The Labeled Segmentation of Printed Books

Lara McConnaughey
Computer Science Division
University of California, Berkeley
laram@berkeley.edu

Jennifer Dai
Computer Science Division
University of California, Berkeley
jenniferdai@berkeley.edu

David Bamman
School of Information
University of California, Berkeley
dbamman@berkeley.edu

Abstract

We introduce the task of *book structure labeling*: segmenting and assigning a fixed category (such as TABLE OF CONTENTS, PREFACE, INDEX) to the document structure of printed books. We manually annotate the page-level structural categories for a large dataset totaling 294,816 pages in 1,055 books evenly sampled from 1750–1922, and present empirical results comparing the performance of several classes of models. The best-performing model, a bidirectional LSTM with rich features, achieves an overall accuracy of 95.8 and a class-balanced macro F-score of 71.4.

1 Introduction

The availability of large-scale book corpora (such as those created by Google Books, the Internet Archive and the HathiTrust) has driven a flurry of work in cultural analytics over the past decade, in which the text contained in historical books has provided the raw material for the analysis of genre (Underwood, 2016), literary character (Bamman et al., 2014), geographic attention (Wilkens, 2013), fame (Michel et al., 2010), and much more.

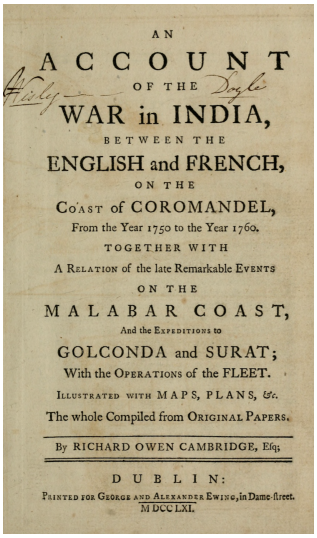
Books, however, are not undifferentiated streams of text in the same way that born-digital documents like tweets or emails are; they are physical objects with materiality (Werner, 2012) and are arranged in a complex structure steeped in a long design tradition, with the core content of the work placed between structured frontmatter (such as a table of contents and introduction) and backmatter (such as an appendix and index). Not all of this content is desirable for all analyses; as we show below, 11% of all pages in books belong to the peritext (Genette, 1987) that surrounds the core content, with wide variability from book to book. For other analyses, such as those addressing questions in book history (Kirschenbaum and Werner, 2014), isolating this structure in a consistent way across historical documents can enable research into how the form of the printed book has, for example, changed over time.

While other work has focused on extracting the idiosyncratic structure inherent in each book, such as recognizing chapter boundaries in order to automatically generate a table of contents, or link a parsed table of contents to positions in a book (Déjean and Meunier, 2005, 2009; Wu et al., 2013; Gao et al., 2009), labeling document segments with a fixed typology has complementary benefits, allowing researchers to identify consistent categories in all books regardless of the names assigned by a specific author or publisher, or popular at a given time.[1]

At the same time, book structure labeling presents real challenges to automatic identification. While large-scale digital collections originate in page scans of the books, the most common form of access by researchers is through the output of OCR; raw image files are prohibitively expensive both in terms of disk space (15.1 million books from the HathiTrust consumes 677 ter-

[1]For example, a section whose function is to outline the structural regions of a book and list the pages on which they begin may be known at different points in history as a "table of contents," "index," or several other terms.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 746–756
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics



AN

ACCOUNT

OF THE

WAR in INDIA,

BETWEEN THE

ENGLISH and FRENCH,

ON THE

COAST of COROMANDEL,

From the Year 1750 to the Year 1760.

TOGETHER WITH

A RELATION of the late Remarkable EVENTS

ON THE

MALABAR COAST,

And the EXPEDITIONS to

GOLCONDA and SURAT;

With the OPERATIONS of the FLEET.

ILLUSTRATED WITH MAPS, PLANS, &c.

The whole Compiled from ORIGINAL PAPERS.

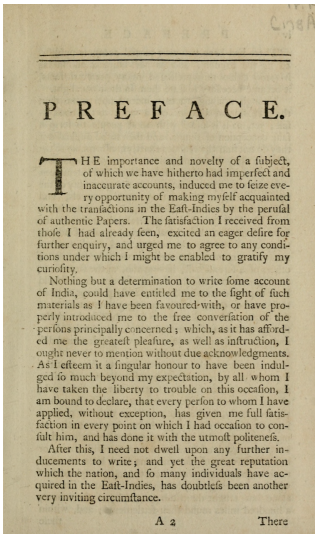
By RICHARD OWEN CAMBRIDGE, Esq;

DUBLIN:

PRINTED FOR GEORGE AND ALEXANDER EWING, in Dame-street.

M DCC LXI.

(a) Title page.



PREFACE.

