BIOfid Dataset: Publishing a German Gold Standard for Named Entity Recognition in Historical Biodiversity Literature

Sajawel Ahmed[1], Manuel Stoeckel[1], Christine Driller[2], Adrian Pachzelt[3], Alexander Mehler[1]
[1]Goethe University Frankfurt
[2]Senckenberg Nature Research Society
[3]Frankfurt University Library
{sahmed,mehler}@em.uni-frankfurt.de

Abstract

The *Specialized Information Service Biodiversity Research* (*BIOfid*) has been launched to mobilize valuable biological data from printed literature hidden in German libraries for over the past 250 years. In this project, we annotate German texts converted by OCR from historical scientific literature on the biodiversity of plants, birds, moths and butterflies. Our work enables the automatic extraction of biological information previously buried in the mass of papers and volumes. For this purpose, we generated training data for the tasks of *Named Entity Recognition* (NER) and *Taxa Recognition* (TR) in biological documents. We use this data to train a number of leading machine learning tools and create a gold standard for TR in biodiversity literature. More specifically, we perform a practical analysis of our newly generated *BIOfid dataset* through various downstream-task evaluations and establish a new state of the art for TR with 80.23% F-score. In this sense, our paper lays the foundations for future work in the field of information extraction in biology texts.

1 Introduction

Data is the gold to any machine learning (ML). Most ML approaches to *Natural Language Processing* (NLP) address modern, high-resource languages (such as English or Chinese) rather than historical, low-resource languages. As a consequence, feasible ML-tools for processing historical documents are still rare. In this paper we consider corpora of historical German texts in order to extract useful information about biological systems in the past (e.g. species, biotopes etc.).

As a contribution to closing the gap between NLP of modern and of historical languages, we present the newly annotated *BIOfid dataset* for *Named Entity Recognition* (NER) and for *Taxa Recognition* (TR) in the domain of biology, the first of its kind concerning the German language. Our approach is especially designed to address the exploration of biodiversity data[1] from historical documents. We perform a large-scale annotation of scanned texts converted by OCR from historical scientific books on the biodiversity of plants, birds, moths and butterflies, thereby creating the necessary training data to accomplish the task of biological NER and TR using various ML algorithms. Our work facilitates an automatic extraction of biological information so far buried in the bulk of papers and volumes (see Table 1). Overall, our newly generated dataset provides a gold standard and hereby lays the foundations for future work, such as relation extraction and classification based on extracted biological named entities and taxa.

Input sentence:
Ahmed observes that Iris grows in Mai in Frankfurt.
TR output:
*Ahmed observes that [Iris]*TAXON *grows in Mai in Frankfurt.*
Biological NER output:
*[Ahmed]*PER *observes that [Iris]*TAXON *grows in [Mai]*TIME *in [Frankfurt]*LOC*.*

Table 1: Example for our selected tasks.

We perform a practical analysis of our dataset via various downstream-task evaluations. First, we generate a baseline for recognizing taxonomic entities by constructing a sequence tagger based on skip-n-grams and external knowledge resources (i.e. WikiData). Secondly, we apply the best publicly available word embeddings for German and use them alongside our BIOfid dataset as an input for training high-performing neural mod-

[1] Biodiversity is the science which measures the variability and diversity of animals and plants.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 871–880
Hong Kong, China, November 3-4, 2019. ©2019 Association for Computational Linguistics

els for NER, namely BiLSTM, ELMo, Flair and BERT (Ahmed and Mehler, 2018; Peters et al., 2018; Akbik et al., 2018; Devlin et al., 2018). By using the optimized BiLSTM model we achieve a new best F-score of 80.23% regarding the recognition of taxonomic entities.

The remainder of the paper is organized as follows: Section 2 reviews related work. Section 3 describes the source texts and the preprocessing pipeline. Section 4 describes the annotation guidelines, process and environment for producing the BIOfid dataset, and methods (n-gram-based sequence tagger, neural models) for evaluating the practical quality of our annotated dataset. Section 5 presents the experimental results. Finally, Section 6 draws a conclusion.

2 Related Work

2018 was a vital year for the task of German NER, following a saturation period from when the last major progress was made by Lample et al. (2016). With the grammar-specific morphological processing and resource-optimization presented by Ahmed and Mehler (2018), the gap between English and German NER was closed. In the same year, with the emergence of multilingual language models such as *ELMo*, *Flair* and *BERT* (Peters et al., 2018; Akbik et al., 2018; Devlin et al., 2018), the performance of various NLP tasks, including NER, was notably improved. Hence, the task of German NER has benefited from these developments.

However, with respect to the availability of a variety of resources, there has not been much progress made until now. Regarding the standard task of NER based on four categories (PERSON, LOCATION, ORGANIZATION, OTHER), the first choice of resources for German is still the *GermEval* dataset (Benikova et al., 2014), followed by the datasets of *CoNLL* and *TüBa-D/Z* (Tjong Kim Sang and De Meulder, 2003; Telljohann et al., 2012). However, their potential for purposes outside of theoretical ML is limited. These datasets do not contain any annotations for taxonomic and temporal entities which are of key interest for biodiversity researchers.

