Nature Precedings : doi:10.1038/npre.2009.3141.1 : Posted 22 Apr 2009

TEXT MINING IN FULL TEXT ARTICLES
METHODICAL AND REPRESENTATION ISSUES

Roman Klinger, Robert Pesch, Theo Mevissen, Juliane Fluck

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 St. Augustin, Germany

In many cases, information from abstracts of biomedical publications is not sufficient for annotation of database entries. Therefore, text mining systems supporting curators of biodatabases should be able to process full text articles. Beside the technical problems arising from full text parsing, the representation of the annotated full text is an important issue. Journal articles are mostly electronically available in PDF or HTML format. Also with more easily manageable XML formats, readers would like to have a visualisation of annotations and semantic enrichment directly in the PDF or HTML. We summarize the technical problems arising from parsing of HTML and PDF journal full texts and show first results of visualisation in both formats.

Textflow problems

Most crucial problems arising from parsing full text in the commonly available PDF and HTML formats:

HTML

- Additional text in sidebars
  - Worst: Various sidebar styles depending on journal type
- Hyperlinking: additional texts in separate web pages
  - Worst: Linking via Java Script
- No well-defined structure like in XML

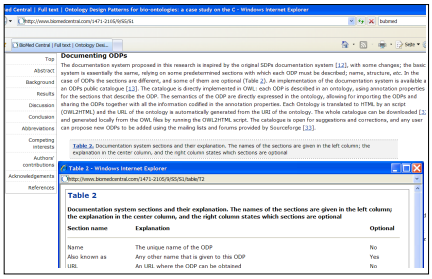
PDF

- Causes problems as it is a graphical format, no logical markup!
- Underlying text not always in correct order
  - No big problem for indexing or named entity recognition
  - Big problem for information extraction, zoning

- Typical flow problems:
  - Multiple columns versus one column
  - Abstract identification
  - Tables
  - Headers and footers
  - Footnotes

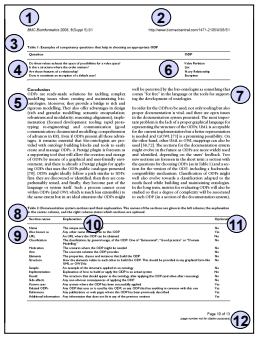
Even more problems arising if OCR is necessary!

HTML representation



Example of an HTML visualisation
- Simple removal of HTML tags would include side bar text in journal text.
- Removal of side bar not trivial as HTML format is not well-defined
- Table as separate Web Page

PDF representation



Example of a PDF to text extraction
Numbering shows detected order of text via Java PDF library

Textual representation and adaptation of named entity recognition tools

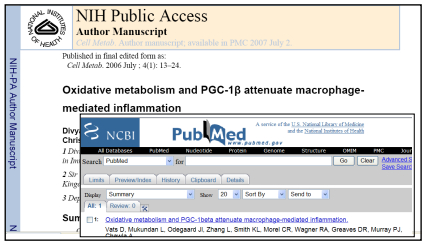
HTML

- Select content to be processed:
- Remove most of the HTML tags
- Convert some of the tags to symbols (ä ₁)
- Convert special characters like beta:
  - some codes for beta are: &#223, &#914 ,&#946, &beta ,&Beta with different amount of leading zeros for the numbers
- Use content of some attributes (<img alter="..">)

PDF

Tokenization issues are main problem for visualization:
- Extra hyphens at page, line, column breaks in between words
- Typically PDF converters tokenize on white space
  - Example problem: 'BCL2-induced'
- Encoding of Greek letters (e.g. <ce><b2> for beta)
- Special handling of glyphs
  - "œ" is not "oe", "IJ" is not "IJ", "Æ" is not "AE", "α" is not "alpha"
  - Mapping of such symbols to positions in PDF

Unicode representation



Example for different representation of special characters in full text and in PubMed abstract
In PubMed all greek letters are converted to the written form of the letter

Different zones in text

Detecting information in areas not of interest reduces the precision:

False matches may occur in
- header lines,
- footnotes,
- authors names and affiliation
- acknowledgements or
- references (cf. annotated PDF)

HTML may provide zoning information in form of class-attributes, but this information is not always available.

PDF documents lack structural information.

ProMiner: dictionary based named entity recognition

- Standard ProMiner software includes the gene and protein dictionaries for human, mouse and a disease dictionary
- Further dictionaries are available

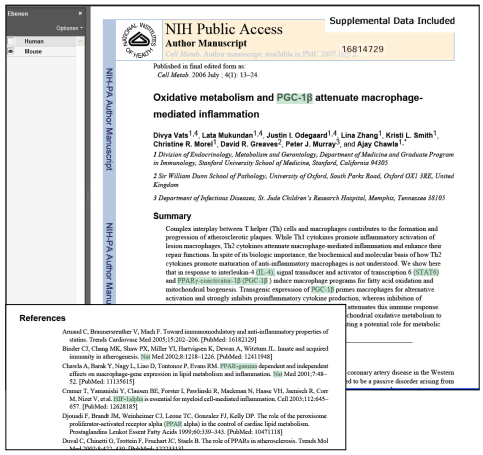
- Good Performance in BioCreative I and II assessments

- ProMiner could easily be integrated in a larger processing pipeline (e.g. as a pre-tagging module for information extraction systems)
- It is available as Java module with defined input and output streams
- Integrated as an annotator service for named entities in the UIMA framework.
- Linkage to other data easily possible through the provided mapping to databases or controlled vocabulary

Further reading:
http://www.scai.fraunhofer.de/prominer.html

Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer and Juliane Fluck
ProMiner: Rule based protein and gene entity recognition. 2005 , BMC Bioinformatics 2005, 6(Suppl 1):S14.
Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu HH, Torres R, Krauthammer M, Lau WW, Liu H, Hsu CN, Schuemie M, Cohen KB, Hirschman L.
Overview of BioCreative II gene normalisation
Genome Biol. 2008;9 Suppl 2:S3. Epub 2008 Sep 1.

Annotated PDF



ProMiner extension for PDF annotation

Established PDF extension for named entity recognition system ProMiner:

- Automatic process of text extraction from PDF
- Annotation of text with different dictionaries
- Integrated Visualization in PDF files

Example on the left:
- Annotation could be separately turned on/off
- Highlighted entities in original PDF
- Pop-Up Annotations with database references
- Links to external databases
- Different encodings of special letters (like Greek symbols)

Remaining problems:
- False matches in non-interest zones
- False matches in affiliations, acknowledgments and references

Contact:
Juliane Fluck Fraunhofer SCAI,
Schloss Birlinghoven,
53754, St. Augustin
juliane.fluck@scai.fraunhofer.de