The Biomaterials Annotator: a system for ontology-based concept annotation of biomaterials text

Javier Corvi [1], Carla V. Fuenteslópez [2], José M. Fernández [1], Josep Lluis Gelpí [1,3], Maria-Pau Ginebra [4], Salvador Capella-Gutierrez [1] and Osnat Hakimi [1,5]

[1] Barcelona Supercomputing Center (BSC), Barcelona, Spain
[2] Institute of Biomedical Engineering, Botnar Research Centre, University of Oxford, UK
[3] Dept. of Biochemistry and Molecular Biology, University of Barcelona, Spain
[4] Dept. of Material Science and Engineering, Universitat Politècnica de Catalunya, Spain
[5] Faculty of Medicine and Health Sciences, Universitat Internacional de Catalunya, Spain

Abstract: Biomaterials are synthetic or natural materials used for constructing artificial organs, fabricating prostheses, or replacing tissues. The last century saw the development of thousands of novel biomaterials and, as a result, an exponential increase in scientific publications in the field. Large-scale analysis of biomaterials and their performance could enable data-driven material selection and implant design. However, such analysis requires identification and organization of concepts, such as materials and structures, from published texts. To facilitate future information extraction and the application of machine-learning techniques, we developed a semantic annotator specifically tailored for the biomaterials literature. The **Biomaterials Annotator** has been implemented following a modular organization using software containers for the different components and orchestrated using Nextflow as workflow manager. Natural language processing (NLP) components are mainly developed in Java. This set-up has allowed named entity recognition of seventeen classes relevant to the biomaterials domain. Here we detail the development, evaluation and performance of the system, as well as the release of the first collection of annotated biomaterials abstracts. We make both the corpus and system available to the community to promote future efforts in the field and contribute towards its sustainability.

Main Features

A semantic annotator specifically tailored for the biomaterials literature.

Designed to recognize named entities from 17 different categories.

Concept recognition relies on manually curated and validated lexical resources.

Semantic resources

Interdisciplinary nature: concepts from various fields such as biology, chemistry, engineering and medicine.

Multiple nomenclatures, vocabularies, and especially ontologies were identified and combined into a single instrument, the Devices, Experimental scaffolds and Biomaterials Ontology (DEB) was used providing the logical schema and the definition of key categories.

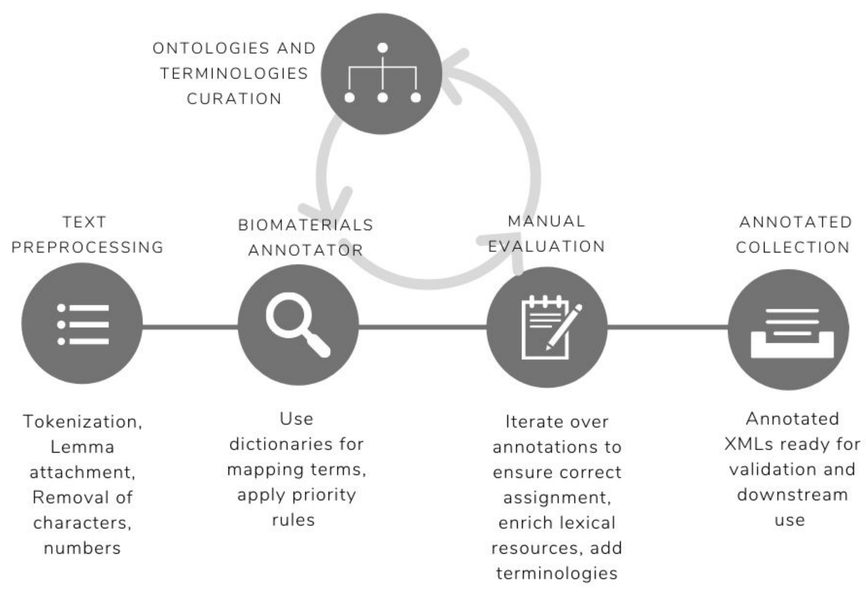

Figure 1: Overview of the workflow used in the development and validation of the Biomaterials Annotator.

