ICDAR 2009 Book Structure Extraction Competition

Antoine Doucet
University of Caen, France
doucet@info.unicaen.fr

Gabriella Kazai
Microsoft Research, Cambridge, UK
gabkaz@microsoft.com

Bodin Dresevic, Aleksandar Uzelac, Bogdan Radakovic, and Nikola Todic
Microsoft Development Center Serbia
{bodind,aleksandar.uzelac,bogdan.radakovic,nikola.todic}@microsoft.com

Abstract

This paper introduces the Book Structure Extraction competition run at ICDAR 2009. The goal of the competition is to evaluate and compare automatic techniques for deriving structure information from digitized books, which could then be used to aid navigation inside the books. More specifically, the task that participants are faced with is to construct hyperlinked tables of contents for a collection of 1,000 digitized books. This paper describes the setup of the competition, the book collection used in the task, and the proposed measures for the evaluation. Results of the evaluation will be presented at the ICDAR 2009 conference and will be published in the INEX 2009 proceedings.

1. Introduction

Mass-digitization projects, such as the Million Book project[1], efforts of the Open Content Alliance[2], and the digitization work of Google[3], are converting whole libraries by digitizing books on an industrial scale [1]. The process involves the efficient photographing of books, page-by-page, and the conversion of each page image into searchable text through the use of optical character recognition (OCR) software.

Current digitization and OCR technologies typically produce the full text of digitized books with only minimal structure information. Pages and paragraphs are usually identified and marked up in the OCR, but more sophisticated structures, such as chapters, sections, etc., are currently not recognized. In order to enable systems to provide users with richer browsing experiences, it is necessary to make available such additional structures, for example in the form of XML markup embedded in the full text of the digitized books.

The Book Structure Extraction competition aims to address this need by promoting research into automatic structure recognition and extraction techniques that could complement or enhance current OCR methods and lead to the availability of rich structure information for digitized books. Such structure information can then be used to aid user navigation inside books as well as to improve search performance [9].

As the competition is still running at the time of writing, this paper only describes the setup of the competition; the results will be published separately once the competition concluded.

The paper is structured as follows. We start by placing the competition in the context of the work conducted at the INEX evaluation forum (Section 2). In Section 3, we describe the setup of the competition, including its goals and the task that has been set for its participants. The book collection used in the task is detailed in Section 4, while the proposed measures for the evaluation of the participating systems' performance is described in Section 5. We conclude with a summary of the competition and our future plans in Section 6.

2. Background

Motivated by the need to foster research in areas relating to large digital book repositories, see e.g., [4], the Book Track was launched in 2007 as part of the Initiative for the Evaluation of XML retrieval (INEX)[4]. INEX was chosen as a suitable forum, as searching for information in a collection of books can be seen as one of the natural application areas of focused retrieval approaches [3], which have been investigated at INEX since 2002 [2]. In particular, focused retrieval over books presents a clear benefit to users,

[1] http://www.ulib.org/
[2] www.opencontentalliance.org/
[3] http://books.google.com/
[4] http://www.inex.cs.otago.ac.nz/

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.280

enabling them to gain direct access to parts of books (of potentially hundreds of pages in length) that are relevant to their information need.

The overall goal of the INEX Book Track is to promote inter-disciplinary research investigating techniques for supporting users in reading, searching, and navigating the full texts of digitized books and to provide a forum for the exchange of research ideas and contributions. In 2007, the track focused on information retrieval (IR) tasks [5]. In 2008, two new tasks were introduced, including the book structure extraction task [6]. The structure extraction task was set up with the aim to evaluate automatic techniques for deriving structure from the OCR texts and page images of digitized books. The first round of the structure extraction task, in 2008, permitted to set up appropriate evaluation infrastructure, including guidelines, tools to generate groundtruth data, evaluation measures, and a test set of 100 books. The second round is currently being run both at INEX 2009 and at the International Conference on Document Analysis and Recognition (ICDAR) 2009. This round builds on the established infrastructure and an extended test set of 1,000 digitized books.

```
<!ELEMENT bs-submission
  (source-files, description, book+)>
<!ATTLIST bs-submission
  participant-id  CDATA  #REQUIRED
  run-id  CDATA  #REQUIRED
  task  (book-toc) #REQUIRED
  toc-creation (automatic |
    semi-automatic) #REQUIRED
  toc-source (book-toc | no-book-toc |
    full-content | other) #REQUIRED>
<!ELEMENT source-files EMPTY>
<!ATTLIST source-files
  xml  (yes|no) #REQUIRED
  pdf  (yes|no) #REQUIRED>
<!ELEMENT description (#PCDATA)>
<!ELEMENT book (bookid, toc-entry+)>
<!ELEMENT bookid (#PCDATA)>
<!ELEMENT toc-entry(toc-entry*)>
<!ATTLIST toc-entry
  title (#PCDATA) #REQUIRED
  page  (#PCDATA) #REQUIRED>
```

Figure 1. DTD of the XML output ("run") that participating systems are expected to submit to the competition, containing the generated hyperlinked ToC for each book in the test set.

