via OCR, with the Taxonomic Name Server (TNS) to identify likely scientific names (13).

Similarly to the BHL, other researchers and institutes are exploiting the power of automatic text processing for the digitisation of natural history collections. Software has been developed to parse OCR output of printed text to formalised DwC entries for archival and retrieval purposes (121). Drinkwater et al. (20) investigate the aid of OCR in the digitisation of herbarium specimen labels, and found a significant increase in time effectiveness using OCR output to (i) sort specimens prior to database submission, and (ii) to add transform labels to minimal database records. Drinkwater et al. explicitly note that OCR is currently only possible for typed and printed labels and not for handwritten text.

As HTR is one of the more challenging tasks within the field of DIAR, mainly due to the huge variety in writing styles and languages, paper degradation, overlapping words and historical handwriting (**Chall.6**). The recognition of named entities can help document understanding and searchability of the text, and can potentially aid HTR (86). Formerly, NERC was a task solely used on digital text, but it has recently also been applied directly to handwritten text (85; 84; 86). Especially when few instances of words exist and a collection consists of many different hands and connected words, making it difficult to create character-based representations, the identification of key words can help make the text searchable, and potentially aid HTR. Moreover, in many cases, full-text transcriptions of entire pages of field books are not required in order to make them digitally accessible.

5.3 Data

Transcribed field books exist online, but (to the best of our knowledge) no segmented and annotated images of handwritten species observations are available online. For this purpose, word images from 240 field notes from a natural history collection have been segmented and semantically annotated. The process of annotation has been carried out in the context of this work. However, the process of segmenting digital images into word

zones has been carried out by the MONK system for the project *Making Sense of Illustrated Handwritten Archives*[1] (19), and this is reflected in Figure 5.1.

From a field book on mammals, we selected field notes from four different writers, to account for different handwriting styles and structures, ensuring a representative dataset to demonstrate how the automated methods perform on heterogeneous, real-world data. The segmented word images were obtained from a nichesourcing effort, with the help of a handwriting recognition system MONK and a group of domain expert labellers. The word images were subsequently manually annotated using four classes, as shown in Table 5.1. Two of four classes are taxonomic entities. The third class refers to the publisher of the taxonomic name, and lastly we have the class *Other*, which includes all words that do not belong to any of the previously mentioned classes.

Table 5.1: Dataset class count

class	Genus	Species	Author	Other	Total
y	0	1	2	3	
n	177	167	144	17309	17797

The final counts of examples per class are shown in Table 5.1. The process of labelling and annotating words is time-consuming and, in our case, requires expert knowledge. Therefore, limited training data is available. As machine learning methods generally require a very large number of labelled samples, methods have to be adjusted to the dataset size to acquire a predictive model that generalises well. These adjustments are described in Section 5.4 and 5.5. This is also one of the challenges of projects working with real-world data where obtaining labelled data is expensive or simply not feasible. Models that use prior knowledge are better able to generalise from noisy data and small samples. The dataset used in this work can be found online.[2]

5.4 Scientific Name Extraction Model

Below we describe our contribution. The full pipeline is shown in Figure 5.1, the blue rectangle indicating the scope of this work.

We used the MONK handwriting recognition system (23; 26), developed by Schomaker, for word segmentation (122; 123; 124). First, the system segments handwritten document images into lines and second, relative to those lines, into word zones that potentially hold words. The system allows the labelling of word images and transcription of sentences by