For biological NER in the German language, there are no predecessor resources available to the knowledge of the authors; only an English counterpart exists, namely the *Copious* dataset (T.H. Nguyen et al., 2019), which has been recently published during our ongoing work. This confirms our research endeavors and shows the necessity of more data in this field. We take the English counterpart as the baseline and compare its dataset and results with our own. Overall, our work constitutes the first effort on enabling a state-of-the-art performance for neural representation learning to biological NER.

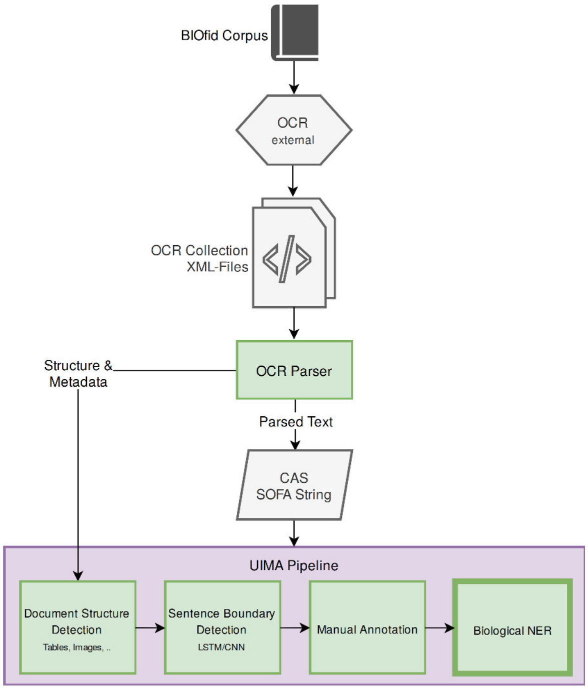


Figure 1: Flowchart showing the data cleaning steps within our preprocessing pipeline.

3 Source Texts & Preprocessing Pipeline

BIOfid Corpus The *BIOfid Corpus* is a collection of historical scientific books on central European biodiversity. It was assembled by a group of German domain experts, denoting a potential pool of relevant print-only journals and publications for historical biodiversity science. However, mainly due to license issues, not all publications could be considered for the corpus.

The available publications were scanned by an external service and subsequently paginated with the software *Visual Library*. Subsequently, every high-resolution page (400 dpi) was digitized with *ABBYY FineReader 8.0 (2005)* to ABBYY-XML, which includes structural information like paragraphs, bold/italic text, images, and table blocks.

OCR Parser The raw OCR data contained various errors, e.g. delivering typical OCR errors such as confusing letters (**ß** → **b**), or delivering gibberish due to the wrong recognition of non-textual elements in scans such as images, figures, or tables. Furthermore, species names or their appended author citation were frequently recognized incorrectly, e.g. ”Lepidium ruderale L.” → ”Lepidium rüderale I.”.

We built the following preprocessing pipeline (see Figure 1) to clean the source data and increase its overall quality. First, the raw OCR data was passed to a parser (labeled ”OCR Parser” in Figure 1). This parser read a given ABBYY-XML into a UIMA CAS, while retaining all structural information in a custom UIMA type system, which was tailored to the ABBYY-XML output.

Using a set of heuristics, the structural information was used to detect erroneous parts in the parsed text, such as page numbers, image and figure blocks mislabeled as text, text margins and table lines parsed as the characters ”I” or ”-”, and tables containing merely non-word characters such as counts of observations[2].

The parser performed further fundamental text segmentation using the information given by the ABBYY-XML, such as tokenization and paragraph splitting. The ABBYY-XML contains tokenization information on the character basis, denoting whether a character is marking the beginning of a word. This information was used alongside plain whitespaces to tokenize the raw text, while further splitting words from non-word characters. All this information was stored in a UIMA CAS using the aforementioned type system and passed down the UIMA pipeline.

Document Structure The BIOfid corpus comprises about 15 journal titles including approximately 410 books. 201 of these books containing 969 articles were selected by domain experts as a representative sample from the entire corpus to generate training data for biological NER.

Sentence Boundary Detection In biological literature, author citations are commonly abbreviated (e.g. Carl von Linné in ”Fagus sylvatica L.”) as well as species names (e.g. ”F. sylvatica” after the first definition). Therefore, standard rule-based tools often fail to detect the correct sentence boundaries in such unstructured raw text documents. Hence, for this task we included the LSTM-based sentence boundary detector *DeepEOS* (Schweter and Ahmed, 2019) in our processing pipeline and trained it with 1,361 sentences, which were manually extracted from the BIOfid corpus. The total amount of training sentences was increased from a preliminary size of 300, since the first experimental results revealed that the SBD is crucial for the performance of our downstream-task.