3 Competition Setup

3.1. Goals

The goal of the structure extraction competition at ICDAR 2009 is to test and compare automatic techniques for deriving structural information from digitized books in order to build hyperlinked tables of contents (ToC) that could then be used to navigate inside the books.

Example research questions whose exploration is facilitated by this competition include, but are not limited to:

- Can a ToC be extracted from the pages of a book that contain the actual printed ToC (where available) or could it be generated more reliably from the full content of the book?
- Can a ToC be extracted only from textual information or is page layout information necessary?
- What techniques provide reliable logical page number recognition and extraction and how logical page numbers can be mapped to physical page numbers?

3.2 Task Description

Given the OCR text and the PDF of a sample set of 1,000 digitized books of different genre and style, the task is to build hyperlinked tables of contents for each book in the test set. The OCR text of each book is stored in DjVu XML format (see Section 4). Participants may employ any techniques and can make use of either or both the OCR text and the PDF images to derive the necessary structure information and generate the ToCs.

Participating systems are expected to output an XML file (referred to as a "run") that contains the generated hyperlinked ToC for each book in the test set. The document type definition (DTD) for the XML output is given in Figure 1.

Participants are invited to submit up to 10 runs, each run containing the ToC for all 1,000 books.

The ToCs created by participants will be compared to a manually built groundtruth; see Section 5 for details on the annotation process and the proposed evaluation measures.

3.3. Participating Organizations

Following the call for participation issued in April 2009, 11 organizations registered, as of 15 May 2009, see Table 1.

Several further organizations have expressed interest but renounced participation due to time constraints. Some of them, however, offered to contribute to the groundtruth creation this year, while preparing their systems for the forthcoming rounds of the competition.

Dublin City University
Exalead Inc.
Fraunhofer Institute
Microsoft Development Center Serbia
Noopsis Inc.
Peking University
Texas A&M
University of Amsterdam
University of Caen
University of Firenze
Xerox Research Centre Europe

Table 1. List of registered participants (as of 15 May 2009)

4. Book Collection

The corpus of the INEX book track contains a collection of 50,239 digitized out-of-copyright books, provided by Microsoft Live Search and the Internet Archive [6].

The set of books used in the book structure extraction competition comprises 1,000 books selected from the INEX book corpus. It contains books of different genre, among which history books, biographies, literary studies, religious texts and teachings, reference works, encyclopedias, essays, proceedings, novels, and poetry.

To facilitate the separate evaluation of techniques based on the analysis of book pages that contain the printed ToC versus techniques that are based on deriving structure information from the full book content, we selected 200 books into the total 1,000 that do not contain a printed ToC. To do this, we used a tool developed by Microsoft Development Center Serbia, which converts the DjVu XML OCR text into BookML, a format in which ToC pages are explicitly marked up. We then selected a set of 800 books with detected ToC pages, and a set of 200 books without detected ToC pages into the full test set of 1,000 books. We note that this ratio of 80:20% of books with and without printed ToCs is proportional to that observed over the whole INEX corpus of 50,239 books.

The uncompressed size of the structure extraction corpus is around 33GB.

Each book is provided in two different formats: portable document format (PDF), and DjVu XML containing the OCR text and basic structure markup as illustrated below:

```
<DjVuXML>
 <BODY>
  <OBJECT data="file..." [...]>
   <PARAM name="PAGE" value="[...]">
   [...]
   <REGION>
    <PARAGRAPH>
     <LINE>
      <WORD coords="[...]"> Moby </WORD>
      <WORD coords="[...]"> Dick </WORD>
      <WORD coords="[...]"> Herman </WORD>
      <WORD coords="[...]"> Melville </WORD>
      [...]
     </LINE>
     [...]
    </PARAGRAPH>
   </REGION>
   [...]
  </OBJECT>
  [...]
 </BODY>
</DjVuXML>
```

An <OBJECT> element corresponds to a page in a digitized book. A page counter, corresponding to the physical page number, is embedded in the @value attribute of the <PARAM> element, which has the @name="PAGE" attribute. The logical page numbers (as printed inside the book) can be found (not always) in the header or the footer part of a page. Note, however, that headers/footers are not explicitly recognized in the OCR, i.e., the first paragraph on a page may be a header and the last one or more paragraphs may be part of a footer. Depending on the book, headers may include chapter/section titles and logical page numbers (although due to OCR error, the page number is not always present).