[1] http://www.makingsenseproject.org
[2] 10.5281/zenodo.2545573

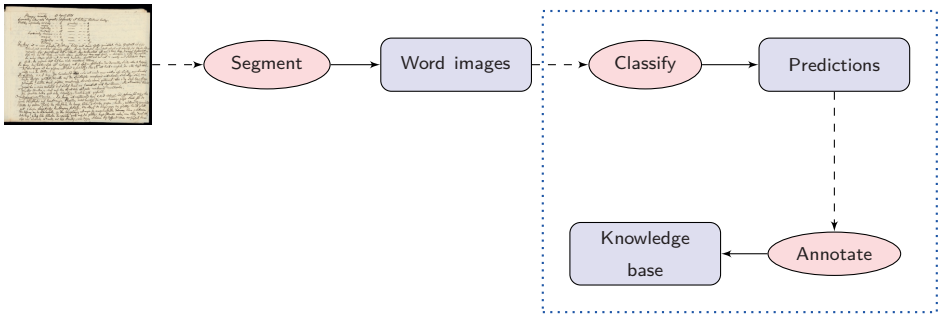


Figure 5.1: The full pipeline: automated semantic annotation of scientific names

domain experts. It then uses these labels for HTR. In this work, the word images were manually annotated using four semantic concepts, or classes: `genus`, `species`, `author` and `other`. The classification of each word image to its corresponding semantic class is discussed in Subsection 5.4.1. In Subsection 5.4.2, we discuss the semantic annotation of the classified word images using the NHC-Ontology[1] for species observations.

5.4.1 Classification of Word Images

To classify the word images to one of four classes, we use three distinct features; *visual structural features, position and context*. We chose to create one single neural architecture, built with help of Keras (125), that could be trained end-to-end, so that the classification error is only propagated once, in contrast to using predictions from multiple classifiers and combining them after training to form a single prediction. The final architecture is explained visually in Figure 5.2, and will be discussed below.

Visual Structural Features. The feature detector that was used in this work for the detection of visual structural features is a CNN (126). It has been shown that CNNs outperform other ANNs on image recognition tasks (127), see Section 2.2.1. The basic network used here is a deep CNN for object recognition developed and trained by Oxford's Visual Geometry Group (VGG) and called the VGG network (127). We use their configuration, with 16 convolutional layers, and import weights pre-trained on the ImageNet task by the VGG (128). Previous work (129) has demonstrated that transferring image representations with CNNs overcomes the problem of training with limited training data, e.g., less than a few thousand training images, despite differences in image statistics between the *source* dataset and *target* dataset. By, for instance, training on the ImageNet task, the VGG model learns filters on various different scales, which can be used as feature extractors for

[1] `http://www.makingsense.liacs.nl/rdf/nhc/`,`https://github.com/lisestork/nhc-ontology/`

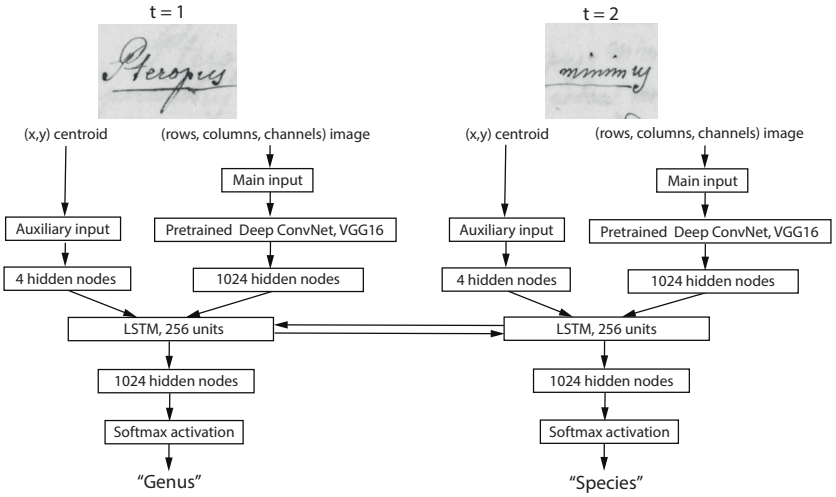


Figure 5.2: The CNN–MLP–BLSTM architecture, "unrolled" for both time steps t.

other types of images. These features, extracted from handwritten documents with help of the convolutional part of the VGG network, are used for training a simple MLP on our task.

Position. In addition to visual features, the position of a word in a document, especially (semi)-structured ones such as field observation records, often provides a good descriptive feature for the recognition of a named entity. The position is therefore often used as a feature in the field of NERC, however, it has been used more often in digital text, e.g., (130) than in digital images, e.g., (85; 84; 86; 83). In this work, we use the relative centroid of the word images, $c = (x, y)$, relative to the image borders, as input features to a simple MLP with two inputs, x and y, and one hidden layer of size 4. To train the entire model end-to-end, we concatenated the last hidden layers of both models. The merged hidden layer therefore has a size of $1024 + 4 = 1028$.