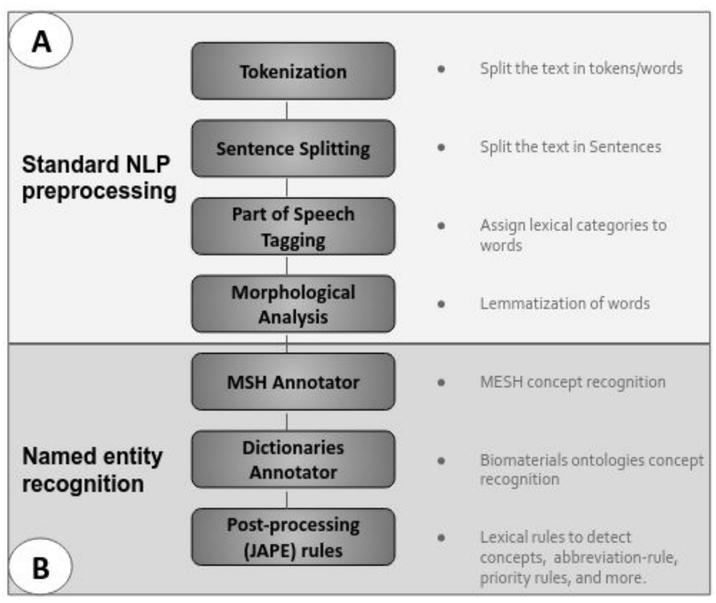

Figure 2: Biomaterials Annotator components; including the standard preprocessing steps (A) and the biomaterials named entity recognition steps (B).

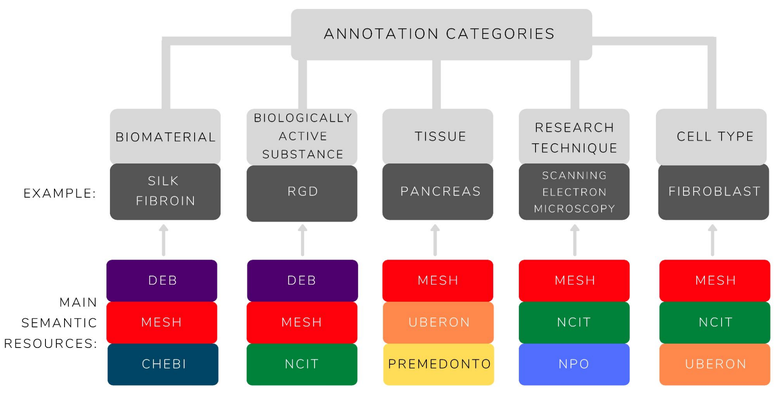

Figure 4: An illustration of part of the annotation schema (showing five out of seventeen categories), which relies on multiple semantic resources for each annotation category.

Figure 3: The appearance of an annotated abstract on GATE's user interface. A) Shows the annotated text and in B) colored labels used to tag annotations by their respective category. C) Information regarding each annotation (type, position, features), and in D) a specific example: "polymers": "BiomaterialType" entity with their corresponding features.

Architecture & Technologies

Components are written in Java and uses the Stanford CoreNLP Natural Language Processing open source toolkit and the General Architecture of Text Engineering (GATE).

The Biomaterials Annotator components are deployed using Docker as software containers; and the execution of the pipeline is orchestrated with Nextflow. By using this architecture, the entire tool, or any of its individual components, can be easily installed and run in heterogeneous environments.

GitHub Actions with Docker Hub are used for continued deployment and integration.
The Biomaterials Annotator and the annotated corpus are available at: https://github.com/ProjectDebbie/Biomaterials_annotator.

