GROTOAP2 — The methodology of creating a large ground truth dataset of scientific articles

Dominika Tkaczyk
d.tkaczyk@icm.edu.pl

Pawel Szostek
pawel.szostek@gmail.com

Lukasz Bolikowski
l.bolikowski@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

Scientific literature analysis improves knowledge propagation and plays a key role in understanding and assessment of scholarly communication in scientific world. In the recent years many tools and services for analysing the content of scientific articles have been developed. One of the most important tasks in this research area is understanding the role of different parts of the document. It is impossible to build effective solutions for problems related to document fragments classification and evaluate their performance without a reliable test set, that contains both input documents and the expected results of classification. In this paper we describe GROTOAP2 — a large dataset of ground truth files built from Open Access Subset of PubMed Central. GROTOAP2 is useful for learning and evaluation of document content analysis-related solutions, such as document zone classification. GROTOAP2 is published under Open Access license and is available at http://cermine.ceon.pl/grotoap2/. The article presents the content of GROTOAP2 and describes the automatic process used to create the dataset.

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 content analysis, zone classification, system evaluation, ground truth dataset

1. INTRODUCTION

The analysis of scientific literature accelerates spreading ideas and knowledge and plays a key role in many areas of research, such as understanding scholarly communication, the assessment of scientific results, identifying important research centers or finding new interesting research possibilities. Scientific literature analysis supports a number of tasks related to metadata and information extraction, scientific data organizing, providing intelligent search tools, finding similar and related documents, building citation networks, and many more. One of the most important tasks in this research area is understanding the roles of different fragments of documents, which is often referred to as document zone or block classification. An efficient zone classification solution needs to be carefully evaluated, which requires a reliable test set containing multiple examples of input documents and the expected results of classification.

In this paper we present GROTOAP2 — GROund Truth for Open Access Publications. GROTOAP2 is a large test set built upon a group of scientific articles from PubMed Central Open Access Subset [2]. The test set contains 3,076 ground truth files holding hierarchical structure of the documents' content along with zone labels. The corresponding source PDF files can be obtained online. GROTOAP2 is distributed under the CC-BY license in the Open Access model and can be downloaded from http://cermine.ceon.pl/grotoap2/.

GROTOAP2 test set is very useful for adapting, training and performance evaluation of document analysis-related solutions, such as zone classification. The test set was built as a part of the implementation of CERMINE [12] — a comprehensive open source system for extracting metadata and parsed bibliographic references from scientific articles in born-digital form. It has been successfully used for training and performance evaluation of CERMINE's extraction process and its two zone classifiers.

In the rest of the paper we present the content of GROTOAP2, compare our solution to existing ones and discuss its advantages and drawbacks. We also describe in detail the automatic process of creating the dataset.

2. PREVIOUS WORK

Existing test sets containing ground truth data useful for 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 [1] is a dataset 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 for performance evaluation of page segmentation and zone classification is very limited. PRImA [5] dataset is also based on document images of various types and layouts, not only scientific papers. Other data sets built upon scanned document images of various layouts and corresponding ground truth data are: MediaTeam Oulu Document Database [10] (containing advertisements, articles, business cards, newsletters, street maps, etc.), UvA dataset [3] (containing scanned magazine pages) and Tobacco800 [8].

Héroux *et al.* [6] proposed an interesting approach for automatic ground truth generation for various document image analysis tasks. In this approach a test set is built for a specific evaluation task based on a derivation of document publishing software. An example implementation deriving the DocBook publishing framework is presented.

Sauvola *et al.* [9] present a distributed system for testing document analysis and understanding applications, that contains a collections of example document images. The system allows to create and manage test cases and execute and control tests of document analysis-related solutions.

GROTOAP2 is a successor of GROTOAP [11], semi-automatically created test set useful for training and performance evaluation of document analysis-related solutions. Unfortunately, since manual correction phase was a part of GROTOAP's creation process, the resulting test set is relatively small and every attempt to expand it is time-consuming and expensive.

