Semester Project: Extraction of Protein Concentration from Scientific Literature

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

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

29 May 2013

Abstract

The main goal of this project is to extract various numerical entities from Scientific Literature. The main focus will be put on protein concentration in specific regions.

1 Overview & Goal

In the interest of various researchers currently engaged in the Blue Brain Project, a tool must be created in order to gather necessary information about the protein concentrations encountered in different cell types. Today, a huge amount of this very specific data of interest is hidden in the bodies of scientific publications and cannot be automatically consumed. The goal of this project is to design and develop a solution prototype which meets the following objectives:

- enable researchers to quickly access information about protein presence (yes or no) and concentrations (absolute or relative) in different cell types and brain regions;
- keep track of the various data resources and fact provenience in order to create optimal conditions for efficient collaboration among researchers;
- give researchers the opportunity to provide their feedback to the application in form of new/corrected annotations on the proposed protein extractions; this way, the system will continue learning and hereby improve its performance; in order to gain researchers' attention the tool should come with a nice and welcoming user interface.

2 Project steps

1. evaluate existing modules
2. perform qualitative analysis on some sample documents, of relations between protein entities and concentrations
3. create a test corpus of protein-concentration relations, based on the literature and interaction with researchers of the domain
4. develop a UIMA chain to extract protein concentrations (using existing available modules & creating new ones)
5. evaluate and iteratively improve the extraction performance of the UIMA chain

3 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).

Existing work:

- UIMA NER modules to identify measures, proteins (BANNER, Gimli), cells and brain regions
- UIMA modules for chunking and syntactic parsing
- BioNumbers (`http://bionumbers.hms.harvard.edu/`)

4 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.

Knowledge in Natural Language Processing is a plus.