4 BIOfid Dataset & Methods

4.1 Annotation Guidelines

Named Entities NEs are real-world objects in a given natural language text which denote a unique individual with a *proper name* (e.g. Frankfurt, Africa, Linnaeus, BHL). This stands in contrast to the class of *common names* which refer to some kind of entities (e.g. city, continent, person, corporation) and *not* a uniquely identifiable object.

The standard task of NER focuses on the former class of proper names. However, it is often not easy to differentiate between both classes. Hence, to support the annotators in making the right decision, we created guidelines which demonstrated the rules for annotations. We gradually developed this document in collaboration with the annotators, until finalizing it as the guidelines for annotating the BIOfid corpus. The appendix shows the material which was provided to the team of annotators. First, in Appendix A some introductory examples from the BIOfid corpus are given. Next, in Appendix B the general guidelines used for producing the NER dataset are shown.

As we essentially extend the standard task of NER to our scope of biodiversity, our guidelines are built upon those used for producing the GermEval dataset (Benikova et al., 2014). For this, we take the original German text and extend it with the important adjustments described in the next paragraphs for the context of biodiversity. In contrast to Benikova et al. (2014), we do not consider derivative or partial NEs as a separate category. As the recent work of Ahmed an Mehler (2018) has shown, discarding subtle details is even beneficial, whereas fine-graded feature engineering for deep neural networks usually deteriorates the final performance.

Time In the standard task of NER, temporal information is not captured by the four base entities. However, the aspect of time is important for the research on biodiversity which is constantly evolving. Therefore, we annotated every text unit

[2] An example of such pages is given in Appendix C.

Dataset	Sentence	PERSON	LOCATION	ORGANIZATION	OTHER	TIME	TAXON
CoNLL	18,933	*5,369*	6,579	4,441	3,968	N/A	N/A
GermEval	31,300	10,807	17,275	8,303	4,557	N/A	N/A
TüBa-D/Z	104,787	55,746	28,582	32,224	12,865	N/A	N/A
Copious	*26,277*	*2,889*	*9,921*	*N/A*	*N/A*	*2,210*	*12,227*
BIOfid	**15,833**	**5,393**	**6,785**	**1,085**	**7,849**	**5,197**	**15,085**

Table 2: Statistics for German NER datasets together with the English biological NER dataset *Copious* (T.H. Nguyen et al., 2019).

which denotes a specific temporal entity with the tag TIME (e.g. *[13.02.1835]*TIME, see more in Appendix B: Table 9). For text units which describe a time interval, we marked the starting and ending points as two distinct temporal entities.

Taxonomy Taxonomy is a field in biology that deals with the systematic classification of organisms by morphological, phenotypic, behavioral and phylogenetic characteristics. Based on a variety of common traits, a group of organisms forms a so-called taxon. A well-known example of this are the Darwin's finches, endemic birds in the Galápagos Islands. The different species (each species represents a taxon) are distinguished primarily by the size and shape of their beaks and the associated specialized diets.

Taxa are classified according to international nomenclature codes[3,4,5,6] and are delineated at different hierarchical levels, also known as taxonomic ranks. Most of us are well acquainted with the distinction between the animal and plant kingdoms, although there are other kingdoms e.g. fungi or bacteria. Subordinate to a kingdom are many more ranks such as phylum, class, order, family, genus and species. According to this, the hierarchical classification of the bird species *Struthio camelus*, the common ostrich, from the lowest to the highest taxonomic rank is as follows: *Struthio camelus* (species), *Struthio* (genus), *Struthionidae* (family), *Struthioniformes* (order), *Aves* (class), *Chordata* (phylum), *Animalia* (kingdom). Each scientific name mentioned here along with its taxonomic rank (in parentheses) represents a taxon, meaning a group of organisms with a set of common characteristics being indicative for a common ancestry.

Due to differing and evolving methods of classification, taxonomies are subject to constant change. This also applies to taxonomic nomenclature. Therefore, among others, synonymy and homonymy also play an important role in biology (e.g. there is a plant genus with the name ”Paris”). The relevance of taxonomy for biodiversity research and conservation is fundamental (Thomson et al., 2018), consequently, we considered it justified to introduce the NE-category of TAXON into the process of NER.

For organisms of all taxonomic ranks, we considered scientific names (both accepted and synonyms) and vernacular names, if referring to a certain taxon, as NEs (e.g. *[Struthio camelus]*TAXON or [common ostrich]TAXON, *[Mirza zaza]*TAXON or [northern giant mouse lemur]TAXON, see more in Appendix B: Table 7). Author citation and year, usually appended to the scientific name of a taxon, were tagged as NEs of the categories PERSON and TIME, respectively (e.g. *[Falco]TAXON [Linnaeus]PER [1758]TIME*). Both author and temporal information embedded within the scientific name, were included in the NE TAXON (e.g. *[Carex praecox [Jacq.]PER var. distans]TAXON [Appel]PER*).

4.2 Annotation Process

We performed a single major series of annotations. Instead of just focusing on some inter-agreement value, we performed double checks on existing annotations on given articles through biological experts. This strategy removed the time overload associated to multi-annotations while ensuring a high quality of data.