Unlike previous examples, GROTOAP2 is a free, large dataset, based on open access born-digital documents. Thanks to the automatic method used to create GROTOAP2 the dataset can be easily expanded in the future.

3. GROTOAP2 TEST SET

GROTOAP2 is an automatically constructed dataset based on documents from PubMed Central Open Access Subset. The dataset contains:

- 3,076 ground-truth files in XML format holding the content of scholarly publications in hierarchical structured form,
- a list of URLs to corresponding scientific articles in PDF format,
- a bash script, that can be used to download PDF files from PMC repository.

The main part of GROTOAP2 are ground truth files built from scholarly articles in PDF form. A ground truth file contains a geometric hierarchical structure that holds the content of an article. In this representation an article consists of a list of pages, each page contains a list of zones, each zone contains a list of lines, each line contains a list of words, and finally each word contains a list of characters. Each structure element can be describes by its text content, position on the page and dimensions. The structure stored in a ground truth file contains also the natural reading order for all structure elements. Additionally labels describing the role in the document are assigned to zones.

The smallest elements in the structure are individual characters. A word is a continuous sequence of characters placed in one line with no spaces between them. Punctuation marks and typographical symbols can be separate words or parts of adjacent words, depending on the presence of spaces. Hyphenated words that are divided into two lines appear in the structure as two separate words that belong to different lines. A line is a sequence of words that forms a consistent fragment of the document's text. Words placed geometrically in the same line of the page, that are parts of neighbouring columns, do not belong to the same line. A zone is a consistent fragment of the document's text, geometrically separated from surrounding fragments and not divided into paragraphs or columns.

All documents in the test set have the Manhattan layout. All bounding boxes are defined by coordinates given in typographic points (1 typographic point equals to 1/72 of an inch). The origin of the coordinate system is the left upper corner of the page.

GROTOAP2 contains 3,076 documents with 27,049 pages and 437,320 zones in total, which gives the average of 8.79 pages per document and 16.17 zones per page. Every zone is labelled with one of 21 labels. Figure 1 lists the labels in the dataset and shows the fraction of documents containing a given label.

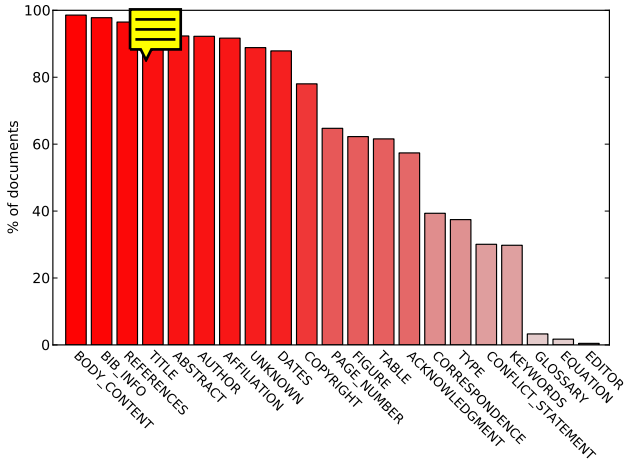

Figure 1: The dataset contains the following zone labels: *body_content*, *bib_info*, *references*, *title*, *abstract*, *author*, *affiliation*, *unknown*, *dates*, *copyright*, *page_number*, *figure*, *table*, *acknowledgment*, *correspondence*, *type*, *conflict_statement*, *keywords*, *glossary*, *equation*, *editor*. **The figure shows the fraction of the documents in GROTOAP2, that contain zones with a given label.**

We used TrueViz format [7] for storing ground truth files. TrueViz is an XML-based format that allows to store the geometrical and logical structure of the document, including pages, zones, lines, words, characters, their content and bounding boxes, and also zone labels and the order of the elements. XML format makes ground truth files easily readable by machines. Unfortunately, there is also a major caveat: ground truth files may grow to enormous size, much greater

than related PDF's. To limit the size of the dataset, only ground truth files are included in it. The corresponding PDF files can be easily downloaded using provided URL list or directly using bash script.

