Data in Brief ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# MiBio: A dataset for OCR post-processing evaluation

Jie Mei, Aminul Islam, Abidalrahman Moh'd, Yajing Wu, Evangelos E. Milios

*Faculty of Computer Science, Dalhousie University, Canada*

ARTICLE INFO

*Article history:*
Received 11 June 2018
Accepted 24 August 2018

ABSTRACT

We introduce a dataset for OCR post-processing model evaluation. This dataset contains fully aligned OCR texts and the ground truth recognition texts of a English biodiversity book. To better used for benchmark evaluation, we extracted the following information in TSV files: 1) 2907 OCR-generated errors with position in the OCR texts and correction in the ground truth text, 2) ground truth word and sentence segmentation of the OCR texts. In this article, we detail the data preprocessing and provide quantitative data analysis.



## Specifications table

| | |
|---|---|
| Subject area | *Computer Science* |
| More specific subject area | *Natural Language Processing* |
| Type of data | *Text, Table* |
| How data was acquired | *OCR texts are generated from scanned book images by an open source OCR engine (Tesseract 3.0.2) and ground truth texts are generated with additional manual correction.*<br>*Tables contain information (i.e. ground truth OCR tokens and OCR error corrections) extracted from the texts.* |

DOI of original article: https://doi.org/10.1016/j.ipm.2018.06.001
*E-mail addresses:* jmei@cs.dal.ca (J. Mei), aminul@louisiana.edu (A. Islam), amohd@cs.dal.ca (A. Moh'd), yajing@cs.dal.ca (Y. Wu), eem@cs.dal.ca (E.E. Milios).

https://doi.org/10.1016/j.dib.2018.08.099




Please cite this article as: J. Mei, et al., MiBio: A dataset for OCR post-processing evaluation, Data in Brief (2018), https://doi.org/10.1016/j.dib.2018.08.099

| | |
|---|---|
| Data format | *OCR & ground truth text: raw text, page separated, tokenized; Table: CSV, analyzed.* |
| Experimental factors | *Word/Character Error Rate of OCR Text, Levenshtein edit distance distribution of OCR errors.* |
| Experimental features | *OCR texts have relatively low error rate but contains complex OCR errors.* |
| Data source location | |
| Data accessibility | *Data is with this article and also in GitHub* https://github.com/jmei91/MiBio-OCR-dataset |
| Related research article | *Jie Mei,* Aminul Islam, Abidalrahman Moh'd, Yajing Wu, Evangelos E. Milios. *Statistical Learning for OCR Error Correction. 2017. Information Processing and Management.* ***in press****.* |

## Value of the data

- This dataset can be used for evaluating and comparing the performance of OCR post-processing models.
- This dataset contains page separated OCR texts and according ground truth texts. The line and paragraph breaks in the source image are preserved in both text versions. Thus each line in both OCR and ground truth texts are fully aligned and can easily refers to each other.
- OCR errors are extracted and listed in table. For each OCR error, we record its correction in the ground truth text and position in the OCR text.
- We provide the ground truth word and sentence segmentation for OCR texts to disambiguate word and sentence boundary and to be served as a reference when evaluating the tokenization performance of post-processing models.

## 1. Data

We made available Mining Biodiversity (MiBio) dataset with 2910 OCR-generated errors along with the OCR and the ground truth recognition texts for benchmark testing. The OCR text was generated from the book titled "Birds of Great Britain and Ireland (Volume II)" [1] and made publicly available by the Biodiversity Heritage Library (BHL) for Europe[1] using Tesseract 3.0.23[2]. The ground truth text is based on an improved OCR output[3] and adjusted manually to match with the original content of the whole book.

The scanned image data of the book contains 460 page-separated files, where the main content is included in 211 pages. The scanned images and different format of raw OCR outputs are online accessible and downloadable on https://archive.org/download/birdsofgreatbrit02butl.