Inside a page, each paragraph is marked up. It should be noted that an actual paragraph that starts on one page and ends on the next is marked up as two separate paragraphs within two page elements. Each paragraph element consists of line elements, within which each word is marked up separately. Coordinates that correspond to the four points of a rectangle surrounding a word are given as attributes of word elements.

5. Evaluation

The ToCs created by participants will be evaluated by comparing them to a manually built groundtruth. This section describes the groundtruth annotation process and the proposed evaluation measures.

5.1. Annotation Process

The process of manually building the ToC of a book is very time-consuming. Hence, to make the creation of the groundtruth for 1,000 digitized books feasible, we resorted to 1) facilitating the annotation task with a dedicated tool, 2) making use of a baseline annotation as starting point and employing human annotators to make corrections, and 3) sharing the workload.

An annotation tool was specifically designed for this purpose and developed at the University of Caen. The tool takes as input a generated ToC and allows annotators to manually correct any mistakes. A screen capture of the tool is shown in Figure 2. In the application window, the right-hand side displays the baseline ToC with clickable (and editable) links. The left-hand side shows the current page and allows to navigate through the book. The JPEG image of each visited page is downloaded from the INEX server at `www.booksearch.org.uk` and is locally cached to limit bandwidth usage.

An important side-effect of making use of a baseline ToC is that this may trigger a bias in the groundtruth, since annotators may be influenced by the ToC presented to them. To reduce this bias (or rather, to spread it among participating organizations), we need to ensure that the baseline ToCs are contributed equally from all participating organizations.

Finally, as we already mentioned, the annotation effort will be shared among all participants. The created groundtruth will then be made available to all contributing participants for use in future evaluations. The minimum required contribution is to annotate 50 books (10 of which will be books without a printed ToC).

5.2. Measures

As we mentioned before, the extracted ToCs will be evaluated by comparing them to the groundtruth. This, first requires the definition of a number of basic concepts:

Definitions. We define the atomic units that make up a ToC as ToC Entries. A ToC Entry has the following three properties: *Title*, *Link*, and *Depth Level*. For example, given a ToC entry corresponding to a book chapter, its *Title* is the chapter title, its *Link* is the physical page number at which the chapter starts in the book, and its *Depth Level* is the depth at which the chapter is found in the ToC tree, where the book represents the root.

Given the above definitions, the task of comparing two ToCs (i.e., comparing a generated ToC to one in the groundtruth) can be reduced to matching the titles, links and depth levels of each ToC entry. This is, however, not a trivial task as we explain next.

Matching Titles. A ToC title may take several forms and it may only contain, e.g., the actual title of a chapter, such as "His Birth and First Years", or it may also include the chapter number as in "3. His Birth and First Years" or even the word "chapter" as in "Chapter 3. His Birth and First Years". In addition, the title that is used in the printed ToC may differ from the title which then appears in the book content. It is difficult to differentiate between the different answers as all of them are in fact correct titles for a ToC entry.

Thus, to take into account the fact that many similar answers may be correct, we adopt vague title matching in the evaluation. We say that two titles match if they are "sufficiently similar", where similarity is measured based on a modified version of the Levenshtein algorithm (where the cost of alphanumeric substitution, deletion and insertion is 10, and the cost of non-alphanumeric substitution, deletion and insertion remains 1) [7]:

Two strings A and B are "sufficiently similar" if

$$D = \frac{LevenshteinDist * 10}{Minlength(A), length(B)}$$

is less than 20% and if the distance between their first and last five characters (or less if the string is shorter) is less than 60%.

Matching Links. A link is said to be correctly recognized if there is an entry with matching title linking to the same physical page in the groundtruth.

Matching Depth levels. A depth level is said to be correct if there is an entry with matching title at the same depth level in the groundtruth.

Matching complete ToC entries. A ToC entry is entirely correct if there is an entry with matching title and same depth level, linking to the same physical page in the groundtruth.

Measures. For a given book ToC, we can then calculate precision and recall measures [8] for each property separately, and for complete entries. Precision is defined as the ratio of the total number of correctly recognized ToC entries and the total number of ToC entries in a generated ToC; and recall as the ratio of the total number of correctly recognized ToC entries and the total number of ToC entries in the groundtruth. The F-measure is then calculated as the harmonic mean of precision and recall.

These measures will be averaged over the two subsets of the 1,000 books (see Section 4, as well as the entire test set to calculate overall performance. The two subsets, comprising of 200 and 800 books, respectively, that do and do not have a printed ToC, will allow us to compare the effectiveness of techniques that do or do not rely on the presence of printed ToC pages in a book. The main measure for result presentation will be the F-measure, reported for complete entries. Participants are, however, welcome to propose alternative measures.