4. THE METHOD OF CREATING GROTOAP2

GROTOAP2 dataset was created automatically from PubMed Central resources:

1. First, a large set of files was downloaded from Open Access Subset of PubMed Central. We obtained both source articles in PDF format and corresponding NLM files containing metadata, full text and references.

2. PDF files were automatically processed and their hierarchical geometric structures along with the natural reading order was constructed.

3. The text content of every zone extracted previously was compared to labelled data from corresponding NLM files, which resulted in attaching the most probable labels to zones.

4. Files containing a lot of zones with unknown labels, that is zones for which the labelling process was unable to determine the label, were filtered out. From the remaining set the final data set was chosen randomly.

4.1 Gathering resources

Open Access Subset of Pubmed Central [2] is a rich source of scholarly publications and related metadata. It contains around 500,000 fulltext life sciences publications, as well as associated metadata in a form of NLM files. It is a relatively small part of PMC, it contains less than 2% of the collection of articles. Nevertheless, in many cases it is more than sufficient for performing machine learning tasks related to analysing scientific publications. It contains a broad variety of page layouts and topics.

We downloaded both source PDF files and associated metadata in a form of NLM files. NLM files contain a rich set of metadata of the document (title, authors, affiliations, keywords, abstract, journal/conference information, etc.), full text (sections, headers and paragraphs, tables and equations, etc.), and also parsed references of the document.

4.2 Ground-truth file generation

Downloaded PDF and NLM files were used to generate the geometric hierarchical structure of the files' content, which was stored using TrueViz format in ground truth files. The process was relatively time-consuming and so far we were able to process 92,501 files.

In the first phase of ground truth generation process we used automatic tools provided by CERMINE [12]. First, the individual characters were extracted from PDF files. Then, the characters were grouped into words, words into lines and finally lines into zones. After that reading order analysis was performed resulting in text elements at each hierarchy level being stored in the order reflecting how people read manuscripts.

In the second phase the text content of each zone extracted previously was matched against labelled fragments extracted from corresponding NLM files. We used Smith-Watermann distance and Cosine distance with thresholds to determine the most probable label for a zone. If the process was unable to find a match among labelled NLM fragments, the zone was labelled as UNKNOWN.

4.3 Filtering files

Data in NLM files vary greatly in quality from perfectly labelled down to containing no valuable information. Poor quality NLM result in sparsely labelled zones in generated TrueViz files, as the labelling process has no data to compare zone content to. Hence, it was necessary to filter documents whose zones are classified in satisfying measure. Figure 2 shows a histogram of documents with specified percentage of zones labeled with concrete classes. One can see that there are many documents (69.07%) having more than 80% of zones labelled.

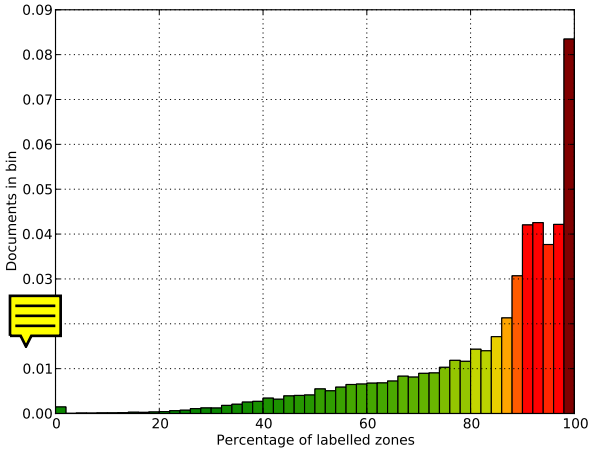

Figure 2: Histogram of documents in the PubMed-based dataset having given percentage of zones with an assigned class value.

