1. Introduction

This project is a system that used to extract information or named-entity recognition from academic papers. There are two parts in this system, the first is extracting information from the raw documents such as PDF into the XML or CSV files, the second is analyzing the useful sentences in these files.

2. Technical Feasibility

The main technologies and tools that are associated with this project are

Eclipse

Pycharm

Grobid

Each of the technologies are freely available and the programming skills required are satisfied by research team.

The Grobid is a machine learning library for extracting, parsing and re-structing raw documents into structed encoded documents with a particular focus on technical and scientific publications. It is written in java and can be called in Python program. This tool has been made available in open source.

Therefore, this project is technical feasible.

3. Possible Solution

To extract information from papers in PDF files, the Grobid library offers the Python client to use this library in Python. Running the Grobid program in Pycharm, the processFulltextDocument service of Grobid processes all the PDF files present under the input directory, and writes the resulting XML TEL files under the output directory. Then the program stores the path of the output directory as the output to be the input of the jar files.

The jar files are designed to analysis the useful sentences in the XML TEL files from Grobid output directory. In Pycharm, the Python uses subprocess or os library in Python to call jar file and uses the path of XML TEL files that stored before as the input of jar. After analyzing by jar files, the useful sentences will display in the Pycharm.

4. Alternate Solution

The alternate solution for calling jar file in Pycharm is using Jython in Python. Jython is a Java implementation of Python. It is freely available. Setting the Jython in the Pycharm, then appending the path of jar files to the Python system library, so in the Pycharm, the Python can import the packages in the jar files. Therefore, the Pycharm can display the analyzing result.

5. Evaluation Criteria

The key aspects of Grobid are following:

High speed of header extraction and references parsing.

Scalability and robustness.

Lazy loading of models and resource.

Robust and fast PDF processing with pdfalto and dedicated post-processing.

Therefore, The Grobid library is effective to this project.

Both subprocess, os library in Python and Jython are freely available. The subprocess and os can be directly used in the Python and Jython needs to append the path of jar files to the Python sys path, then can be applied in the program. However, which one would be suitable for this project still need to be tested when jar files finished.

6. Conclusion

The aim of this project is to design NER/Information extract system, the two part---extracting information/NER from PDF papers files to output XML files and analyzing useful sentences from XML files can be realized.