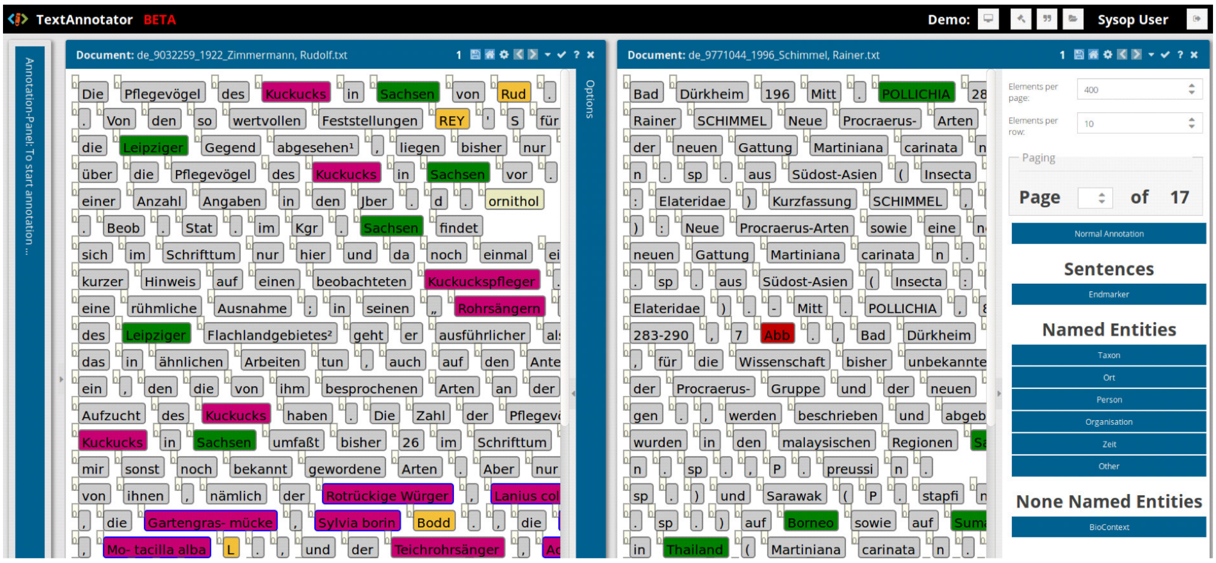
For this scheme, a group of annotators consisting of two researchers from the project team were employed. Both researchers were native speakers of German, and, additionally had a profound background in biology. Besides, two further student assistants with similar profiles were employed to provide further assistance.

[3] http://iczn.org/code

[4] http://www.iapt-taxon.org/nomen/main.php

[5] http://www.the-icsp.org/

[6] http://talk.ictvonline.org/taxonomy/



Die Pflegevögel des Kuckucks in Sachsen von Rud .
. Von den so wertvollen Feststellungen REY ' S für
die Leipziger Gegend abgesehen¹ , liegen bisher nur
über die Pflegevögel des Kuckucks in Sachsen vor .
einer Anzahl Angaben in den Jber . d . ornithol
. Beob . Stat . im Kgr . Sachsen findet
sich im Schrifttum nur hier und da noch einmal ei
kurzer Hinweis auf einen beobachteten Kuckuckspfleger .
eine rühmliche Ausnahme ; in seinen „ Rohrsängern
des Leipziger Flachlandgebietes² geht er ausführlicher al
das in ähnlichen Arbeiten tun , auch auf den Ante
ein , den die von ihm besprochenen Arten an der
Aufzucht des Kuckucks haben . Die Zahl der Pflegev
Kuckucks in Sachsen umfaßt bisher 26 im Schrifttum
mir sonst noch bekannt gewordene Arten . Aber nur
von ihnen , nämlich der Rotrückige Würger , Lanius col
, die Gartengras- mücke , Sylvia borin Bodd . , die
, Mo- tacilla alba L . , und der Teichrohrsänger , Ac

Bad Dürkheim 196 Mitt . POLLICHIA 28
Rainer SCHIMMEL Neue Procraerus- Arten
der neuen Gattung Martiniana carinata n
n . sp . aus Südost-Asien (Insecta
: Elateridae) Kurzfassung SCHIMMEL ,
) : Neue Procraerus-Arten sowie eine n
neuen Gattung Martiniana carinata n .
. sp . aus Südost-Asien (Insecta :
Elateridae) . - Mitt . POLLICHIA , 8
283-290 , 7 Abb . , Bad Dürkheim
, für die Wissenschaft bisher unbekannte
der Procraerus- Gruppe und der neuen
gen . , werden beschrieben und abgeb
wurden in den malaysischen Regionen Sa
n . sp . , P . preussi n .
sp .) und Sarawak (P . stapfi n
. sp .) auf Borneo sowie auf Suma
in Thailand (Martiniana carinata n .

Figure 2: Working environment for annotating the BIOfid corpus (figure taken from (Abrami et al., 2019)).