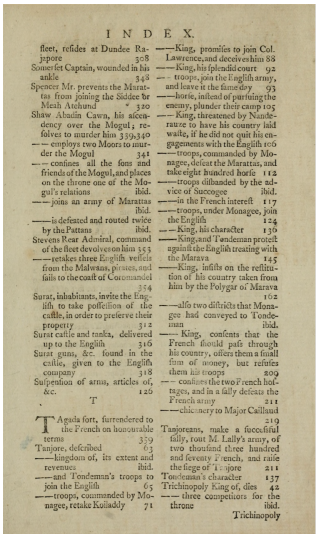
THE importance and novelty of a subject, of which we have hitherto had imperfect and inaccurate accounts, induced me to seize every opportunity of making myself acquainted with the transactions in the East-Indies by the perusal of authentic Papers. The satisfaction I received from those I had already seen, excited an eager desire for further enquiry, and urged me to agree to any conditions under which I might be enabled to gratify my curiosity.

Nothing but a determination to write some account of India, could have entitled me to the sight of such materials as I have been favoured with, or have properly introduced me to the free conversation of the persons principally concerned; which, as it has afforded me the greatest pleasure, as well as instruction, I ought never to mention without due acknowledgments. As I esteem it a singular honour to have been indulged so much beyond my expectation, by all whom I have taken the liberty to trouble on this occasion, I am bound to declare, that every person to whom I have applied, without exception, has given me full satisfaction in every point on which I had occasion to consult him, and has done it with the utmost politeness.

After this, I need not dwell upon any further inducements to write; and yet the great reputation which the nation, and so many individuals have acquired in the East-Indies, has doubtless been another very inviting circumstance.

A 2 There

(b) Preface.



INDEX.

Trichinopoly

(c) Index.

Figure 1: *An Account of the War in India, Between the English and French* (1761). From the HathiTrust.

abytes of space[2]) and in the resources required for computational processing. While people are able to distinguish the different sections of a book with ease, the degraded nature of the OCR output (especially for historical books) blurs the clear markers that signal the category to human readers—both in terms of the lexical signals like "Preface" or "Index" that head a page, and its visual structure as well. Figure 1 illustrates an example of three pages from a single book drawn from the HathiTrust; figure 2 displays the corresponding OCR output; the degradation introduced by OCR affects not only the accuracy of character and word identification, but also the structural layout as well.

To address these limitations and enable research that depends on reasoning over fine-grained document structure within books, we introduce the task of labeled segmentation, and make the following contributions:

- We create an human-annotated gold standard of 294,816 pages in 1,055 printed books drawn from the HathiTrust Digital Library.
- We approach this problem in the most common resource-deficient scenario researchers most frequently encounter: with access only to the pre-existing output of OCR.
- We compare several different classes of models, including a fast independent predictor (a random forest), a simple linear sequence labeling model (CRF), and a sequence labeling bidirectional LSTM that can capture non-linearities in the feature space. All data and pre-trained models are openly available to the public at https://github.com/dbamman/book-segmentation.

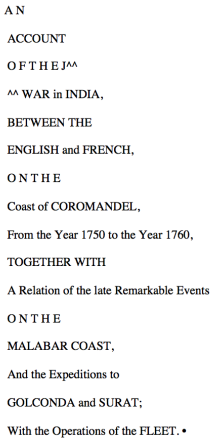
[2] https://www.hathitrust.org/statistics_visualizations

2 Data

In order to support the analysis and prediction of labeled document structure, we present a manually annotated dataset of 1,055 books, where each page has been labeled according to one of 10 categories described in §2.1 below. All books originate in the HathiTrust Digital Library. In order to capture historically representative phenomena, we use the decade-stratified sample of 1,075 books from Bamman et al. (2017), in which each decade from 1750-1922 is roughly equally represented. From this sample of 1,075 apparent books, we exclude all non-book records (including digitized newspaper clippings, unbound pamphlets and reports, opera programs, etc.) to yield a total labeled dataset of 294,816 pages in 1,055 books.

2.1 Categories

While there is no codified form of the standard categories that are present in print books, modern book designers generally adhere to a tradition involving a typical sequence of parts (Wilson, 1993; Lee, 2009). We draw on this tradition to inform our set of the following ten categories; to contextualize its prevalence, each category is listed with its description and the fraction of the 1,055 books



AN

ACCOUNT

OF THE J^^

^^ WAR in INDIA,

BETWEEN THE

ENGLISH and FRENCH,

ON THE

Coast of COROMANDEL,

From the Year 1750 to the Year 1760,

TOGETHER WITH

A Relation of the late Remarkable Events

ON THE

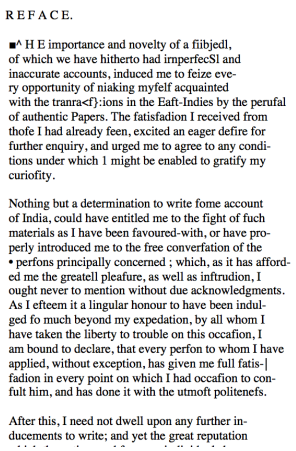
MALABAR COAST,

And the Expeditions to

GOLCONDA and SURAT;

With the Operations of the FLEET. •

(a) Title page.



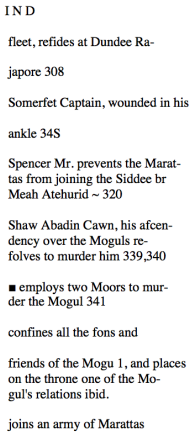
REFACE.

