Priming Digitisation: Learning the Structure in Historical Texts

Lise Stork[1,2], Fons Verbeek[1], Aske Plaat[1], Jaap van den Herik[2], and Katherine Wolstencroft[1]

[1]Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
[2]The Leiden Centre for Data Science, Leiden, The Netherlands

Keywords: Named Entities, Object Recognition, Neural Networks, Conditional Random Fields, Biodiversity Research

The number of specimens stored globally in natural history museums is estimated to be between 1.2 and 2.1 billion. Of these historical records, a mere 3% is currently digitally available [1]. A considerable amount of work is required to digitise the archival content, such as field books, diaries and specimen labels. The historical context is not easily reconciled with current knowledge and paleographers are required to decipher the historical handwriting. Hence, most time is spent geo-referencing dubious place names, resolving amgibuous names and processing, entering, and maintaining records in databases [10, 6].

Some researchers address the task by first fully transcribing field books [7, 1, 2], after which Natural Language Processing (NLP) methods are applied to structure the data. We argue that full transcription is not necessary to make the field books accessible for research, and that a more targeted approach for the transformation of field books into structured databases is required. To this purpose, we propose to complement a handwriting recognition system with intuition concerning the structure of the text. Our implementation is twofold: (1) we use a top down semantic model that formalises the semantics of field notes - the named entities and links between them, in combination with (2) bottom up image-based statistical methods that automatically extract the named entities. The latter is based on spatial context: we implement a recogniser that adaptively learns probabilistic templates. High probabilities are assigned to areas in the field-note images where named entities are likely to occur.

Our method provides two benefits over full text transcription: (1) a handwriting recogniser can be focussed to words that elucidate on the informative content of a field note such as a taxonomical name or locality, and (2) by obtaining knowledge about the semantic class of a word, a handwriting recognition system is biased towards word candidates that are semantically more likely than others. By priming the digitisation process we equip natural history museums with a new, more focussed approach to facilitate and speed up the conversion of their collections to queriable databases.

Our semantic model,i.e., an ontology for field books, is constructed from domain expertise and existing community standards[8]. The ontology enables us to bias the recognition process and structure its output, making it machine readable and aggregatable. An example of named entities and links as defined in our ontology are the elements of a taxonomical name. In binomial nomenclature, a taxonomical name is composed of two parts, the *genus* and the *specific epithet*. These two parts are then followed by a person name, being the publisher of the taxonomical name.

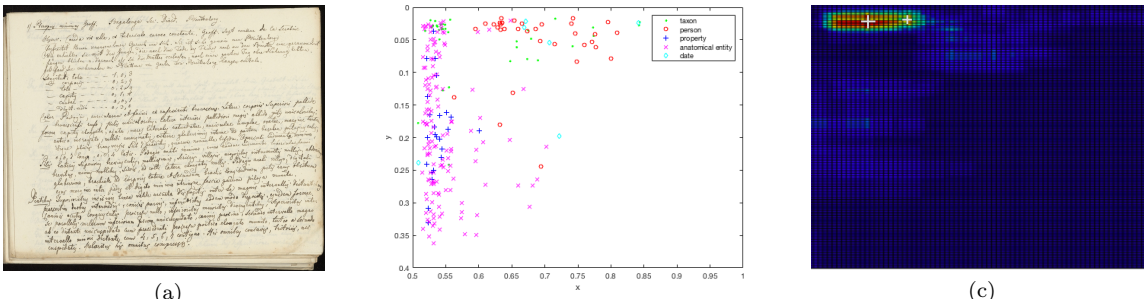
With the ontology we manually annotated a field book from our use case[1], harvesting 389 annotations and 9500 triples[2]. The annotation process provided us with relevant data necessary to study the automatic recognition of named entities. Simultaneously, it facilitated the elucidation of the content of one field book which is now queriable by users via a SPARQL endpoint[3]. Figure 1 shows a plot for the centroids -i.e., the x and y coordinates of the centre - of the annotated named entities. It is notable from figure 1 that properties and anatomical entities are usually located on the left side of the page, parallel to the y-axis, while the taxonomical names, locations and dates are located parallel to the x-axis at the top of page.

From figure 1 it is clear that a spatial arrangement exists in our data. The use of spatial context for object recognition has proven to increase the robustness of a recognition process identifying objects in a scene [4, 5, 3], as strong spatial relations between objects increase the recognition of individual objects. It is particularly useful when visual information cannot provide sufficient discriminative information for an accurate separation of classes.