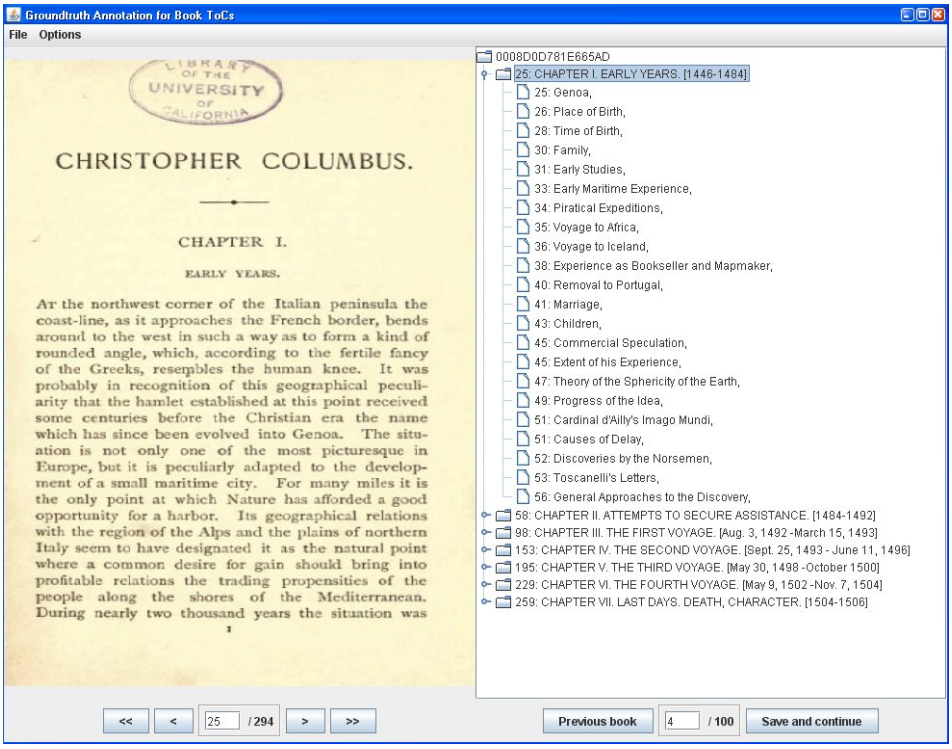

Figure 2. A screen shot of the groundtruth annotation tool

6. Summary and future plans

At the time of writing, the test collection has been distributed to the registered participants. They are expected to submit the automatically generated ToCs for the test set by the end of June, after which the creation of the groundtruth data set will commence. The results of the competition will be announced at the ICDAR conference, 26-29 July 2009, in Barcelona, Spain. The results will also be made available online[5], and will also be published as part of the proceedings of the INEX 2009 Workshop, to be held in December 2009 in Brisbane, Australia. Participants are invited to submit a paper for publication at INEX 2009, describing the approach they have taken in solving the task.

In future years, we aim to investigate the usability of the extracted ToCs. In particular, we will explore the use of qualitative evaluation measures in addition to the current precision/recall measures. This would enable us to better understand what properties make a ToC useful and which are important to users engaged in reading or searching. Insights as such are expected to contribute to future research into providing better navigational aids to users of digital book repositories.

[5] http://users.info.unicaen.fr/~doucet/StructureExtraction2009/

References

[1] K. Coyle. Mass digitization of books. *Journal of Academic Librarianship*, 32(6):641–645, 2006.

[2] S. Geva, J. Kamps, and A. Trotman, editors. *Advances in Focused Retrieval: 7th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2008)*, volume 5613 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, Heidelberg, 2009.

[3] J. Kamps, S. Geva, and A. Trotman. Report on the SIGIR 2008 workshop on focused retrieval. *SIGIR Forum*, 42(2):59–65, 2008.

[4] P. Kantor, G. Kazai, N. Milic-Frayling, and R. Wilkinson, editors. *BooksOnline '08: Proceeding of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories*, New York, NY, USA, 2008. ACM.

[5] G. Kazai and A. Doucet. Overview of the INEX 2007 Book Search Track (BookSearch'07). *ACM SIGIR Forum*, 42(1):2–15, 2008.

[6] G. Kazai, A. Doucet, and M. Landoni. Overview of the INEX 2008 Book Track. In Geva et al. [2].

[7] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707+, February 1966.

[8] C. J. van Rijsbergen. *Information retrieval*. Butterworths, London, 2 edition, 1979.

[9] R. van Zwol and T. van Loosbroek. Effective use of semantic structure in XML retrieval. In G. Amati, C. Carpineto, and G. Romano, editors, *ECIR*, volume 4425 of *Lecture Notes in Computer Science*, pages 621–628. Springer, 2007.