GROTOAP: GROund Truth for Open Access Publications

Dominika Tkaczyk
d.tkaczyk@icm.edu.pl

Artur Czeczko
a.czeczko@icm.edu.pl

Krzysztof Rusek
kr292291@students.mimuw.edu.pl

Łukasz Bolikowski
l.bolikowski@icm.edu.pl

Roman Bogacewicz
r.bogacewicz@icm.edu.pl

Centre for Open Science, Interdisciplinary Centre for Mathematical and Computational Modelling, Univ. of Warsaw
ul. Prosta 69, 00-838 Warszawa, Poland

ABSTRACT

The field of digital document content analysis includes many important tasks, for example page segmentation or zone classification. It is impossible to build effective solutions for such problems and evaluate their performance without a reliable test set, that contains both input documents and expected results of segmentation and classification. In this paper we present GROTOAP — a test set useful for training and performance evaluation of page segmentation and zone classification tasks. The test set contains input articles in a digital form and corresponding ground truth files. All input documents included in the test set have been selected from DOAJ database, which indexes articles published under CC-BY license. The whole test set is available under the same license.

Categories and Subject Descriptors

D.2.10 [**Software Engineering**]: Design—*Quality analysis and evaluation*; I.7.5 [**Document and Text Processing**]: Document Capture—*Document analysis*

General Terms

Performance

Keywords

document analysis, page segmentation, zone classification, system evaluation, ground truth

1. INTRODUCTION

Page segmentation and zone classification tasks play very important role in document analysis. The goal of page segmentation is to build a geometric hierarchical structure of the document by dividing the document's pages into zones, lines and words. Classifying zones means finding their function in the document and assigning corresponding labels (eg. title, authors, affiliation) to them. An efficient page segmentation or zone classification solution needs to be carefully evaluated, which requires a reliable test set containing multiple examples of input documents and expected results of segmentation and classification.

JCDL'12, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

In this paper we present GROTOAP — a test set useful for adapting, training and performance evaluation of document analysis-related solutions, such as page segmentation and zone classification. The test set is built upon a group of scientific articles from Directory of Open Access Journals (DOAJ) database [1] and contains original publications in PDF format, ground truth files holding hierarchical structure and zone labels extracted from publications, and finally the SegmEdit tool useful for editing the ground truth files. The whole set is distributed under the CC-BY license in the Open Access model and can be downloaded from *http://ceon.pl/en/research/solutions/*.

Existing test sets containing ground truth data useful for page segmentation or zone classification are usually based on scanned document images instead of born-digital documents. For example UW-III [4] contains various document images along with structure-related ground truth information. Unfortunately UW-III is not free and difficult to purchase. MARG [2] is a data set containing scanned pages from biomedical journals. The main problem is that it contains only the first pages of documents and only a small subset of zones is included, and as a result its usability is very limited. Other data sets built upon scanned document images of various layouts and corresponding ground truth data are: PRImA [5], MediaTeam Oulu Document Database [7], UvA [3] and Tobacco800 [6].

GROTOAP test set was built as part of the implementation of the metadata extraction process [8] for born-digital documents and has been successfully used for training and performance evaluation of the process and its individual steps.

2. GROTOAP TEST SET

GROTOAP test set consists of three parts:

- input documents in PDF format, that contain scientific articles from DOAJ database,
- ground truth files in XML format, that contain the geometric hierarchical structure of input documents along with their zone labels,
- SegmEdit tool, useful for editing ground truth files.

To make the test set useful for performance evaluation of page segmentation and zone classification tasks, the ground truth files store their typical output: the hierarchical structure holding all the pages, zones, lines, words and characters

of the document, and also zone labels. We have used TrueViz format for storing all ground truth data.