■^ H E importance and novelty of a fiibjedl, of which we have hitherto had irnperfecSl and inaccurate accounts, induced me to feize every opportunity of niaking myfelf acquainted with the tranra<f}:ions in the Eaft-Indies by the perufal of authentic Papers. The fatisfadion I received from thofe I had already feen, excited an eager defire for further enquiry, and urged me to agree to any conditions under which 1 might be enabled to gratify my curiofity.

Nothing but a determination to write fome account of India, could have entitled me to the fight of fuch materials as I have been favoured-with, or have properly introduced me to the free converfation of the • perfons principally concerned ; which, as it has afforded me the greatell pleafure, as well as inftrudion, I ought never to mention without due acknowledgments. As I efteem it a lingular honour to have been indulged fo much beyond my expedation, by all whom I have taken the liberty to trouble on this occafion, I am bound to declare, that every perfon to whom I have applied, without exception, has given me full fatis-| fadion in every point on which I had occafion to confult him, and has done it with the utmoft politenefs.

After this, I need not dwell upon any further inducements to write; and yet the great reputation

(b) Preface.



IND

fleet, refides at Dundee Ra-

japore 308

Somerfet Captain, wounded in his

ankle 34S

Spencer Mr. prevents the Marattas from joining the Siddee br Meah Atehurid ~ 320

Shaw Abadin Cawn, his afcendency over the Moguls refolves to murder him 339,340

■ employs two Moors to murder the Mogul 341

confines all the fons and

friends of the Mogu 1, and places on the throne one of the Mogul's relations ibid.

joins an army of Marattas

(c) Index.

Figure 2: OCR output for the page scans illustrated in fig. 1.

in our dataset in which it appears (for example, 47.8% of books have an annotated preface).

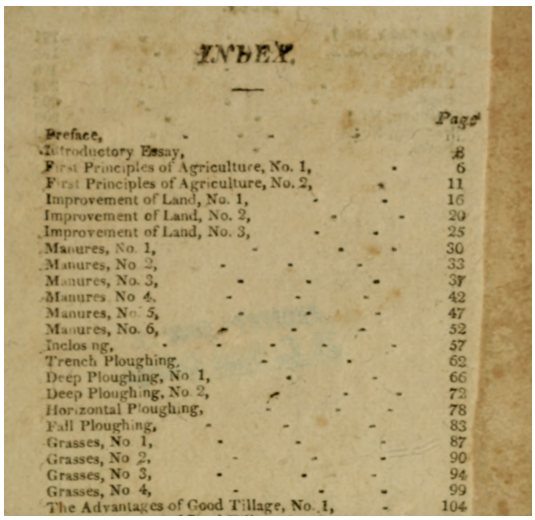
- TITLE PAGE (93.0%), which lists the title of the work and (optionally) other information including the names of the author, translator, and others involved in its creation. In this category we group the primary title page along with the HALF-TITLE (a page that generally only presents the title of the work, often preceding the main title page or first chapter).
- AD CARD (18.1%), which lists other works by the author or publisher; or, more generally, any other object that is advertised for sale.
- PUBLISHER (39.9%), which includes the modern COPYRIGHT page (typically on the verso side of the title page) and also the COLOPHON (an imprint often appearing at the end of a work).
- DEDICATION (17.5%), an inscription by the author dedicating the work to another.
- PREFACE (47.8%), which includes a FOREWORD, PREFACE, and INTRODUCTION. While modern designers articulate prescriptive distinctions among these categories primarily in their subject matter and authorial voice,[3] we do not find a strong distinction among these sub-categories evident when labeling the text. We therefore follow Genette (1987) in grouping all together as prefatory material.
- TABLE OF CONTENTS (46.8%), which includes "an accurate listing of all textual matter which follows it and the page on which the parts of the book commence" (Wilson, 1993).
- TEXT (99.3%), which includes the main contents of the book. TEXT is naturally the most frequent category, but only accounts for 89.4% of pages in all books in our dataset. We also see wide variability from book to book; the average TEXT ratio in books is 0.82, with a standard deviation of 0.18.
- APPENDIX (14.4%) includes a heterogeneous mix of other minor categories that appear infrequently in different books. These include: NOTES (1.1%) (which "have the character of footnotes which, because of their extent, are placed at the back of the book" (Wilson, 1993)); BIBLIOGRAPHY (1.7%), "a listing of the books and periodicals, etc., which the author has used as source material or which he recommends as supplementary reading matter" (Wilson, 1993); GLOSSARY (0.6%), "a list of definitions of terms used in the text" (Wilson, 1993), ERRATA (4.1%), mistakes corrected in the printing of the book, and SUBSCRIBERS (1.7%), a list of individuals who have committed to purchasing the work in advance (a historical category not frequently seen in modern texts). We annotate each of these subcategories individually for future work, but collapse them into the single category of APPENDIX for the work below.
- INDEX (19.2%), which "serves to catalogue,

[3] "A preface is written by the author and is generally about the writing of the book. A foreword is a comment on the book and/or the author by another person. An introduction, which may by the author or another, may contain such matter, but it's primarily a preparation for, or explanation of, the content" (Lee, 2009)

with page indications, all the references which an author wishes to identify" (Wilson, 1993).