Results

- An overall performance of 0.77 avg. F1-score among 17 biomaterials related categories.
- Categories above 0.8 are satisfactorily covered, e.g. Structure, Biomaterial Type and Tissue.
- Lower scores (e.g. Biomaterial, Biologically Active Substance and Cell) are related to missing concepts, ambiguity and ontology disagreements:
  - Biomaterials and Biologically active substance include many ambiguous concepts.
  - Some materials may act as a biomaterial in one set-up, but can also be measured in terms of cell expression or non-biomaterial in another set-up (e.g. collagen).

Category	Precision strict	Recall strict	F-score strict	Precision lenient	Recall lenient	F-score lenient	Precision average	Recall average	F-score average
Adverse Effects	0.94	0.75	0.82	1	0.8	0.87	0.97	0.77	0.85
Associated Biological Process	0.88	0.68	0.77	0.94	0.73	0.82	0.91	0.71	0.79
Biologically Active Substance	0.58	0.43	0.49	0.7	0.52	0.59	0.64	0.48	0.54
Biomaterial	0.76	0.47	0.57	0.83	0.52	0.63	0.79	0.49	0.6
Biomaterial Type	0.92	0.88	0.9	0.98	0.93	0.95	0.95	0.9	0.92
Cell	0.76	0.59	0.66	0.84	0.65	0.73	0.8	0.62	0.69
Effect On Biological System	0.96	0.69	0.79	1	0.72	0.82	0.98	0.71	0.8
Manufactured Object	0.96	0.86	0.9	0.96	0.86	0.9	0.96	0.86	0.9
Manufactured Object Component	0.91	0.84	0.86	0.91	0.84	0.87	0.91	0.84	0.87
Manufactured Object Features	0.68	0.59	0.62	0.71	0.61	0.65	0.69	0.6	0.64
Material Processing	0.78	0.6	0.67	0.83	0.63	0.71	0.81	0.61	0.69
Medical Application	0.68	0.49	0.57	0.82	0.6	0.69	0.75	0.54	0.63
Research Technique	0.81	0.63	0.71	0.87	0.68	0.76	0.84	0.66	0.73
Species	0.97	0.79	0.87	0.99	0.81	0.89	0.98	0.8	0.88
Structure	0.93	0.77	0.84	0.95	0.79	0.86	0.94	0.78	0.85
Study Type	0.96	0.95	0.96	0.99	0.97	0.98	0.98	0.96	0.97
Tissue	0.8	0.77	0.78	0.85	0.82	0.83	0.82	0.8	0.81
Global	**0.84**	**0.69**	**0.75**	**0.89**	**0.73**	**0.79**	**0.86**	**0.71**	**0.77**

Table 1: The performance of the Biomaterials Annotator in a test set of 199 abstracts validated manually by 9 experts.

Category	Count
Adverse Effects	657
Associated Biological Process	6231
Biologically Active Substance	7709
Biomaterial	5726
Biomaterial Type	1543
Cell	6839
Effect On Biological System	972
Manufactured Object	5967
Manufactured Object Component	2307
Manufactured Object Features	4200
Material Processing	2728
Medical Application	3868
Research Technique	3701
Species	2089
Structure	4136
Study Type	1806
Tissue	9997
Entities	**70476**
Tokens	**392605**
Sentences	**15979**
Abstracts	**1222**

Table 2: Annotated biomaterials corpus statistics.

Upcoming actions

- DEBBIE (Database of Biomaterial and their biological effect); automatic workflow for MEDLINE abstracts indexation using the Biomaterials Annotator.
- Results stored in DEBBIE Database for posterior presentation to biomaterials experts.
- Additional steps:
  - Annotating relations and linking identified concepts to manufactured biomaterials objects.
  - Incorporate additional categories using controlled resources.
  - Enhanced performance in key categories such as Biomaterials and Cells.

Contact

Javier Corvi: javier.corvi@bsc.es
Osnat Hakimi: osnat.hakimi@gmail.com
Salvador Capella: salvador.capella@bsc.es
Barcelona Supercomputing center.
C/Jordi Girona, 29, 08034 Barcelona