Semester Project:
Content Extraction from PDF Scientific Articles

J.-C. Chappelier
IC/LIA
jean-cedric.chappelier@epfl.ch

R. Richardet
Blue Brain Project (BBP)
renaud.richardet@epfl.ch

16 Nov. 2012

Abstract

The main goal of this project is to push one step further the State-of-the-Art technology in PDF to text conversion.

1 Description

1.1 Context

One goal of the BlueBrain NLP team is to parse scientific articles and extract structured content from them.

So far, only article abstracts have been used. Our next step is to use the full-text of the articles (Pubmed corpus). A major drawback however, is that most of these articles are available only in PDF format, which is a purely presentational format.

Still, most low-level problems (encoding, parsing) could already be solved by using a specialized, high-quality library to parse PDF [1]. The main goal of the proposed semester project is to go one step further using this technology and provide enhanced recognized content.

1.2 Tasks

In the context described above, the following tasks remains open:

1. better content **cleanup**, like removing non-informative headers and footers, collapsing contiguous paragraphs, etc;

2. **parsing tables**: tables in scientific articles contain highly interesting information, but the lack of a good table parser prevents us from using that information;

3. **segmentation**: most article are structured in chapters (introduction, materials and methods, results, discussion, ...); we would like you to develop (based on software tools) a probabilistic machine learning model to classify the different sections of a paper.

2 Environment

This semester project is co-supervised between IC-LIA and SV-BlueBrain Project.

Material and working environment will be provided at BBP (QIJ 3).

3 Required Skills

The target student is a typical Computer or Communication Science, Bachelor or Master Student.

No specific skills are required but a willing interest in the topic.

4 References

[1] http://snowtide.com/