- N/A. For each of the nine categories above, we annotate the beginning and end pages present in a book; any page not contained within a labeled section receives the label `N/A`.

As Genette (1987) articulates, each of these structural categories mediates the relationship between the text and its audience, and each serves a different illocutionary purpose. The TITLE is addressed to the general public (not necessarily the readers) and is not only informational (informing of the name of the work), but also serves as important marketing material as well; PREFACES are addressed to readers, and may be written either by the author of the core content (*authorial*) or by another (*allographic*) and communicate the intention or interpretation of the work; the illocutionary force of a DEDICATION, in contrast, is performative: its very presence is a speech act that serves to dedicate the work to another.


INDEX.

	Page
Preface,	iii
Introductory Essay,	5
First Principles of Agriculture, No. 1,	6
First Principles of Agriculture, No. 2,	11
Improvement of Land, No. 1,	16
Improvement of Land, No. 2,	20
Improvement of Land, No. 3,	25
Manures, No. 1,	30
Manures, No. 2,	33
Manures, No. 3,	37
Manures No. 4,	42
Manures, No. 5,	47
Manures, No. 6,	52
Inclosing,	57
Trench Ploughing,	62
Deep Ploughing, No. 1,	66
Deep Ploughing, No. 2,	72
Horizontal Ploughing,	78
Fall Ploughing,	83
Grasses, No. 1,	87
Grasses, No. 2,	90
Grasses, No. 3,	94
Grasses, No. 4,	99
The Advantages of Good Tillage, No. 1,	104

Figure 3: Table of contents page listed as "Index."

For all categories, we label based on the tenor of the category's meaning, and not on the title of the section that may appear on the page. Figure 3 illustrates one such example of this distinction—a page whose function is to serve as a table of contents but is headed as an "Index" (and also appears at the back of the book, like contemporary indices); rather than functioning as an index in providing references to concepts within the text, it outlines the organizational structure of the sections (as a table of contents does).

Human judgments of these ten categories are relatively uncontroversial; to calculate the coherence of the task, we calculated the inter-annotator agreement rate for two annotators on 25 books, and find a chance-corrected Cohen's $\kappa = 0.83$, suggesting a very high level of agreement.[4] All books then receive a single judgment of page-level annotations by a single annotator.

3 Methods

To explore our ability to label book structure automatically, we test three different feature-rich classifiers. All make use of the same set of features.

3.1 Features

Keywords. Most words on a page are not predictive of the category to which it belongs; a word like *Britain* in a biography of Churchill may distinguish that book from other books, but will also equally be found on the title page, table of contents, preface, content, index, or any other category. Some words, however, are discriminative, such as the titles of the categories ("index," "preface," "dedication," etc.). To identify these terms, we train a unigram logistic regression classifier on the training-only partition of the data (described in section 4 below) and manually select keywords with high face validity. We create two sets of features from these keywords: presence of a keyword in the header of the page (the first four lines) and the presence of a keyword anywhere on the page.

Longest increasing subsequence. As figure 3 shows, tables of contents are distinguished from indices in that the page numbers generally increase from the top of the page to the bottom, corresponding to the linear order of the book. To capture this, we create a feature for the longest increasing subsequence (LIS) of numbers on the page. The LIS for any set of n randomly permuted numbers converges to a Tracy-Widom distribution (Baik et al., 1999); to enable feature value comparisons across pages with different total numbers, we conduct a permutation test by shuffling the numbers on the page and recalculating the LIS for that resample; we set the feature value to be 1 only when the observed LIS is greater than 5% of the LIS scores for the permutations (i.e., $p < 0.05$).

[4] Using the non-parametric bootstrap to account for the size of the sample in our confidence of the agreement rate, we find a 95% confidence interval for κ to be within the interval [0.65, 0.94].

Alphabetical sort. Indices, in contrast, are distinguished from tables of contents in that their lines are sorted alphabetically (from the top to the bottom of the page). To capture this, we create a feature measuring the degree to which the lines on a page are sorted, operationalized as the Spearman rank correlation coefficient (ρ) between the set of lines in their original order and the lines in sorted order. Perfectly sorted lines have a $\rho = 1$; inversely sorted lines have $\rho = -1$ and randomly ordered lines have an expected $\rho = 0$. To account for random sorting that take place by chance, we set this feature value to be ρ only when its p value (rejecting $H_0 \equiv \rho = 0$) < 0.01.

Letter distribution. In addition to measuring the degree which the full page is alphabetized, we can also capture important structural qualities of indices by measuring the degree to which initial letters in words are overrepresented on the page. We calculate this by measuring the empirical distribution of initial downcased letters `[a-z]` for all words in the book, and measure the degree to which the empirical distribution on the page overrepresents any individual letter. Rather than comparison the full distributions (using e.g., Jensen-Shannon divergence), we calculate the number of letters whose frequency on the page deviates from the book frequency for that letter by a z-score (accounting for the number of times we observe the letter) corresponding to a critical value $\alpha \leq 0.05$.