4.3 Annotation Environment

We used the *TextAnnotator* (Abrami et al., 2019), a browser-based annotation tool specifically adjusted for this project. Figure 2 shows the working environment which was provided to the annotators. On the left-hand side of the *QuickAnnotator* view, the raw OCR text from the BIOfid corpus is displayed, separated from the choice of annotation tags on the right-hand side. As sentence splitting was part of the annotation task, we did not provide a sentence view. Instead, we provided the whole article, further allowing the annotators to use contextual information while making their decisions.

4.4 Quality of Data

4.4.1 Quantitative Characteristics

Table 2 shows the total amount of annotated sentences along their six NE-categories and compares this with the three major public datasets for German NER. For our BIOfid dataset, we can see the high value of `TIME` and `TAXON` entities which, so far, do not exist for any publicly available dataset.

4.4.2 Data Format

We use the 4-column CoNLL-format which writes each word of a sentence horizontally along its lemma, POS tag and gold label, separating each sentence by an empty new line. For the tagging scheme, we opt for BIO (IOB2). Listing 1 shows an excerpt of the train file in which the entities `TIME`, `PERSON`, `LOCATION`, `TAXON` are marked by our team of annotators for a given sentence from the BIOfid corpus.

Listing 1: Sample sentence from BIOfid dataset

```
Mein          mein          PPOSAT    O
Sohn          Sohn          NN        O
konnte        können        VMFIN     O
am            an            APPRART   O
3             3             CARD      B–TME
.             —             $.        I–TME
1             1             CARD      I–TME
.             —             $.        I–TME
23            23            CARD      I–TME
den           der           ART       O
Fabrikanten   Fabrikant     NN        O
Walter        Walter        NE        B–PER
Schmidt       Schmidt       NE        I–PER
aus           aus           APPR      O
Geithain      Geithain      NE        B–LOC
bei           bei           APPR      O
einem         ein           ART       O
Spaziergang   Spaziergang   NN        O
auf           auf           APPR      O
dem           der           ART       O
Rochlitzer    Rochlitzer    NN        B–LOC
Berge         Berg          NN        I–LOC
auf           auf           APPR      O
eine          ein           ART       O
Ringamsel     Ringamsel     NN        B–TAX
,             —             $,        O
Turdus        Turdus        NN        B–TAX
torquatus     torquatus     ADJD      I–TAX
L             L             NN        B–PER
.             —             $.        O
,             —             $,        O
hinweisen     hinweisen     VVINF     O
.             $.            —         O
```

We split the BIOfid dataset into train, dev, test files by the common ratio of 80:10:10 percentages after randomizing its order of sentences. These final data files are utilized for training and evaluating our models, which are described in the next section.

4.5 Methods

For the evaluation of the BIOfid dataset, we use six different approaches and compare each others results: one classic *rule-based* model and five high-performing *embedding-based* models.

4.5.1 N-Gram Tagger for TR

We develop a naive sequence tagger as a baseline for the recognition of taxonomic entities in the BIOfid dataset. The baseline is only for a subtask of the full task of biological NER, described in the previous Section 4.1. Our sequence tagger is built on the k-skip-n-grams (with $k = 1$) which are constructed from the tokens of taxonomic entries in the comprehensive *Latin* and *German* gazetteers of biology. Both gazetteers consist of 83,348 taxonomic entries from various biological systematics such as of *aves*, *lepidoptera* and *vascular plant*. In addition, we consider *WikiData*[7] and construct an additional gazetteer by extracting 2,663,995 German and Latin taxonomic entries from the online resource by selecting all entries from a XML-dump that are subjects (`?s`) in the following two SPARQL triple patterns[8]:

- `?s instance-of taxon.`
- `?o subclass-of taxon.`
  `?s instance-of ?o.`

For each gazetteer entry consisting of at least three tokens ($n \geq 3$), we take all tokens as an input and create a list of 1-skip-n-grams. For example, for the taxonomic entry *iris kashmiriana b.*, we create four n-grams *(iris kashmiriana), (iris b.), (kashmiriana b.)* and *(iris kashmiriana b.)*. In this way, we construct 3,023,270 unique n-grams in total from 2,682,959 merged taxonomic entries, while dropping 140,432 duplicate n-grams entirely. Next, we map all these n-grams to the BIOfid test file by standard string matching and thus find the taxonomic occurrences in the target set of text data.

4.5.2 Neural Models for NER

Our neural models consist of two separately trained components: a) foundational word embeddings, modeling the general knowledge from large unlabeled text corpora, and b) various task-specific neural architectures, modeling the domain knowledge from the labeled training data. In this section, both components are presented briefly.