Context. As a third feature type, we introduce context: the characteristics of adjacent word images, specifically *bi-grams*. Figure 5.3 shows frequencies for word image bi-grams. First, horizontal pairwise alignment was calculated per word $w^{(i)}$ and $w^{(j)}$. They were seen as horizontally aligned if $y1^{(i)} < yc^{(j)} < y2^{(i)}$, where i and j indicate the i-th and j-th word image, $y1^{(i)}$ the first y coordinate of $w^{(i)}$, $y2^{(i)}$ the second, and $yc^{(j)}$ the y coordinate of the centroid of $w^{(j)}$. Second, the right neighbouring word of $w^{(i)}$ was retrieved by calculating all pairwise vertical distances for the horizontally aligned words: $dist_{ij} = cx^i - cx^j$, where cx^i refers to the x coordinate of the centroid of $w^{(i)}$.

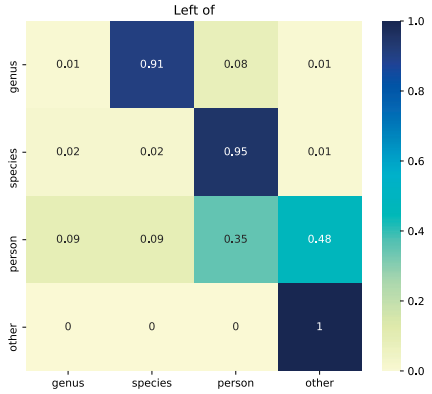


Figure 5.3: Adjacency matrix that shows frequencies for word bi-grams (sequences of two adjacent words). E.g., 'genus' was **left of** 'species' 91% of the time 'genus' was encountered.

The smallest negative distance, within a certain bound, indicated right adjacency. The adjacency matrix only takes into account instances that actually have an adjacent word, as it could be that a word is surrounded by white space on every side.

As expected, the different classes have strong co-occurrence dependencies. Therefore, we converted the dataset to sequences of size two (bi-grams), and added a last layer to the model architecture for sequence prediction. For an adequate prediction we used a BLSTM neural network (discussed in Subsection 2.2.1) that is able to learn long-term dependencies between features. By using the bidirectional variant of the LSTM, dependencies can be learned in both horizontal orientations, see Figure 5.2.

5.4.2 Semantic Annotation of Word Images

The NHC-Ontology[1] is an ontology for species observations, based on the DSW ontology, and written in OWL.[2] The ontology is centered around the description of meta-data relating to the observation of an organism, and allows a researcher to describe as which various taxon groups an organism has been identified. The model uses the Web Annotation Vocabulary[3] to link bounding boxes of word images to their semantic labels. In the example listing below, Listing 5.1, two images refer to a genus and a species, which together constitute one taxonomic name `ex:taxon1` of rank `ex:species`. They are linked to the publisher of the name with the *nhc:scientificNameAuthorship* property.

[1] `http://www.makingsense.liacs.nl/rdf/nhc/`,`https://github.com/lisestork/nhc-ontology/`
[2] `https://www.w3.org/OWL/`
[3] `https://www.w3.org/TR/annotation-vocab/`

```
@prefix nhc: <http://makingsense.liacs.nl/rdf/nhc/> .
@prefix ex: <http://example.org/terms/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix dwc: <http://rs.tdwg.org/dwc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

ex:taxon1 rdf:type dwc:Taxon ;
          nhc:scientificNameAuthorship ex:author1 ;
          nhc:taxonRank ex:species .

ex:author1 rdf:type foaf:Person .

ex:anno1 rdf:type oa:Annotation ;
         oa:hasBody ex:taxon1 ;
         oa:hasTarget ex:image1.jpg#xywh=x,y,h,w ;
         oa:hasTarget ex:image1.jpg#xywh=x,y,h,w .

ex:anno2 rdf:type oa:Annotation ;
         oa:hasBody ex:author1 ;
         oa:hasTarget ex:image1.jpg#xywh=x,y,h,w .
```

Listing 5.1: Example of a semantically annotated species name

5.5 Experiments and Results

To analyse the influence of the three features on the predictive performance of the model, we conducted multiple experiments where we tested the performance of the pre-trained CNN, CNN–MLP and CNN–MLP–BLSTM.