## 2. Experimental design, materials, and methods

### 2.1. OCR and ground truth recognition texts preprocessing

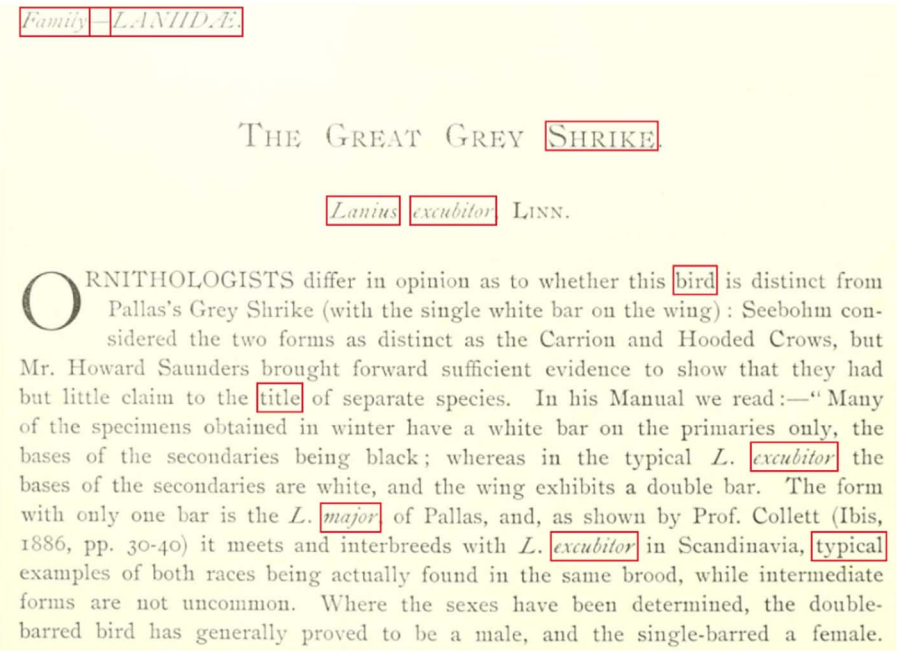
The dataset is generated from two OCR outputs for book "Birds of Great Britain and Ireland (Volume II)" [1]. One version is generated from the standard BHL-Europe recognition workflow, which OCR technique is based on Tesseract 3.0.23. We manually correct the OCR errors in the OCR outputs to

[1] http://www.biodiversitylibrary.org/item/35947#page/13/mode/1up
[2] https://github.com/tesseract-ocr/tesseract
[3] http://www.bhle.eu/en/results-of-the-collaboration-of-bhl-europe-and-impact

Please cite this article as: J. Mei, et al., MiBio: A dataset for OCR post-processing evaluation, Data in Brief (2018), https://doi.org/10.1016/j.dib.2018.08.099

be the ground truth. We then remove footnotes and page numbers in both versions to keep the conte fluency over pages.

a



*Family*—*LANIIDÆ*.

THE GREAT GREY SHRIKE.

*Lanius excubitor* LINN.

ORNITHOLOGISTS differ in opinion as to whether this bird is distinct from Pallas's Grey Shrike (with the single white bar on the wing) : Seebohm considered the two forms as distinct as the Carrion and Hooded Crows, but Mr. Howard Saunders brought forward sufficient evidence to show that they had but little claim to the title of separate species. In his Manual we read :—" Many of the specimens obtained in winter have a white bar on the primaries only, the bases of the secondaries being black ; whereas in the typical *L. excubitor* the bases of the secondaries are white, and the wing exhibits a double bar. The form with only one bar is the *L. major* of Pallas, and, as shown by Prof. Collett (Ibis, 1886, pp. 30-40) it meets and interbreeds with *L. excubitor* in Scandinavia, typical examples of both races being actually found in the same brood, while intermediate forms are not uncommon. Where the sexes have been determined, the double-barred bird has generally proved to be a male, and the single-barred a female.

b

```
Faiuiiv^LAXIILKl-:.

The Great Grey Shrh^e.

Lauius txciibilor, LiNN.

ORNITHOLOGISTS differ in opinion as to whether this biixl is distinct from
Pallas's Grey Shrike (with the single white bar on the wing) : Seebohm con-
sidered the two forms as distinct as the Carrion and Hooded Crows, but
Mr. Howard Saunders brought forward sufficient evidence to show that they had
but little claim to the title of separate species. In his Manual we read : - " Many
of the specimens obtained in winter have a white bar on the primaries only, the
bases of tlie secondaries being black ; whereas in the typical L. cxcubitor the
bases of the secondaries are white, and the wing exhibits a double bar. The form
with only one bar is the L. viajor, of Pallas, and, as shown by Prof. Collett (Ibis,
1886, pp. 30-40) it meets and interbreeds with L. exaibiior in Scandinavia, t3'pical
examples of both races being actually found in the same brood, while intermediate
forms are not uncommon. Where the sexes have been determined, the double-
barred bird has generally proved to be a male, and the single-barred a female.
```

**Fig. 1.** A image segment (a) and its according OCR-generated text (b) of the evaluation dataset. The recognition errors are highlighted in red.

Please cite this article as: J. Mei, et al., MiBio: A dataset for OCR post-processing evaluation, Data in Brief (2018), https://doi.org/10.1016/j.dib.2018.08.099

### 2.2. OCR error extraction

When generating the error list, we adopted the following rules in extracting the OCR errors in aligned contents from the OCR and the ground truth texts:

- when segmenting an OCR-generated string into substrings that match with tokens in the ground truth text, the separating positions are approximated manually to make the best guess. For example, given an OCR string "fFrinn̄HluurJ" aligns with "(Fringillinæ)" in the ground truth, we separated this string into three error-correction mappings: $< \text{f} \rightarrow ) >$, $< \text{FrinHluurJ} \rightarrow \text{Fringillinæ} >$, and $< \text{J} \rightarrow ) >$. In another example, given an OCR string "countr}" and "country," in the ground truth, we split it as two error-correction mapping: $< \text{countr}\} \rightarrow \text{country} >$ and $< \rightarrow , >$.
- *Two ASCII substitution of unicode characters are allowed*: (æ, ae) and (Æ, AE). Note that the dataset is generated from a biodiversity book, which contains terminologies with non-English characters, for example, Corvidæ or ORIOLIDÆ. We accept these two ASCII substitutions in order to match the original terminologies to their English counterparts.
- *The aligned two words with different cases is not treated as an error.* Observed that the standard BHL-Europe recognition workflow is tend to lowercase the non-heading characters in some entirely capitalized words. Thus, we do not categorized this type of mismatchings as error. Such change in capitalization form is also hard to detect by human readers with only input text when page layout is eliminated.
- *The extra whitespaces between tokens are allowed.* It is also observed that the standard BHL-Europe recognition workflow generate extra whitespaces between tokens. We do not categorize this type of mismatch as error unless the inserted whitespace leads to a splitting or merging error.