Word Embeddings The language model of continuous space word representations (*word2vec*) (Mikolov et al., 2013) and its variations by (Levy and Goldberg, 2014; Komninos and Manandhar, 2016) are the foundations of most ongoing research in NLP with neural networks. Based on the context, the model embeds words, phrases or sentences into high dimensional vector spaces. We use the model of *Wang2vec* (Ling et al., 2015) and its morphological extension (Ahmed and Mehler, 2018) which explores syntactic data specific for German and, thus, better suites the task of NER. We use the recently published German language word embeddings from the TTLab[9] which are pre-trained with the morphological extension of the Wang2vec algorithm on the COW corpus (Schäfer, 2015), the largest collection of German texts extracted from web documents with over 617 Mio. sentences. Out of the six published variants of embeddings, we opt for token-based embeddings (*COW.lower.wang2vec*), as they delivered the best results for German NER according to the publishers.

BiLSTM We provide a brief overview of the configurations for the five neural models which we use throughout this paper. The model *BiLSTM-CRF* is similar to the one used in (Ahmed and Mehler, 2018), which goes back to the work of (Lample et al., 2016). The neural network consists of stacked LSTM and CRF layers. The *base layer* combines for a given word its (pre-trained) word embedding with its character-based embedding. These features are forwarded to the *prediction layer* which produces the final NE tag.

Model	Emb.	Language Model	Train Data
BiLSTM-a	COW	N/A	BIOfid
Flair Wang2v.	COW	PCE	BIOfid
Flair ELMo	COW	PCE+Leipzig	BIOfid
Flair BERT	COW	PCE+BERT-Base	BIOfid
BiLSTM-b	COW	N/A	All

Table 3: Overview of the model inputs. For BiLSTM-b we consider all merged training data (i.e. BIOfid + GermEval + CoNLL)

Flair Wang2vec We further train a sequence labeling model using Flair[10]. We build the model in

[7] http://www.wikidata.org/

[8] All results can be acquired with the following WikiData queries: http://w.wiki/3u3 and http://w.wiki/3ud

[9] http://www.texttechnologylab.org/resources2018/

[10] http://github.com/zalandoresearch/flair

the same fashion as used by (Akbik et al., 2018) following the guide given by the authors for the task ”CoNLL-03 Named Entity Recognition (German)”, while keeping the pooled contextualized embeddings (PCE) and exchanging the GloVe embeddings employed by the authors with Wang2vec embeddings trained on the COW corpus.

Flair ELMo In addition to the previous model, we train a Flair Sequence Tagging model by stacking an ELMo embedding layer on top of the Flair Wang2vec model. The ELMo embeddings were trained on a section of the *Leipzig Corpora Collection* (Goldhahn et al., 2012) containing 100,000 sentences from Wikipedia using default parameters.

Flair BERT Similarly, we added BERT (Devlin et al., 2018) to the Flair Wang2vec model. We used the recently published *BERT-Base, Multilingual Cased*[11] pre-trained model for this purpose.

Hyperparameters We take the original neural models and keep the hyperparameters as described in their references. The only adjustments we make to the models are on the input level, i.e. we perform variations for the pre-trained word embeddings, the pre-trained language models, and the training data (see Table 3).

5 Results

We evaluate the performance of all models with the official script from the shared task of CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003). All our experiments were run on Nvidia's *GTX 1080 Ti* GPUs.

5.1 Baseline for TR

N-Gram Tagger Applying the gazetteer to the BIOfid test file gives us the respective baseline for the recognition of taxonomic entities. For evaluation, we use the CoNLL-script and contrast it with easing the conditions by evaluating only the NE predictions and ignoring the prefixed BIO-tagging scheme to every NE. The evaluation does not take into account the other words and is based only on the actual words annotated as TAXON.

Table 4 displays the results for the n-gram tagger. We can nicely see that the increase in size of gazetteers leads to an increase in the final performance. More specifically, for the eased condition, every incremental step from *Latin* to *Latin+German*, and the next step to *All* (i.e. *Latin+German+WikiData*) leads to an increase of +6.64% and +6.56% F-scores, respectively. This matter of fact demonstrates that for the n-gram tagger the resource-size matters.

Gazetteer	CoNLL-Eval	Pr. [%]	Re. [%]	F1 [%]
Lat.	standard	61.50	34.71	44.37
Lat.+Ger.	standard	65.83	45.42	53.75
WikiData	standard	69.05	53.91	60.55
Lat.	eased	92.48	46.04	61.06
Lat.+Ger.	eased	92.94	54.55	67.70
WikiData	eased	95.55	58.87	72.85
All	standard	69.20	55.75	61.75
All	eased	95.57	60.72	74.26

Table 4: Baseline for TR on the BIOfid test file with the N-Gram sequence tagger.

Furthermore, for the eased condition, we see very high scores for precision, however, the recall values are relatively low. This result demonstrates a classic problem of rule-based approaches; as there is no learning process involved, we assume that the performance of the n-gram tagger is highly limited on the features extracted from the source of knowledge (i.e. the amount of information contained in the gazetteer). Besides, no transfer learning is possible from related resources, demonstrating the downsides of non-learning methods.

5.2 Biological NER

We report here the results of our comprehensive survey of five current embedding-based high-performers for biological NER in historical biodiversity literature[12].