5.5.1 Experimental Methodology

Before training, the images were scaled by dividing them by 255 so that they would fall within the range [0-1]. All images were re-sized to the average image dimensions: $y = 74$, $x = 139$. No data augmentation was used. Based on horizontal adjacency, as explained in Subsection 5.4.1, image bi-grams were constructed, sequences of $l = 2$, as input to the BLSTM.

The word images were shuffled, keeping together word images from the same page, and thereafter split into a train and test set. As one word image could occur in two bi-grams, we hereby avoid that word images from the test set were also in the training set, which would bias the classification results. However, by shuffling the pages, we still ensure that the model does not overfit to one writing style or structure. We used 80% of the word images for training and the remaining partition as test set, making sure that 20% of the scientific name elements were in the test set. As classes in the word bi-grams were

highly imbalanced, we used random minority oversampling with replacement, to increase the counts of samples from minority classes in the training data. When training a CNN, oversampling is thought to be the best method to deal with imbalanced datasets with few examples in minority classes, and appears to work best if the oversampling totally eliminates the imbalance (52).

However, as we are dealing with sequences rather than singular samples, we chose to oversample sequences, e.g., *species–author*. Converted back to singular images, this would result in a *step imbalance* with a small imbalance ratio $p = \pm 1.1$ rather than a large imbalance ratio of $p = \pm 16$ (52).

The networks were all trained using the Adam classifier with a learning rate of 10^{-4} and categorical cross-entropy loss. Each network was trained using early stopping with patience 2, meaning that training was stopped when, for two epochs, the validation error was increasing. Per epoch, the weights were only stored if the predictive performance had increased compared to the previous epoch. In the testing phase, thresholding was applied to the output of the networks to compensate for oversampling the data during training, as oversampling alters prior probability distributions. One way to perform thresholding is to simply correct for these prior probabilities, by dividing the output of the network for each class, then seen as posterior probabilities, by the estimated prior probabilities. In our case, the imbalance was not completely eliminated, so the thresholds were calculated as the ratio between the original class counts and those after oversampling.

As a final step, the output of the model that performed best was used to test the whole pipeline. Word images from the test set, that were classified as scientific names, were assigned IRI e.g., `ex:taxon1`. The names were linked and semantically enriched using terms from the ontology and transformed to the RDF format. The code can be found online.[1]

5.5.2 Results and Discussion

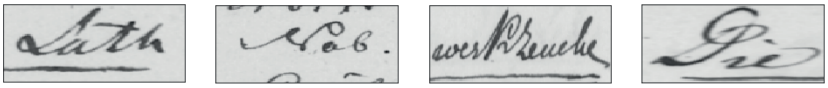
Table 5.2 summarises the final classification results for each network. Due to a large class imbalance, precision and recall were used to assess the predictive power of the classifier. Reporting accuracies would be misleading, as they would portray the underlying distribution, rather than the predictive power of the model (if the model would always predict "*other*", it would be a bad predictor for the task, but the accuracy would be 93%, as the "*other*" class accounts for 93% of the data).

[1] `https://github.com/lisestork/asa-species-names`

Table 5.2: Classification precision, recall and $F1$ results for each network. Support indicates the number of actual occurrences of that class in the given subset.

Method	Class	Precision	Recall	F1-score	Support
1. CNN	Genus	0.80	0.78	0.79	36
	Species	0.64	0.97	0.77	33
	Author	0.78	0.78	0.78	32
	Other	1.00	0.97	0.98	525
	avg / total	0.82	0.77	0.80	626
2. CNN–MLP	Genus	0.85	0.81	**0.83**	36
	Species	0.81	0.88	**0.84**	33
	Author	0.78	0.78	0.78	32
	Other	0.99	0.99	0.99	525
	avg / total	0.96	0.96	0.96	626
3. CNN–MLP–BLSTM	Genus	0.86	0.89	**0.88**	36
	Species	0.94	0.91	**0.92**	33
	Author	0.78	0.88	**0.82**	32
	Other	1.00	0.99	0.99	525
	avg / total	0.98	0.97	0.98	626

