### **BOOTStrep Annual Report**



### www.bootstrep.eu

BOOTStrep aims at providing re-configurable and extensible knowledge resources and software tools which support the automatic content analysis of documents from the biomedical domain. Such resources are a collection of terms from the biomedical domain (BioLexicon) organised as an electronic lexicon, an ontology for the semantic representation of domain knowledge related, in particular, to gene regulatory events, and a suite of text mining components embedded in a common architecture framework (UIMA). Emphasis has been put on the automatic support for setting up and maintaining all knowledge resources. Major services provided from the BOOTStrep systems comprise the extraction of biologically relevant named entities (cells, proteins, genes, etc.), relations and events (e.g., protein-protein interactions, regulation of gene expression). Selected information from the scientific literature will continuously be assembled and updated in an automatically generated database which holds biological facts and assertions and links them to already existing biological terminologies and databases.

# **Summary of Activities**

The BOOTStrep project has a strong resource orientation both with respect to text analysis infrastructure and biomedical applications. As far as infrastructure developments are concerned, the consortium supplied a high-coverage biological terminology and a more focused sublanguage lexicon, with corresponding linguistic specifications. At the conceptual level, an ontology for gene regulation has been set up, while the whole natural language processing cycle is covered by comprehensive text analysis pipelines which incorporate term handling (tokenisation, morphological normalisation, lexicon look-up), syntactic analysis (POS tagging, chunking, parsing), semantic processing (term recognition, the extraction of named entities, relations, and events etc.), up to the level of discourse analysis (anaphora resolution). It should be pointed out that manual intervention was intentionally kept to a minimum, while major emphasis was put on automatic learning procedures for creating terminological, lexical and database content. BOOTStrep's major bio application consists of an automatically generated fact and relation database which captures major portions of the contents locked in natural language documents. Web interfaces are available for all major resources. While basic functionality has been provided at the component level in 2007, the main efforts of the project work in the next year will target the evaluation of the major resources being developed (biological lexicon, ontology, fact database) in the light of various use cases for biologists to access the information, including information retrieval and extraction tasks.

### • Creating a comprehensive biological lexicon (BioLexicon)

The BioLexicon is a semantically enriched resource encoding lexical information relevant for biology. To achieve semantic interoperability, we adopted the ISO Lexical Markup Framework as reference meta-model. The BioLexicon combines general purpose lexical items, biological terminology extracted from both bioresources and specialised corpora (E-coli). Machine learning based named entity and term normalisation techniques facilitate both the population of the BioLexicon and the mapping between resources, acquired terms from text and the BioLexicon. The current version comprises 1,309,496 lexical entries, lemmas and senses; 2,047,903 orthographic variants and 1,697,804 semantic relations. The BioLexicon also contains part-of-speech and inflectional linguistic information, as well as rich domain specific syntactic and semantic subcategorisation frames for biologically relevant verbs and nominalisations (e.g. *inhibition*).

### • Creating an ontology for gene regulation (BioOntology)

The Gene Regulation Ontology (GRO) covers the relevant concepts underlying the biological description of gene regulation processes. This topic is highly relevant to different biological research domains and can be dissected into a limited number of events that can be well described in an ontological resource to feed domain knowledge into an advanced information extraction solution. Relevant OBO ontologies have been considered for the construction of GRO (in particular, Gene Ontology, Sequence Ontology, Molecule Role Ontology, Event Ontology, ChEBI) and the harvested concepts were integrated under the OBO Foundry policy (using only a limited set of well-defined conceptual relations), within a description logics framework (OLW-DL). The current version of GRO encompasses 420 concept classes, 9 semantic base relations and 310 relation instances. We currently are in the process of submitting GRO to the OBO platform, thus making this resource widely available for prospective biological users (such as database curators, experimental biologists, etc.).

#### • Creating the language processing infrastructure

We assembled a large variety of alternative software components whose entirety covers the major steps of comprehensive content-oriented text analysis. The analysis streams have been bundled under UIMA in two major tool collections, viz. one building on OpenNLP tools and software developments within the Jena Group, the other building on software as developed in Manchester and the University of Tokyo, which also includes OpenNLP tools. Both streams incorporate, to a different degree though, linguistic information available from the biological terminology and the biological lexicon, as well as the Gene Regulation ontology, together with an anaphora resolution engine from Singapore (I2R). As a major innovative achievement from the software engineering perspective all of the components (currently, up to 25) were embedded in the UIMA middleware, thus allowing easy exchangeability and reuse of existing software. From a theoretical perspective, a major breakthrough is achieved by the definition of type systems for the entire language processing cycle, which constitute the target language for wrappers for each software component. Such type systems will form the basis for international standardisation activities for large-scale language analysis systems.