Roman numerals. Frontmatter preceding the main content is often paginated with roman numerals, rather than the arabic numerals found in the content. To capture this, we create a binary feature identifying the presence of roman numerals in the first four lines (header) or last four lines (footer) of the page, using the resources of Underwood (2017).

Page density. Content pages are relatively dense with characters (both letters and numbers); title pages and tables of contents are defined by greater volume of whitespace. To capture this differential, we introduce features for the ratio of words and numbers among all (whitespace-delimited) tokens and for the overall number of tokens observed.

Position. We create a set of binary features marking the position of the page within the book (appearance in the first ten pages, last ten pages, and in which quintile it appears), and its real-valued positional ratio within the book (page number divided by the total pages).

Page Sequence. Not all books distinguish frontmatter from the main content with roman numerals; to address this, we identify the page with the first marked page number and create a feature that captures whether a page appears before or after that first marked page.

TextTiling While all words are not indicative of the categories on their own, they can provide a natural segmentation of the book into discrete discourse chunks, in that the language that characterizes a given main content section may differ from that within an introduction (and certainly from more structured sections like indices or tables of contents). To capture this, we create a feature for each page derived from TextTiling (Hearst, 1997): for a given page at position i, we calculate the cosine similarity between the intervals [page$_1$, page$_{i-1}$] and [page$_i$, page$_n$].

The feature classes above total 172 features for each individual page. When representing a page as input to the models below, we also conjoin information about all pages within a window of three pages around the target page; each page is thus represented by a total of 7×172 distinct features. All non-binary features are standardized to standard normals, whose means and variances are estimated using the distribution observed in the training-only partition of the data.

3.2 Models

We compare three different model classes: a random forest (Breiman, 2001), which can capture complex nonlinearities in the feature space but is constrained to make independent predictions; an ℓ_2-regularized conditional random field (Lafferty et al., 2001), which can account for temporal dependencies in the predictions but is limited to linear relationships; and a bidirectional sequence labeling LSTM (Graves, 2012; Ma and Hovy, 2016), which can reason over sequential information while also capturing more complex non-linearities. The observed input to all methods for each page x_i is the same feature representation $f(x_i)$; the CRF also includes information about label transition features, decoding the entire sequence using Viterbi decoding; and the bidirectional LSTM captures persistent state information for each page as two H-dimensional hidden layers, one for the forward direction h_f and one for

Method	Accuracy	Macro precision	Macro recall	Macro F
Majority class	0.888	0.089	0.100	0.094
Random Forest	0.959 [0.947, 0.969]	0.866 [0.831, 0.894]	0.593 [0.555, 0.632]	0.677 [0.641, 0.715]
CRF	0.940 [0.915, 0.959]	0.654 [0.615, 0.695]	0.744 [0.683, 0.835]	0.686 [0.644, 0.740]
BiLSTM	0.958 [0.947, 0.968]	0.776 [0.741, 0.807]	0.670 [0.630, 0.709]	0.714 [0.679, 0.747]

Table 1: Full segment labeling, along with 95% bootstrap confidence intervals.

the backward direction h_b (we set $H = 25$ in these experiments). Predictions for each time step i are made using the vector concatenation of $[h_f^i; h_b^i]$.

4 Evaluation

We compare the performance of the three models described above at the task of page-level labeling: both the multiclass classification problem of predicting which of the 10 labels applies to each page, and the binary task of {TEXT, NON-TEXT} prediction, in which the nine front- and backmatter labels are collapsed into the single label NON-TEXT; while the former allows access to fine-grained categories of (e.g.) indices and tables of contents, the latter covers the common scenario where researchers are interested only in isolating where the core text begins and ends.

Experimentally, we divide the full training data into two partitions: a training-only partition of 400 books, on which we experiment with feature and model development, and a test partition of the remaining 655 books. All results presented below are the result of tenfold cross-validation on the test partition. Each fold trains on $\frac{8}{10}$ of the test data, uses $\frac{1}{10}$ of the 655 books as development for model selection (e.g., to optimize the ℓ_2 regularization parameter for the CRF, terminate training for the BiLSTM, and optimize the depth of the random forest), and uses $\frac{1}{10}$ of the 655 for test. We supplement each training fold with the 400 books from the training-only partition, but this data is never used for evaluation below.

In total, we evaluate the performance on 655 books and calculate 95% confidence intervals for each metric using the non-parametric bootstrap, drawing $B = 10,000$ resamples of books (not individual pages) in order to account for the statistical dependence between page-level predictions.

4.1 Full segment labeling

Table 1 presents the results for full multiclass segment labeling. To contextualize these results, we also provide a simple baseline of predicting the majority class (TEXT) for all pages; since most pages in a book are core content, this yields a high absolute accuracy against which to compare, but a low macro precision/recall/F score (which evenly weights the importance of each class).

All three methods achieve relatively similar performance when measured by absolute accuracy (though the room for improvement over the baseline is small). When treating all classes as equally important and measuring by the macro F score, both sequence labeling methods (CRF and bidirectional LSTM) show slight improvements over the independent predictions of a random forest, but not significantly so, suggesting that the feature representation of the book (which they all share as identical input) is perhaps a strong enough signal that mitigates the label dependencies.