[1] A collection which includes specimens and archival material of all expeditions undertaken by the Committee for Natural History of the Netherlands Indies, recorded in Indonesia between 1820 and 1850. The collection contains roughly 10,000 specimens, and about 8,000 field book pages and is physically stored at Naturalis Biodiversity Centre (NBC). More information can be found on: http://makingsenseproject.org.

[2]Besides the textual and semantic label, annotation provenance data were stored during the annotation process.

[3]http://makingsense.liacs.nl/rdf4j-server/repositories/NC



(a) (b) (c)

Figure 1: a) one example of a field note from our use case, b) the centroids of all annotated named entities, and c) a heatmap of the location of the taxonomical names (left cross) and person names (right cross).

In our study this can be illustrated by the recognition of a word as a taxonomical name. The words *Pteropus* and *Gymnonotus* are both instances of a taxonomical name, with taxon rank *genus*, although visually they do not share many features. In the field books, a genus is underlined and often has the -us suffix. However, its spatial context can provide more relevant information to recognise both names as belonging to the class *taxon* and rank *genus*. Both appear in the top left corner of the page and are followed by an epithet and person name. Hence, as a proper spatial arrangement exists, we can use spatial context in combination with intrinsic visual features to perform named entity recognition.

So, our dataset provides us with a collection of field book pages $P_1, ...P_n$ with labelled word zones[4] $W_1, ...W_t$. A Convolutional Neural Network (CNN) is trained to learn a posterior probability distribution per word zone. The distribution assigns a probability $P(e_i|x_1, x_2, ..., x_n)$ to each named entity $e_i \in E$ from the ontology, based on the observed pixel values $x_i \in X$. Subsequently, a Conditional Random Field (CRF) is used to calculate the conditional distribution over the whole sequence of all class predictions: $P(e_1, e_2, ..., e_t|W_1, W_2, ...W_t)$. The posterior distribution is dependent upon the learned spatial context given the class predictions, i.e., (i) the location of a named entity on the page, represented by its centroid c_i in relative coordinates x and y, (ii) the co-occurrence of entities, represented by a frequency count histogram per page P_i, and (iii) their pairwise relative locations on the page L_{ij}, $i \neq j$. Calculating the posterior distribution over the whole sequence enables the calculation of a configuration with the highest probability.

The method described has been developed as part of the *Making Sense project*[1], and will be applied in cooperation with the handwriting recognition system MONK [9]. The location of a word is recognised by such a system and before possible word candidates are calculated for each word zone, the probability of it belonging to a specific class - or no class - is calculated and the most likely configuration is adopted. Using our method, we are able to bootstrap the field book digitisation process, in an effort to make more of the historical biodiversity data available to researchers and the general public.

References

[1] Ariño, A. H. (2010). Approaches to estimating the universe of natural history collections data. Biodiversity Informatics, 7(2). ISO 690

[2] Bloom, D., Thomer, A., Vaidya, G., Guralnick, R., Russell, L.: From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks. ZooKeys. 209, 235-253 (2012).

[3] Cao, Y., Wang, C., Li, Z., Zhang, L., & Zhang, L.: Spatial-bag-of-features. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 3352-3359). IEEE. (2010)

[4] Galleguillos, C., Rabinovich, A., & Belongie, S.: Object categorization using co-occurrence, location and appearance. In Computer Vision and Pattern Recognition, CVPR. (pp. 1-8). (2008)

[5] Heitz, G., & Koller, D.: Learning spatial context: Using stuff to find things. Computer VisionECCV 2008, 30-43.(2008)

[6] Lister, A.: Natural history collections as sources of long-term datasets. Trends in Ecology & Evolution. 26(4), 153-154 (2011)

[7] Nakasone, S., Sheffield, C.: Descriptive Metadata for Field Books: Methods and Practices of the Field Book Project. D-Lib Magazine. 19(11/12), 1 (2013).

[8] Stork, L., Weber, A. Gassó Miracle, E., Verbeek, F., PLaat, A., van den Herik, J. & Wolstencroft, K. Semantic Annotation of Natural History Collections. *In review*

[4] A named entity labelled by a bounding box

[9] Schomaker,L.:Design considerations for a large-scale image-based text search engine in historical manuscript collections. Information Technology, 58(2), 80-88. (2016).

[10] Vollmar, A., Macklin, J. A., & Ford, L. (2010). Natural history specimen digitization: challenges and concerns. Biodiversity informatics, 7(2).