BOOTStrep is a research project funded by the IST program of the European Commission that brings together experts in the field of natural language processing, lexicon design, ontology development and bioinformatics. The goal is to discover new ways of ontology design and processing of scientific literature to deliver a range of facts to researchers in biomedical sciences. The first focus in the project is the use of ontologies for information extraction tasks, the second focus is the generation of lexicon resources from sets containing biomedical terms and the third focus is the generation of a fact database representing knowledge from the scientific literature. Basis is a text analysis system which continuously increases its coverage by analysing biological documents.

Summary of Activities

During the first seven months of the BOOTStrep project, the partners have identified a biological research topic of great importance that serves as first proof of principle for the anticipated information extraction solutions. The selected topic is 'Gene Regulation' in selected species. As part of ongoing work relevant to this topic all relevant terminology has been identified, a specific corpus has been generated from the public literature, the structure of the Bio-Lexicon has been defined, and the general IT infrastructure for the Human Language Technology (HLT) has been chosen based on the industry standard UIMA (Unstructured Information Management Architecture, IBM).

In 2007 BOOTStrep mainly aims at implementing a first prototype version of the IBM UIMA based system architecture. Using documents from the application domain Gene Regulation, first versions of the Bio-Lexicon and the Bio-Ontology will be populated and integrated into the HLT infrastructure to extract facts from the scientific literature.

The topic ‘Gene Regulation’ is highly relevant

Gene regulation is part of ongoing research in molecular biology. A large number of biologists investigate in special gene regulation mechanisms that control the development of the cell and the organism. On an even larger scale gene regulation is relevant to the development of diseases in humans and in model organisms.

The importance of gene regulation in biology is also reflected by the large number of terms relevant to this domain. Via our analyses we identified a large range of terms relevant to this domain: 1,710 gene ontology terms, 3,216 InterPro entries, 2,624 IntAct entries, 6,520 Medical Subject Headings terms, 56 SwissProt Keywords, 14,067 OMIM entries and over 100,000 UniProt/SwissProt proteins. Currently these terminologies are prepared to be integrated into the modular structure of the Bio-Lexicon (ongoing interdisciplinary work).

Technical Infrastructure

The IBM Unstructured Information Management Architecture (UIMA) was chosen as the technical backbone architecture of the project. Common annotation types for the integration of Human Language Technology (HLT) components in the UIMA infrastructure were designed. The types for the document meta-information, document structure and morpho-syntactic annotation are widely established. Different components (Sentence Splitter, Tokenizer, POS Tagger, Chunker, Parsers, Acronym Recognizer, Named Entity Tagger) were collected, in combination with their detailed description accordant to the usage in the project. First versions of machine learning-based methods to enable classification and information extraction issues were implemented and tested.

Bio-Lexicon & Bio-Ontology

The generation of the Bio-Lexicon and the Bio-Ontology is central to the BOOTStrep project. The Bio-Lexicon delivers terms from biological data resources, such as the UniProtKb database and the gene ontology, in a representation that is optimized for human language technologies, in order to improve the performance. The Bio-Ontology will be linked to the Bio-Lexicon to provide semantic information with the terms in the Bio-Lexicon, which is one of the main research ideas of the BOOTStrep project.

Apart from the generation of the terminological lists, research work in the project is concerned with the resolution of polysemy, i.e. terms having different meanings. A first set of lexical objects with a first repository of data categories (conformant to the most recent LMF ISO proposal) has been provided by the consortium. The Bio-Lexicon model has been instantiated both in XML (with the relevant DTD) and in RDF (with the relevant RDF-schema).

The BOOTStrep project partners endeavoured to construct a top-level Bio-Ontology, called BioTop. Two BioTop related publications to international conferences were accepted.

The definition of the Bio-Lexicon and the Bio-Ontology are ongoing work. Together, these resources will embody the resource-oriented perspective of the project. In particular, the Bio-Lexicon will provide language-specific information, while the Bio-Ontology will represent language-independent knowledge of terms. Finally, these two resources will be made available to the community as a final result.

User Involvement, Promotion and Awareness

Publications evolved out of the BOOTStrep context were represented at major conferences and workshops, like the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (Coling/ACL06), the Conference on Empirical Methods in Natural Language Processing (EMNLP06), the International Conference on Formal Ontology in Information Systems (FOIS06), or the fifth international Conference on Language Resources and Evaluation (LREC06). The papers mainly covered topics like Information Extraction and co-reference resolution, Ontologies for the Bio-domain, and models for multilingual biomedicine lexicons. The ISO community has been informed about the adoption of LMF by BOOTStrep. Feedback is expected as a possible improvement to the model.

Dr Ananiadou from the University of Manchester was awarded 1 of 20 IBM UIMA Innovation Awards.

Press releases found their way into newspapers like “Welt”, and led to a dissemination of the project start and goals in the community and publicity.

Future Work or Exploitation Prospects, as appropriate

In 2007, project work will be further separated into tool- and resource-oriented tasks. On the one hand, BOOTStrep will provide a software infrastructure needed for text analysis proper and knowledge capture from texts. On the other hand, an inventory of terms will be generated from a variety of existing resources, which may serve as links between facts from the unstructured literature and corresponding entries in the structured biomedical databases. The representational model for the terminological lexicon (Bio-Lexicon) and the ontological repository (Bio-Ontology) will have to be further populated.