An Architecture for Linguistic and Semantic Analysis on the arXMLiv Corpus

D. Ginev, C. Jucovschi, S. Anca, M. Grigore, C. David, M. Kohlhase
http://kwarc.info/projects/lamapun/
Jacobs University Bremen, Germany

The arXMLiv corpus is a remarkable collection of text containing scientific mathematical discourse. With more than half a million documents, it is an ambi- tious target for large scale linguistic and semantic analysis, requiring a generalized and distributed approach. In this paper we implement an architecture which solves and automates the issues of knowledge representation and knowledge management, providing an abstraction layer for distributed development of semantic analysis tools. Furthermore, we enable document interaction and visualization and present current implementations of semantic tools and follow-up applications using this architecture.

We identify five different stages, or purposes, which such architecture needs to address, encapsulating each in an independent module. These stages are determined by the different properties of the document formats used, as well as the state of processing and linguistic enrichment introduced so far. We discuss the need of migration between XML representations and the challenges it would pose on our system, revealing the benefits and trade-off of each format we employ.

In the heart of the architecture lies the Semantic Blackboard module. The Semantic Blackboard comprises a system based on a centralized RDF database which can facilitate distributed corpus analysis of arbitrary applications, or analysis modules. This is achieved by providing a document abstraction layer and a mechanism for storing, reusing and communicating results via RDF stand-off annotations deposited in the central database.

Achieving a properly encapsulated and automated pipeline from the input corpus document to a semantically enriched output in a state-of-the-art representation is the task of the Preprocessing, Semantic Result and Output Generation modules. Each of them addresses the task of format migration and enhances the document for further semantic enrichment or aggregation. The fifth module, targeting Visualization and Feedback, enables user interaction and display of different stages of processing.

The overall architecture purpose is to facilitate the development and execution of semantic analysis tools for the arXMLiv corpus, automating the migration of knowledge representation and establishing a complete pipeline to both a presentation and content enriched document representation.

Additionally, we present three applications based on this architecture. Mathematical Formula Disambiguation (MFD) embodies an analysis module that uses heuristic pattern matching to disambiguate symbol and structure semantics. Context Based Formula Understanding (CBFU) is another Semantic Blackboard module which in turn focuses on establishing context relationships between symbols, helping to disambiguate their semantics. We also present the Applicable Theorem Search (ATS) system, a follow-up application that performs search functions, retrieving theorem preconditions for the user.