### 2.3. OCR text tokenization

Tokenizing OCR text is one internal step in OCR post-processing. The tokenization performance affect downstream error detection and correction. Since intra-word characters of OCR errors can be misrecognized as punctuation, it is hard to disambiguate the misrecognized punctuation with true punctuation in an OCR text and thus lead to high token boundary ambiguities. We thus provide the ground truth OCR tokens for evaluating the tokenization performance of OCR post-processing models. The ground truth tokens are generated by first tokenizing the ground truth recognition text and maps the segmentation positions to the OCR texts.

### 2.4. Dataset analysis

To have a close look at the OCR input/output, we sample a segment of OCR-generated text with original scanned image In Fig. 1. Table 1 shows the OCR performance, measured by precision and recall, indicating a high quality OCR output with low error rate in both word- and character-level measurements.

Referencing to the ground truth OCR tokens in the dataset, we quantitative analysis the tokenization performance on the OCR texts by tokenization different schemes including the Whitespace, Penn Treebank, WASTE [2] and Elephant [5]. The results are shown in Table 2. Observe that some OCR errors are orthographically far from their correction, we further analysis the distribution of error words with respect to Levenshtein edit distance [3] in Table 3.

**Table 1**
The precision and recall of OCR generated text.

| Measure | Character-wise | Word-wise |
|---|---|---|
| **Precision** | 1 − 6362/409236 = 98.45% | 1 − 2906/101700 = 97.14% |
| **Recall** | 1 − 6362/407194 = 98.44% | 1 − 2906/98097 = 97.04% |



Please cite this article as: J. Mei, et al., MiBio: A dataset for OCR post-processing evaluation, Data in Brief (2018), https://doi.org/10.1016/j.dib.2018.08.099

**Table 2**
The performance of different tokenization schemes on the OCR text.

| Tokenization method | Measure [%] | | | |
|---|---|---|---|---|
| | Prec. | Rec. | F1 | Err. |
| **Whitespace convention** | 85.5 | 73.14 | 78.84 | 94.93 |
| **Penn Treebank convention** | 94.33 | 93.94 | 94.13 | 11.08 |
| **WASTE** (Jurish 2013) | 95.18 | 93.14 | 94.15 | 11.05 |
| Elephant (Evang et. al. 2013) | 95.17 | 93.18 | 94.16 | 11.03 |

**Table 3**
The Levenshtein distance distribution of the errors in the OCR texts.

| Edit distance | Error statistics | | Sample OCR error | |
|---|---|---|---|---|
| | Number | Percent [%] | Correction | Error |
| **1** | 889 | 30.58 | galbula | ga/bula |
| **2** | 1376 | 47.35 | yellowish | j̈ellowish |
| **3** | 307 | 10.56 | bents | Ijcnts |
| **4** | 148 | 5.09 | my | ni}' |
| **5** | 70 | 2.41 | Lanius | Lioiiits |
| **6** | 51 | 1.75 | minor | )iii > iof |
| **7** | 28 | 0.96 | garrulus | f;ay > ///us |
| **8** | 16 | 0.55 | curvirostra | iUi'7'iyosira |
| **9** | 5 | 0.17 | Nucifraga | Aiiii/ruťd |
| **> = 10** | 16 | 0.55 | pomeranus | poiiui-iVtiis |
| **Total** | 2906 | 100 | | |

Q9 **Uncited reference**

[4].

## References

[1] Arthur G. Butler, Frederick William Frohawk, H. Gr¨onvold, Birds of Great Britain and Ireland. Order Passeres, Brumby & Clarke, Hull, 1907.
[2] B. Jurish, Word and sentence tokenization with Hidden Markov Models, J. Lang. Technol. Comput. Linguist. 28 (2) (2013) 61–83.
[3] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, Sov. Phys. Dokl. 10 (1966) 707.
[4] Jie Mei, Aminul Islam, Abidalrahman Moh'd, Yajing Wu, Evangelos E. Milios, Statistical learning for OCR error correction, Inf. Process. Manag. (2018) (In press).
Q10
[5] K. Evang, V. Basile, G. Chrupała, J. Bos, Elephant: Sequence Labeling for Word and Sentence Segmentation, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013) pp. 1422–1426.



Please cite this article as: J. Mei, et al., MiBio: A dataset for OCR post-processing evaluation, Data in Brief (2018), https://doi.org/10.1016/j.dib.2018.08.099