The Gold Standard Table 5 contains a detailed summary of all results. In that table, we the report the results which are given by T.H. Nguyen et al. (2019). For the optimized *BiLSTM Tagger*, we achieve excellent results and establish a new state-of-the-art for the first task of TR with 80.23% F-score (see Table 5: BiLSTM-a). For biological NER, we outperform the English counterpart *Co-*

[11] http://github.com/google-research/bert

[12] Our manual inspection of the training data showed that the annotations are content-wise homogeneous, except for the category OTHER. The annotators reported its usage as a residual NE-category for everything which is biologically interesting (e.g. morphology, animal behavior, reproduction, development) but does not fall under the definition of the five major categories. Initial experimental results confirmed its heterogeneous quality. Therefore we omitted OTHER (3,143 sentences) from our further experiments which in turn increased the final performance of NER.

Model	Scores [%]	TAXON	PERSON	LOCATION	ORGANIZATION	TIME	*Overall*
Copious Nguyen (2019)	Precision	77.42	58.92	85.05	N/A	70.67	77.49
	Recall	69.67	48.44	85.63	N/A	54.36	71.89
	F1	73.34	53.17	85.34	N/A	61.45	74.58
BiLSTM-a	Precision	81.33	63.19	66.20	60.24	91.16	75.62
	Recall	79.16	77.45	57.35	67.57	88.16	74.98
	F1	**80.23**	69.60	61.46	63.69	89.63	75.30
Flair Wang2vec	Precision	75.94	61.25	67.58	61.64	90.59	73.58
	Recall	81.37	76.09	62.89	58.11	85.24	75.89
	F1	78.08	71.89	62.63	56.95	87.89	74.30
Flair ELMo	Precision	75.64	67.16	58.31	56.82	90.49	73.05
	Recall	79.92	79.89	65.06	60.81	86.02	76.50
	F1	77.88	69.34	66.30	61.22	88.25	75.01
Flair BERT	Precision	76.63	65.30	66.96	58.00	92.21	74.98
	Recall	77.38	81.02	61.89	58.00	90.33	76.22
	F1	77.01	72.31	64.32	58.00	**91.26**	75.59
BiLSTM-b	Precision	80.45	88.61	72.72	81.21	87.63	79.35
	Recall	76.65	89.40	84.02	70.74	81.17	75.38
	F1	78.50	**89.00**	**77.96**	**75.61**	84.27	**77.31**

Table 5: Results for the task of German biological NER with various neural networks models along the English baseline on the Copious dataset (T.H. Nguyen et al., 2019). All models are trained on the BIOfid dataset and evaluated with the official CoNLL-2003 eval script.

pious for all categories except for LOCATION. For the latter category, the *Copious* dataset contains *9,921* training samples whereas ours has 3,136 fewer samples. We assume that this lower amount results into the lower performance.

With the popular deep language models *Flair*, *ELMo* and *BERT*, we interestingly stay below the performance of the BiLSTM model (except for TIME). Although we utilize the same pre-trained COW word embeddings for all models, we assume that the lower performance arises due to the language models themselves being trained on only a relatively small corpus (ELMo: 100,000 sentences). However, for training ELMo on larger corpora, such as the COW corpus, we would require many months of training time. For the pre-trained Flair and BERT, we can only fine-tune the last tagging layer, not the whole language model itself. This stands in contrast to the BiLSTM model which can be wholly targeted to our domain-specific training data. Hence, this demonstrates the downside of such heavy language models; although they might deliver the top performances, it is difficult to adjust them for lightweight processes, making them impractical for the context of low-resources scenarios.

Data Merging for BiLSTM Tagger For BiLSTM-a, it can be noted that the performance of the standard categories PERSON, ORGANIZATION, and, especially LOCATION is inferior. Therefore, we performed resource-optimization by merging high quality data with our BIOfid dataset in order to increase the training samples for the low performing categories. We merge the datasets of GermEval and CoNLL with our annotated sentences, resulting in train, dev, and test sizes of *46,857*, *6,629*, and *9,437* sentences, respectively. Table 5: BiLSTM-b shows the improvements in performance with the increased dataset. Our results demonstrate the effectiveness of our approach; we do not need to modify the model, rather it is sufficient to perform data-driven optimization. Considering the overall performance, we outperform the English counterpart by $+2,73\%$ F-Score and thus establish a new state-of-the-art for the task of biological NER.

Error Analysis We manually analyze the errors made by the ensemble of neural models. We observe three major issues that compose the absolute majority of errors: a number of *missing annotations* from our experts, *OCR erros* in the raw text and *rare words* that occur frequently in our test dataset. An example of an OCR error is the annotated text span *[1, Juni 1967]* TIME which is misclassified by all models as *1, [Juni 1967]* TIME due to the comma in the date format. Another example is *[KLeebend]* LOC which is not tagged due to the capital "L". Further, the word *[Venn-*

fußfläche] LOC occurs 17 times in the test dataset, but only twice in the training set. It is a three word compound of the words *Venn*, *Fuß* and *Fläche*, that describes a part of the landscape *Vennvorland* in Germany. We conclude that the preprocessing pipeline has to be further refined to remove the OCR errors, while a re-annotation of the data could solve the missing annotations and a more thorough shuffle may solve the rare word issue.

