Fraunhofer SCAI

FRAUNHOFER INSTITUTE FOR ALGORITHMS AND SCIENTIFIC COMPUTING SCAI

PROCESSING FULL TEXT PUBLICATIONS IN PORTABLE DOCUMENT FORMAT

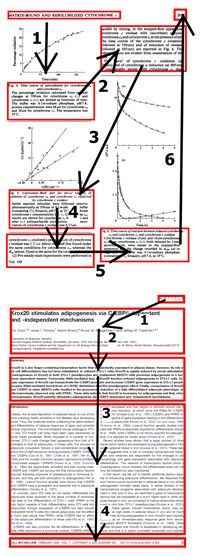
Roman Klinger, Robert Pesch, Heinz Theodor Mevissen, and Juliane Fluck

Introduction and Motivation

- Portable Document Format (PDF) is the main file format for publishing full text in the scientific community
- Number of electronically accessible publications is increasing (open access)
- More entities can be found in full text compared to abstracts [3]
- Visualization of enriched entities in original PDF layout helps understanding and interpretation
- Besides textual information in full text documents, additional sources for extracting information can also be considered:
  - The supplementary material with additional results
  - Tables summarizing important facts
  - Figure captions providing short summarizations of depictions
  - Footnotes providing additional information

⇒ Not all of that is available in provided XML files (e.g. by PMC)

Challenges



Weird text orders

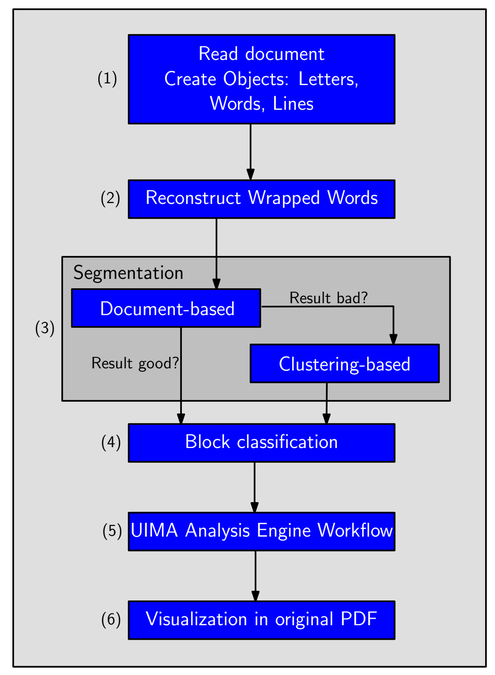
PDF is a graphically oriented format
⇒ extraction of content is non-trivial!

- No logical markup: tables, figures, headers, footers and footnotes interrupt text flow
- Stored text order corresponds only roughly to reading order
  - Order of paragraphs can be incorrect
  - Column-wise layout is not always stored column-wisely
- Hyphens are added to wrap words
- Encoding issues, mapping to standard characters needed
- Content included, which is invisible

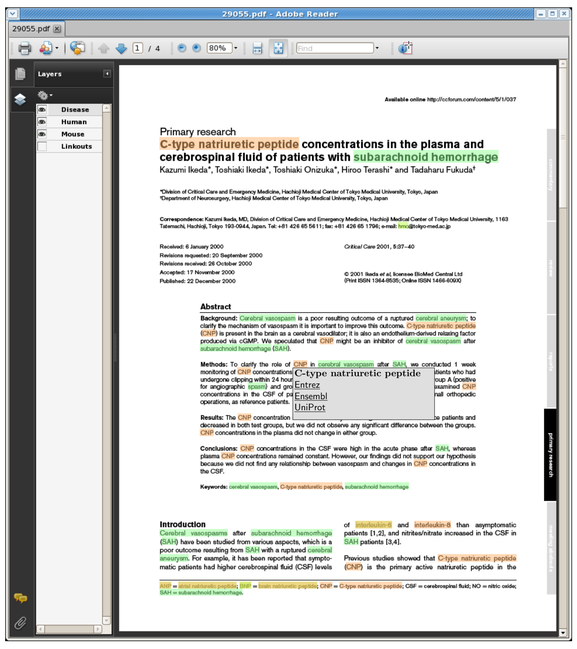
Methods

Implementations to read/write PDF in UIMA framework:

1. Transform document into suitable data structure (PDFBox)
2. Reconstruct wrapped words with TEX heuristic and Wikipedia-lookup
3. The reading order is reconstructed and text blocks are created
   (a) Use stored order unchanged or
   (b) Apply hierarchical clustering on words (followed by sorting)
4. Text blocks are classified rule-based and separated from main text
   - Abstract; Header/Footer; Tables/Figures; Captions; Footnotes
5. Entities are annotated using a UIMA Analysis Engine of interest (e. g. to annotate gene names)
6. Entities are highlighted in the PDF document and link outs are attached to the found entities (iText, JavaScript)



Examples



Visualization Example

Block Classification Example
removed in online version

Results

Classification results on randomly selected independent set from PubMed Central Open Access Corpus

- 71 pages (excluding front pages)
- 77 front pages
- 37 complete PDF publications

Category	Prec.	Rec.	F_1
Header/Footer	0.99	0.94	0.96
Abstract	0.83	0.90	0.86
Table caption	0.61	0.99	0.75
Table cell	0.95	0.91	0.93
Figure caption	0.99	0.93	0.96
Text in figure	0.91	0.68	0.78
Footnote	0.90	0.79	0.84

Evaluation on 800 full documents sampled from PubMed Central Open Access Corpus

- Includes 27660 wrapped words!

Search for Genes/Proteins [1] and IUPAC Names [2] in PDF and provided XML

- Gene mentions:
  90180 in PDF, 52987 in XML
- IUPAC mentions:
  1537 in PDF, 829 in XML
- Unique Gene mentions:
  35375 in PDF, 18499 in XML
- Unique IUPAC mentions:
  1098 in PDF, 646 in XML

Summary

- Work flow enables processing full text PDF documents in UIMA
- Visualization in PDF document helps understanding and interpretation
- Elements like tables and figures are classified in the document
- Extendable work flow and applicable to different identification tasks

Further Work

- Develop document structure recognition
- Compare the performance of common identification tasks in parts of documents
- Evaluate performance of existing annotation engines on PDF documents

References

[1] R. Klinger, C. M. Friedrich, J. Fluck, and M. Hofmann-Apitius. Named Entity Recognition with Combinations of Conditional Random Fields. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 89–91, Madrid, Spain, April 2007.

[2] R. Klinger, C. Kolářik, J. Fluck, M. Hofmann-Apitius, and C. M. Friedrich. Detection of IUPAC and IUPAC-like Chemical Names. *Bioinformatics*, 24(13):i268–i276, 2008. Proceedings of the International Conference Intelligent Systems for Molecular Biology (ISMB).

[3] M. J. Schuemie, M. Weeber, B. J. Schijvenaars, E. M. van Mulligen, C. C. van der Eijk, R. Jelier, B. Mons, and J. A. Kors. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604, November 2004.

This work has been partially funded in the framework of the European integrated project @neurIST, which is cofinanced by the European Commission through the contract no. IST-027703 (see http://www.aneurist.org)

Contact

Roman Klinger
roman.klinger@scai.fraunhofer.de
http://www.scai.fraunhofer.de