### • Initialising the biological fact database

The database serves the purpose to bring together the different resources from the BOOTStrep project: the BioLexicon, the BioOntology (GRO) and the extracted information from the literature which is kept in the biological fact database. As core topic the fact database will contain information relevant to gene regulation, but wherever necessary, the scope of the fact database will be extended. The biological fact database will deliver results from the basic functionality of the language infrastructure. This database contains term occurrences of biologically relevant named entities (as defined by the biological terminology and GRO), as well as

relations from the chosen application domain, viz. gene regulation in E. coli and other species.

According to our plans, the current and preliminary version will be replaced by a larger fact repository which will be based on a streamlined version of the NLP architectures, in particular taking high-throughput analysis requirements of several millions of documents into account. We are considering the following modifications: we will deal with binary relations, as well as with additional constraints, i.e., experimental conditions (e.g., temperature, pressure, humidity, purity of cell cultures) that influence the validity of biological statements (hypothetical statements). Another major challenge of future work in 2008 will be to link these automatically generated findings to already existing annotation data as manually provided by human annotators.

### **User Involvement, Promotion and Awareness**

The BOOTStrep consortium has paid considerable attention to the needs of prospective users through close alliances with experimental and system biologists of two project member institutions – the EBI and the University of Manchester (National Centre for Text Mining located at the Manchester Interdisciplinary Biocentre (MIB)). We will continue to broaden and deepen this collaboration by keeping EBI and systems biology research centres (via MIB) involved in several use cases for the fact database, the BioLexicon and the BioOntology under construction. Furthermore, we are currently setting up a *Users Advisory Board* which should bring in different views on how the resources being built in the project might have impact on biologists' work. Also, next year's work will be heavily directed at several evaluation scenarios whose constraints and set ups are currently under design.

Among the activities to raise the visibility of the project, first of all, we count the number of peer-reviewed publications. At the end of the year 2007, the BOOTStrep consortium has published 10 workshop papers, 27 conference papers, and 11 journal articles under the BOOTStrep contract, the latter covering prestigious ones such as *Bioinformatics*, *BMC Bioinformatics*, *Artificial Intelligence in Medicine*, etc. From these 48 publications, 5 were authored by at least two authors from different project member groups, and 1 by at least three authors from different project member groups.

Since UIMA, the basis of the language processing backbone of the BOOTStrep project, is becoming more and more a *de facto* standard for language technology, BOOTStrep member activities in this field are highly visible. Udo Hahn from Jena University has coorganised an UIMA workshop at the national conference for natural language processing in Germany and has been awarded one of the 2007 UIMA Innovation Awards (donated with 24,000 USD). Sophia Ananiadou from University of Manchester won for the second time running one of the 2007 UIMA Innovation Awards (\$22,000 USD). Jena University provides (under the very liberal CPL license) open-source access to one of the full NLP pipelines being developed within the BOOTStrep project. The University of Manchester and I2R in Singapore provide either open-source access or Web Services for UIMA-wrapped software to the other full pipeline for NLP developed within the BOOTStrep project.

All these repositories and sites are highly visited (as log data reveals) and sources are (re)used by external users. In addition, in order to capitalise on the impact of the BOOTStrep project, the University of Manchester jointly with the University of Tokyo and the University of Colorado have launched an international initiative (U-Compare) based on the creation of a set of common shared types infrastructure by which a higher level of inter-operability of NLP tools is to be achieved, e.g. tools by other groups can be automatically combined to form larger workflows. In particular, it enables users to compare using systematically the performances of individual tools and their combined workflows.

### **Future Work or Exploitation Prospects**

The major activities of the BOOTStrep consortium in 2008 will focus on increasing the robustness of the NLP backbone and, thus, direct our efforts to set up a high-volume biological fact database based on automatically extracting its contents from natural language documents. Another major goal will then be the evaluation of this fact database under different usage scenarios – as defined by the consortium in close interaction with the yet to be established *Users Advisory Board*.

Moreover, the BOOTStrep consortium will develop an infrastructure for ontology alignment/population/query answering. The aim of this platform is to leverage on existing ontology alignment frameworks, and to make the implementation of methods for ontology alignment/enrichment easily reusable in this context.

In parallel, all major knowledge resources (terminology, lexicon, ontology) and software tools will be made available for external test purposes by users other than BOOTStrep members.

#### **Further Information**

The home page of the BOOTStrep project is continuously updated and contains the project's results in terms of publications, resources and software. In particular, the following links might be useful:

#### Access to the biological lexicon

- www.ilc.cnr.it/BootStrepD22 (password protected)
- wiki.ilc.cnr.it/tomcat/BLWI/ (access to the BL web interface: still in progress)

#### Access to the Gene Regulation ontology

• a sourceforge website for the Gene Regulation Ontology (GRO) is going to be set up, and the GRO will also be included in the OBO framework

## Access to the language processing pipelines

- JENA: www.julielab.de
- UoM: www.nactem.ac.uk
- I2R: various NLP tools are available at http://nlp.i2r.a-star.edu.sg/demo ie.html