6 Conclusion

In this study, we presented a newly annotated *BIOfid* dataset for German NER in historical biodiversity literature and performed a comprehensive evaluation of the quality of our dataset with five competing neural models. We come to the conclusion that the value of our dataset does not rely solely on the two new entities of TIME and TAXON. By generating domain-related annotation data typical for historical biodiversity literature, we increase the potential performance for biological NER, even for the four standard NE categories. This was demonstrated by the limited scope of the rule-based approach which could not come close to the performance delivered by the neural models and which, in turn, established a new state-of-the-art for both of our selected tasks of TR and NER.

In the course of the annotation process, we discovered that there are further information entities in the BIOfid corpus which do not fall into the definition of standard NE-categories, albeit they are useful from the perspective of biodiversity researchers. For future work, we plan to increase the semantic granularity of the BIOfid dataset by mapping and re-annotating the existing six NE-categories to the top-level hierarchy of *WordNet* (Miller, 1995). This includes 26 categories that can be either *abstract entities* or *concrete entities* (i.e. NE) and can be assigned to specific biological entities, such as morphology, habitat, reproduction, behavioral traits, or species community. By re-annotating the dataset we additionally plan to deliver an inter-agreement value for both the current NER-dataset and the much smaller WordNet-dataset (which is planned to contain an up to 9 times higher amount of annotated information per sentence). Furthermore, we plan to extract all biological entities with the trained neural models from the BIOfid corpus and perform on them the task of *relation extraction* based on current embedding methods.

Overall, our work mobilizes data from undigitized literature leading to huge potentials for biodiversity researchers. It enables cartographic research on the distribution of Central European biodiversity ranging from the pre-modern time up to our current ever increasingly digitizing age.

Acknowledgments

This work was funded by the German Research Foundations (DFG) as part of the *BIOfid* project[13] (DFG-326061700). The BIOfid dataset, the taxonomic gazetteers, the source code, and the supporting appendix (which includes the full annotation guidelines) are uploaded on GitHub[14] for the research community. Special thanks go to W. A. Hemati and G. Abrami for setting up the annotation environment, and to R. Roller, S. Löbbert, Dr. G. Kasperek, Dr. A. Hausinger, Dr. T. Hörnschemeyer and further project members for supporting the annotation process.

References

Giuseppe Abrami, Alexander Mehler, Andy Lücking, Elias Rieb, and Philipp Helfrich. 2019. TextAnnotator: A flexible framework for semantic annotations. In *Proceedings of the Fifteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation, (ISA-15)*, ISA-15.

Sajawel Ahmed and Alexander Mehler. 2018. Resource-Size matters: Improving Neural Named Entity Recognition with Optimized Large Corpora. In *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Darina Benikova, Christian Biemann, and Marc Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *LREC*.

Armin Burkhardt. 2004. 2004. Nomen est omen? : zur Semantik der Eigennamen. In *Landesheimatbund Sachsen-Anhalt e. V. (Hrsg.): "Magdeburger Namenlandschaft" : Orts- und Personennamen der Stadt und Region Magdeburg*.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named Entity Recognition: A Maximum Entropy Approach

[13] http://biofid.de/en/

[14] https://github.com/texttechnologylab/BIOfid

Using Global Information. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*.

Alexandros Komninos and Suresh Manandhar. 2016. Dependency Based Embeddings for Sentence Classification Tasks. In *HLT-NAACL*, pages 1490–1500.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *NAACL-HLT*.

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In *ACL (2)*, pages 302–308.

Wang Ling, Chris Dyer, Alan Black, and Isabel Trancoso. 2015. Two/Too Simple Adaptations of word2vec for Syntax Problems. In *NAACL-HLT*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster. UCREL, IDS.

A Schiller, S Teufel, C Stöckert, and C Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS [Guidelines for tagging German corpora of written language with STTS]. Technical report, Technical Report. Stuttgart, Germany: Institut für maschinelle Sprachverarbeitung [Institute for Machine Language Processing].

Stefan Schweter and Sajawel Ahmed. 2019. Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*. Accepted.

Heike Telljohann, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2012. Stylebook for the Tübingen treebank of written German (TüBa-D/Z).

Scott A Thomson, Richard L Pyle, Shane T Ahyong, Miguel Alonso-Zarazaga, Joe Ammirati, Juan Francisco Araya, John S Ascher, Tracy Lynn Audisio, Valter M Azevedo-Santos, Nicolas Bailly, et al. 2018. Taxonomy based on science is necessary for global conservation. *PLoS biology*, 16(3):e2005075.

Nhung T.H. Nguyen, Roselyn S. Gabud, and Sophia Ananiadou. 2019. COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal*, 7:e29626.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.