Category	Precision	Recall	F	True n
Title	0.782	0.751	0.766	887
Dedication	0.630	0.489	0.551	188
Pubinfo	0.697	0.590	0.639	261
Ad card	0.642	0.516	0.572	717
TOC	0.844	0.842	0.843	1,139
Preface	0.736	0.643	0.686	2,253
Text	0.971	0.991	0.981	160,721
Index	0.894	0.628	0.737	2,586
Appendix	0.688	0.412	0.515	2,460
N/A	0.894	0.801	0.845	9,791

Table 2: Individual category results, BiLSTM.

Table 2 lists the precision, recall and F-score results by category for the best-performing model (bidirectional LSTM). Several categories are worth calling out: the precision and recall for recognizing table of contents is high (≥ 0.84 for both metrics), suggesting that this method may provide a solid foundation for work in book structure extraction that relies on an identified table of contents in order to recognize the idiosyncratic structure of books. Title page and index recognition are also relatively high (0.89 precision/0.63 recall); what these three categories have in common are strong structural features (the distribution of ink and whitespace on the page; regularities in the numbers and the degree of alphabetization).

While dedications and publication information

Method	Accuracy	Macro P	Macro R	Macro F
Majority class	0.888	0.444	0.500	0.470
Chop	0.857	0.804	0.692	0.725
Random Forest	0.966 [0.953, 0.976]	0.947 [0.937, 0.956]	0.877 [0.836, 0.911]	0.908 [0.877, 0.931]
CRF	0.963 [0.949, 0.973]	0.887 [0.843, 0.921]	0.920 [0.896, 0.941]	0.902 [0.872, 0.926]
BiLSTM	0.965 [0.953, 0.974]	0.938 [0.924, 0.951]	0.881 [0.843, 0.913]	0.907 [0.880, 0.928]

Table 3: Content/non-content labeling, along with 95% bootstrap confidence intervals.

are both relatively infrequent (often occupying a single page in a book), the greatest point of confusion is in separating the main content from the structurally similar pages that typically precede it (in the preface) and follow it (in the appendix). While confusion between PREFACE/TEXT and APPENDIX/TEXT account for most of the errors, figure 4 illustrates several difficult cases and exemplary mistakes in the other categories: fig. 4a is a page that blurs the line between an index and table of contents; fig. 4b is an advertisement for a book "in the press and speedily will be published"; and fig. 4c is a dedication that, without strong lexical indicators, is mistaken for a title page.

In order to understand the contribution that individual features make on the predictions, we carry out an ablation test for each feature class, in which we remove a feature class from the model and perform exactly the same training and test procedure as described in section 4: we train a model on the training fold supplemented with the 400 books in the training partition, perform hyperparameter optimization on development data, and report accuracy on the held-out test fold, repeating ten times, once for each fold in cross-validation.

Feature	Δ Macro F-score
–Keywords	-0.15
–Position	-0.03
–Density	-0.02
–Window	-0.01
–Roman	-0.01
–Letter	-0.01
–LIS	0.00
–TextTiling	0.00
–Page sequence	0.00
–Alphabetical	0.01

Table 4: Feature ablation results for the BiLSTM model, illustrating the change in macro F-score that results by removing a given feature class from the full model.

The simplest features are the most informative: the small set of keywords learned from the training partition (which include common section labels like *preface*, *content*, *index*, *advertisement*, other informative markers such as *dedicated*, *copyright*, and currency markers like $, £), the position of the page in the book, and the density of characters (including words and numbers) on the page.

4.2 Content/non-content segment labeling

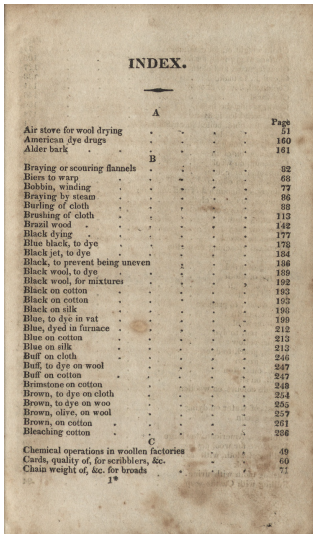
In order to assess the ability of these different features and models to demarcate the core text of a work, we binarize the multiclass label, assigning TEXT to all pages labeled TEXT in the multiclass setting, and NON-TEXT to all other pages. We train all three classifiers again on these binarized labels and repeat the training procedure for each model outlined in section 4.

Table 3 shows the results for the binary task of {TEXT, NOT-TEXT} prediction. Here again we contextualize these results with two simpler baselines: a simple majority class predictor (always predict TEXT), and a model that identifies the average start and end positions in a book for the first and last text page (respectively) within the training data, and predicts TEXT for pages within that range (roughly within the [0.10, 0.94] interval), and NON-TEXT for all pages outside of it. This corresponds to a heuristic that chops off the first 10% of a book and the last 6% as NON-TEXT.