Bold $F1$ scores indicate statistical superiority over $F1$ scores for that same class within the cell of the preceding method. The table indicates that the BLSTM produced the highest average $F1$ scores for each class. The addition of the BLSTM layer specifically increases precision and recall scores for the author names. This makes sense; without context these appear similar to regular words. The input of centroid data to the network does not have an effect on the recall or precision of author names, but does increase precision for the retrieval of species names. Figure 5.4 shows 4 images from the test set that were misclassified. While both the CNN and CNN–MLP network misclassify most of the same word images, the output of the CNN–MLP–BLSTM is quite different. Image **(a)** and **(b)** were both misclassified by the networks without the BLSTM layer, but were correctly classified by the final model. Image **(a)** for example, was classified as "*species*", while actually being labelled as an author name. Visually, it resembles a species name; it is underlined and appears in a similar position on the page. Without context of other words it is challenging to correctly classify such images without proper historical knowledge of the domain. Image **(b)** was misclassified as "*other*", but correctly identified as an author name in the BLSTM model, most likely due to the visual characteristics of the word image that is left adjacent. On the other hand, image **(c)** and **(d)** are together misclassified as a species name and its author by the BLSTM network, while they were correctly classified by the other networks. Examining the images, we see that they are adjacent and visually resemble these classes (capitals, underlining).



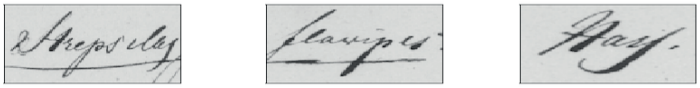
(a) $y = 2$, $\hat{y} = 1$ **(b)** $y = 2$, $\hat{y} = 0$ **(c)** $y = 3$, $\hat{y} = 1$ **(d)** $y = 3$, $\hat{y} = 2$

Figure 5.4: Four misclassified examples. Classlabels relate to those discussed in table 5.1

In Table 5.3, we present retrieval scores for the identification of complete scientific names from field book pages. A python script parsed the recognised species elements from the test set, and connected them together using the NHC-Ontology. A total of 27 out of 36 species names were retrieved, with an *F1* score of 0.86. Interestingly, there were no false-positives among the final predictions. Figure 5.5 shows one of the correctly classified scientific names. The final RDF dataset can be queried through our online SPARQL endpoint.[1]

Table 5.3: Final classification precision, recall and $F1$ results for the detection of scientific names.

Method	Class	Precision	Recall	F1-score	Support	Total
CNN–MLP–BLSTM	Scientific names	1.0	0.75	0.86	27	36



(a) $y = 0$, $\hat{y} = 0$ **(b)** $y = 1$, $\hat{y} = 1$ **(c)** $y = 2$, $\hat{y} = 2$

Figure 5.5: A correctly classified scientific species name: **(a)** *Genus* **(b)** *Species* **(c)** *Person*

5.6 Conclusions and Future Work

In this chapter we show that we can accurately identify and classify components of handwritten species observation records from different features: visual structural features, position and context. We show that our methods are applicable even though the dataset contains four authors with different handwriting styles and different processes of recording their species observations. A major challenge of working with handwritten text is its irregularity. Our results show that we can mitigate this challenge by building up multiple pieces of evidence for classification by learning from multiple features. Each of the different

[1] `http://makingsense.liacs.nl/rdf4j-server/repositories/SN`, can be queried through a query editor such as: `https://yasgui.org/`

features we examine in our model adds information and improves the overall results. In addition, as the results are extracted and structured in RDF format as part of the process, they are immediately available for search and comparison with other archives - historical or present day.

The dataset used for experiments in this chapter is part of the same expedition archive (the NC collection, see Subsection 2.3.2). Although we represent multiple authors and styles, the next step would be to demonstrate the generic nature of our results by analysing biodiversity records from other expeditions. Once we establish that, we will extend our methods to identify other common classes from biodiversity data, for example, locations, dates and anatomical entities.

It is our aim to integrate the new methods with established methods for automated handwriting recognition, using a fruitful dialogue between our system and a system for HTR, in which the hypotheses (highest confidence values) of both systems work together for the transcription and semantic annotation of named entities in manuscripts.