The process of creating the test set required a manual phase of correcting the structure of the documents stored in ground truth files. To make this as easy as possible, we have created SegmEdit, a tool for editing TrueViz files. SegmEdit consists of two parts: SegmEditGUI and SegmServer. SegmEditGUI is a desktop program that allows to view, edit and save TrueViz files. It may be used as a standalone application, in such case PDFs and ground truth files are stored and saved locally. An alternative is to use SegmServer, a HTTP server that is able to store all files and serve them to SegmEditGUI. In such case the server communicates with multiple SegmEditGUI instances over the network sending files and receiving them partly or completely processed, which makes editing ground truth files by multiple users very easy.

The input documents and ground truth files are published in Open Access model under the CC-BY license. The source code of SegmEdit tool is published under the GPL v3 license.

3. BUILDING THE TEST SET

The process of creating GROTOAP test set was semi-automatic. It consisted of three main phases (the second and third phase were performed twice):

1. selecting and downloading a set of publications based on metadata from DOAJ database,
2. automatic extraction of hierarchical structure using metadata extraction tools,
3. manual correction of the results of automatic structure extraction with the use of SegmEdit tool.

All the publications in the test set have been taken from journals distinguished with SPARC Europe Seal for Open Access Journals. We have harvested Directory of Open Access Journals for basic metadata of all articles from said journals, including links to full texts. Next, we have pseudo-randomly selected articles to be downloaded and included in two test groups. The first group contains one article from each four journals published by the same publisher, 113 articles in total. Articles published by the same publisher have usually very similar layout, and as a result the layout distribution in the first test group is similar to the layout distribution in the entire DOAJ database. The second group consists of 115 articles and has been compiled by randomly choosing one article from each publisher, as a result it contains most layouts that appear in the DOAJ database. Currently only the first group is included in the test set. The second group will be processed and added to the test set in the near future.

To minimize the time needed for manual correction, first we processed selected publications using tools implemented in our metadata extraction process [8] in order to automatically extract their structure. Currently the process requires a document in PDF format on the input. First individual characters and their bounding boxes were extracted using iText library, then a modified Docstrum algorithm was used to segment pages. Finally, a classifier based on Hidden Markov Models found labels for all zones of the document.

During the last phase TrueViz files created automatically were corrected by a group of people with the use of SegmEdit tool. We used one SegmServer instance to store and serve PDF and ground truth files to multiple SegmEditGUI instances run on client computers. As a result of choosing this architecture we did not have to manually distribute files to a group of people and gather the results together. The correction phase included verifying words, lines and zones generated by metadata extraction tools, splitting incorrectly merged objects and merging incorrectly split ones, verifying and correcting labels assigned to zones. To minimize human errors we performed an additional checking phase, which included an inspection and approval by an independent judge.

4. CONCLUSIONS

We have presented GROTOAP — a test set useful for training and evaluation of content analysis tasks like page segmentation and zone classification. We have described in details the contents of the test set and the process of creating it. Two main features distinguishing GROTOAP from earlier efforts are:

- usefulness in testing algorithms optimized for processing born-digital content,
- and reliance on Open Access publications which guarantees easy distribution of both original material and derived ground truth data.

5. ACKNOWLEDGMENTS

The work is supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the Strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

We would like to thank a number of people involved in the process of manual correction of the ground truth files. We would also like to thank Klaudia Grabowska for her help in preparing information about publishing documents in the Open Access model.

6. REFERENCES

[1] DOAJ. http://www.doaj.org/.
[2] MARG. http://marg.nlm.nih.gov/.
[3] UvA. http://www.science.uva.nl/UvA-CDD/.
[4] UW-III. http://www.science.uva.nl/research/dlia/datasets/uwash3.html.
[5] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher. A Realistic Dataset for Performance Evaluation of Document Layout Analysis. *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300, 2009.
[6] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a Test Collection for Complex Document Information Processing. In *Proc. 29th Annual Int. ACM SIGIR Conference*, pages 665–666, 2006.
[7] J. Sauvola and H. Kauniskangas. MediaTeam Document Database II, a CD-ROM collection of document images, 1999.
[8] D. Tkaczyk, L. Bolikowski, A. Czeczko, and K. Rusek. A modular metadata extraction system for born-digital articles. In *Proceedings of the 10th IAPR International Workshop on Document Analysis Systems*, 2012.