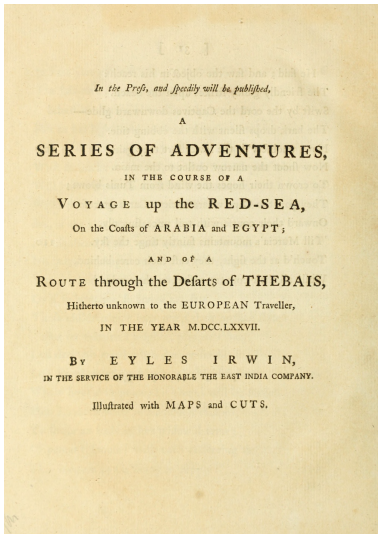
The chop heuristic performs worse than the majority class predictor in terms of absolute accuracy, but improves over the class-balanced macro scores. All three feature-rich models show substantial improvements over all metrics, but are indistinguishable from each other, each achieving nearly identical performance. For this reduced purpose, any of the three classifiers are sufficient for segmenting TEXT from NON-TEXT, even a random forest making independent predictions for each page.

5 Related work

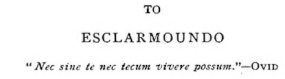
The work described here has points of intersection with several other threads of research. The most direct originates in work that grows out of the INEX and ICDAR book structure extraction com-



(a) Index/TOC.



(b) Ad predicted as text.



(c) Dedication predicted as title.

Figure 4: Exemplary mistakes in prediction.

petitions (Kazai et al., 2009, 2010; Doucet et al., 2011, 2013), in which participants are challenged to recognize the fine-grained structure present in documents (recognizing, for example, that the current article has sections entitled "Abstract," "Introduction," "Data, "References," etc.). The most successful systems recognize structure by parsing the table of contents (Déjean and Meunier, 2005, 2009; Wu et al., 2013; Gao et al., 2009) rather than relying on the content of the book itself. Our work primarily differs in the fundamental design choice of prescribing a fix set of categories into which we classify pages (in order to enable comparison across documents) rather than prioritizing the idiosynractic structure of a book (which is useful for generating new tables of contents).

Given the relatively constant page-level categories that printers use to describe book design, we formulate our problem as a classification task into a set of pre-established categories. An alternative is to take an unsupervised approach, and learn the set of categories empirically from the data; this general problem of book segmentation in its unlabeled form shares functional similarity with other work in general unsupervised topic or discourse segmentation (Hearst, 1997; Utiyama and Isahara, 2001; Chen et al., 2009)—most notably, the work of Eisenstein and Barzilay (2008) (for whom the section labels may be considered a form of "cue words" akin to discourse markers). Given the amount of data in large-scale book collections, we see this as an interesting path forward (either in a fully unsupervised or semi-supervised setting); an unsupervised approach that includes aspects of metadata such as country of publication or publisher may also be fruitful in accommodating variation in printer's rules as a function of time and geographical location (books by French publishers, for example, often place the table of contents at the back of the book).

As figure 2 illustrates, one of the primary challenges that we face with the labeled segmentation of books is the degraded nature of the input; unlike contemporary business documents for which OCR is largely a solved problem, historical books present several challenges due to their binding, age, and significant variation in font and printing. Much work has focused on overcoming these limitations from several perspectives, including creating ground truth for historical books (Papadopoulos et al., 2013), bootstrapping their alignment with existing resources (Feng and Manmatha, 2006; Yalniz and Manmatha, 2011), exploiting the fact that books often have multiple scans or printings that could be leveraged (Smith et al., 2011; Wemhoener et al., 2013) or developing methods that account for variation in the printing process (Berg-Kirkpatrick et al., 2013; Berg-Kirkpatrick and Klein, 2014).

Large-scale book corpora are increasingly being used as the raw material for linguistic analysis, especially those focused on measuring historical change (Hamilton et al., 2016a,b; Kulkarni et al., 2015; Mitra et al., 2014; Mihalcea and Nastase, 2012; Kim et al., 2014). These studies use not only the observed word frequencies pro-

vided by the Google Ngram dataset, but also increasingly structured representations of language as well (Lin et al., 2012; Goldberg and Orwant, 2013). The task of labeled book segmentation may be helpful in reducing the noise inherent in the use of statistics aggregated from these large datasets—both in terms of filtering out the 11% of all pages that are not the core content (e.g., such as indices), and also in grounding the text at the appropriate date for historical analysis (such as deriving statistics only from the core content, and not from an introduction written years afterward).

6 Conclusion

We introduce in this work the task of *book structure labeling*, the problem of assigning to each page in a printed book its membership in one of a set of predetermined categories. In annotating a large dataset of books, we are able to empirically assess the ability to accurately segment and label books from a range of historical time periods.