We also wanted to be sure that the layout distributions in the whole processed set and the selected subset are similar. If poor quality metadata was associated with particular publishers or journals, choosing only highly covered documents could result in eliminating particular publishers and their layouts, which was to be avoided. We calculated the similarity of journal distributions of two sets using the following formula:

$$sim = \sum_{j \in J} min(d_A(j), d_B(j))$$

where J is the set of all journals in our document set and $d_A(j)$ and $d_B(j)$ are the percentage share of a given journal in sets A and B, respectively. The formula yields 1.0 for identical distributions, and 0.0 in the case of two sets, which do not share any journals. It turns out that the similarity of the whole processed set, and a subset of documents with at

Journal	Number of articles	Percentage
PLoS ONE	278	9.04%
Acta Crystallographica	242	7.87%
Nucleic Acids Research	81	2.63%
The Journal of Cell Biology	71	2.31%
BMC Public Health	54	1.76%
British Journal of Cancer	43	1.40%
The Journal of Experimental Medicine	42	1.37%
BMC Bioinformatics	41	1.33%
BMC Genomics	35	1.14%
Critical Care	31	1.01%

Table 1: The most popular journals included in GROTOAP2 test set.

least 80% labelled zones is 0.85, thus the distributions are indeed similar.

Due to the large size of ground truth files, we were not able to include all filtered documents in the final dataset. GROTOAP2 contains 3,076 documents chosen randomly from all documents with at least 80% of successfully labelled zones. The journal distribution similarity of the final set and the set processed originally is 0.68. The table 1 lists 10 most popular journals in GROTOAP2 dataset.

5. KNOWN PROBLEMS

The process of creating GROTOAP2 was entirely automatic. The lack of human experts supervision allowed us to create a large dataset, but also caused the following problems:

- segmentation errors resulting in incorrectly recognized zones and lines and their bounding boxes,
- labeling errors resulting in incorrect zone labels,
- the process was unable to assign a concrete label for 7% of all zones in the set.

6. CONCLUSIONS AND FUTURE WORK

We presented GROTOAP2 — a test set useful for training and evaluation of content analysis-related tasks like zone classification. We described in details the contents of the test set and the automatic process of creating it, we also discussed its advantages and drawbacks. The main features distinguishing GROTOAP2 from earlier efforts are:

- usefulness in testing algorithms optimized for processing born-digital content,
- reliance on Open Access publications which guarantees easy distribution of both original material and derived ground truth data,
- the possibility to expand the test set in the future thanks to the automatic generation method.

Our future plans include:

- generating a dataset of parsed bibliographic references by extracting reference strings from PDF files and labelling their fragments based on NLM data,
- assigning more specific body labels, especially for section headers of different levels.

7. 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".

8. REFERENCES

[1] MARG. http://marg.nlm.nih.gov/.
[2] PubMed. ftp://ftp.ncbi.nlm.nih.gov/pub/pmc.
[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] P. Heroux, E. Barbu, S. Adam, and E. Trupin. Automatic Ground-truth Generation for Document Image Analysis and Understanding. In *Ninth International Conference on Document Analysis and Recognition*, pages 476–480, Sept. 2007.
[7] C. H. Lee and T. Kanungo. The architecture of TrueViz: a groundTRUth/metadata editing and VIsualiZing ToolKit. *Pattern Recognition*, 15, 2002.
[8] 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.
[9] J. Sauvola, S. Haapakoski, H. Kauniskangas, T. Seppänen, M. Pietikäinen, and D. Doermann. A distributed management system for testing document image analysis algorithms. In *ICDAR '97 Proceedings of the 4th International Conference on Document Analysis and Recognition*, 1997.
[10] J. Sauvola and H. Kauniskangas. MediaTeam Document Database II, a CD-ROM collection of document images, 1999.
[11] D. Tkaczyk, A. Czeczko, K. Rusek, L. Bolikowski, and R. Bogacewicz. Grotoap: ground truth for open access publications. In *12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 381–382, 2012.
[12] D. Tkaczyk, P. Szostek, P. J. Dendek, M. Fedoryszak, and L. Bolikowski. CERMINE - automatic extraction of metadata and references from scientific literature. In *Proceedings of the 11th IAPR International Workshop on Document Analysis Systems*, 2014.