The ten categories that form our typology are drawn from printers' guides and informed by contemporary criticism, but still reflect our historical present; while we have in part let our categories be shaped by our experience labeling texts (so that we have preserved in our annotations historical categories not in contemporary use, such as SUBSCRIBERS), we recognize that the act of categorization glosses over meaningful distinctions—for example, while we have grouped sections marked ADVERTISEMENT, TO THE READER, PREFACE, INTRODUCTION, FOREWORD and others into the single category of PREFACE, such labels may have historically significant differences that may be worth preserving for some analyses. Nevertheless, we expect the coarse distinctions we outline here to occasion research that requires access to those broad categories. Potential uses of this work include using the categories directly to answer questions in book history (e.g., charting the historical prevalence of advertisements and their variation across time), improving the task of idiosyncratic structure detection by identifying tables of contents, and identifying the fine-grained topics of books by parsing recognized indices.

In this work, we deliberately focus on the resource-deficient scenario most commonly encountered by researchers working with large book corpora, in which books are represented as the output of errorful OCR. In providing a labeled dataset for others to use, we hope to encourage other work that reasons about the structure present in alternative representations (such as images) as well.

Acknowledgments

Many thanks to the anonymous reviewers and Hannah Alpert-Abrams and for their valuable feedback, and to the HathiTrust Research Center for their assistance in enabling this work. The research reported in this article was supported by a grant from the Digital Humanities at Berkeley initiative and resources provided by NVIDIA.

References

Jinho Baik, Percy Deift, and Kurt Johansson. 1999. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 4.

David Bamman, Michelle Carney, Jon Gillick, Cody Hennesy, and Vijitha Sridhar. 2017. Estimating the date of first publication in a large-scale digital library. In *Proceedings of the ACM/IEEE Annual Joint Conference on Digital Libraries*.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217, Sofia, Bulgaria. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick and Dan Klein. 2014. Improved typesetting models for historical OCR. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 118–123, Baltimore, Maryland. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Mach. Learn.*, 45(1):5–32.

Harr Chen, S. R. K. Branavan, Regina Barzilay, and David R. Karger. 2009. Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 371–379, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hervé Déjean and Jean-Luc Meunier. 2005. Structuring documents according to their table of contents. In *Proceedings of the 2005 ACM Symposium on Document Engineering*, DocEng '05, pages 2–9, New York, NY, USA. ACM.

Hervé Déjean and Jean-Luc Meunier. 2009. On tables of contents and how to recognize them. *International Journal of Document Analysis and Recognition (IJDAR)*, 12(1):1–20.

Antoine Doucet, Gabriella Kazai, and Jean-Luc Meunier. 2011. ICDAR 2011 book structure extraction competition. In *ICDAR*.

Antoine Doucet, Gabriella Kazai, and Jean-Luc Meunier. 2013. Overview of the ICDAR 2013 competition on book structure extraction.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 334–343, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shaolei Feng and R Manmatha. 2006. A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 109–118. ACM.

L. Gao, Z. Tang, X. Lin, X. Tao, and Y. Chu. 2009. Analysis of book documents' table of content based on clustering. In *2009 10th International Conference on Document Analysis and Recognition*, pages 911–915.

Gérard Genette. 1987. *Paratexts: Thresholds of Interpretation*. Cambridge University Press.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, volume 1, pages 241–247.

Alex Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational models of semantic change. In *Proceedings of EMNLP*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Marti A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64.

Gabriella Kazai, Antoine Doucet, Marijn Koolen, and Monica Landoni. 2009. Overview of the inex 2009 book track. In *INEX*.

Gabriella Kazai, Antoine Doucet, Marijn Koolen, and Monica Landoni. 2010. *Overview of the INEX 2009 Book Track*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Matthew Kirschenbaum and Sarah Werner. 2014. Digital scholarship and digital studies: The state of the discipline. *Book History*, 17.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 625–635, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Marshall Lee. 2009. *Bookmaking: Editing, Design, Production (Third Edition)*. W. W. Norton & Company.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 259–263. Association for Computational Linguistics.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, Baltimore, Maryland. Association for Computational Linguistics.

Christos Papadopoulos, Stefan Pletschacher, Christian Clausner, and Apostolos Antonacopoulos. 2013. The IMPACT dataset of historical document images. In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, HIP '13, pages 123–130, New York, NY, USA. ACM.

David A. Smith, R. Manmatha, and James Allan. 2011. Mining relational structure from millions of books: Position paper. In *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing*, BooksOnline '11, pages 49–54, New York, NY, USA. ACM.

Ted Underwood. 2016. The life cycles of genres. *Cultural Analytics*.

Ted Underwood. 2017. Datamunging Github repository. https://github.com/tedunderwood/DataMunging.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 499–506, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Wemhoener, Ismet Zeki Yalniz, and R Manmatha. 2013. Creating an improved version using noisy OCR from multiple editions. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 160–164. IEEE.

Sarah Werner. 2012. Where material book culture meets digital humanities.

Matthew Wilkens. 2013. The geographic imagination of Civil War-era American fiction. *American Literary History*, 25(4):803–840.

Adrian Wilson. 1993. *Design of Books*. Chronicle Books.

Z. Wu, P. Mitra, and C. L. Giles. 2013. Table of contents recognition and extraction for heterogeneous book documents. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1205–1209.

Ismet Zeki Yalniz and Raghavan Manmatha. 2011. A fast alignment scheme for automatic OCR evaluation of books. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 754–758. IEEE.