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
javier.corvi@bsc.es
osnat.hakimi@gmail.com

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.

1 Introduction

The last two decades saw the field of biomaterials and tissue engineering grow from a small niche of biomedical research to an extensive domain, covering topics such as functional materials, cell-material interaction, nanomaterials and medical devices. The expanding scientific data generated by the field is primarily available in text documents, such as peer-reviewed research papers, patents and conference abstracts. This ever-growing knowledge is increasingly harder for researchers to efficiently discover, organize and use. For example, systematically reviewing the applications and scaffolds made of a commonly used polymer such as poly-lactic-glycolic-acid (PLGA), requires skimming through >12,000+ abstracts (MEDLINE search on October 2020). Among the different alternatives for the automated processing of available texts, Natural Language Processing (NLP) workflows for information retrieval and indexing offer a much needed automated solution. Such computational workflows facilitate information discovery, information extraction and organization, saving researchers time and minimizing manual tasks.

Central to information retrieval and indexing is the extraction of concepts of interest, also known as Named Entity Recognition (NER). NER is an integral part of NLP workflows as it allows the automated identification of concepts in unstructured text and its assignment to a pre-defined category or class. For example, in the field of biomaterials, categories may include 'Biomaterials' ('PLGA'), 'Structures' (such as 'fibre' or 'sponge') and 'Tissues' (such as 'tendon' or 'bone'). The use of NER to automatically recognize entities enables several downstream applications, including machine translation, information retrieval and indexing as well as automated question-answering mechanisms.

The recognition of concepts in the biomaterials domain is complicated by language and terminology originating from multiple scientific disciplines (chemistry, engineering, biology, medicine). A significant challenge lies in identifying and combining

Proceedings of the Second Workshop on Scholarly Document Processing, pages 36–48

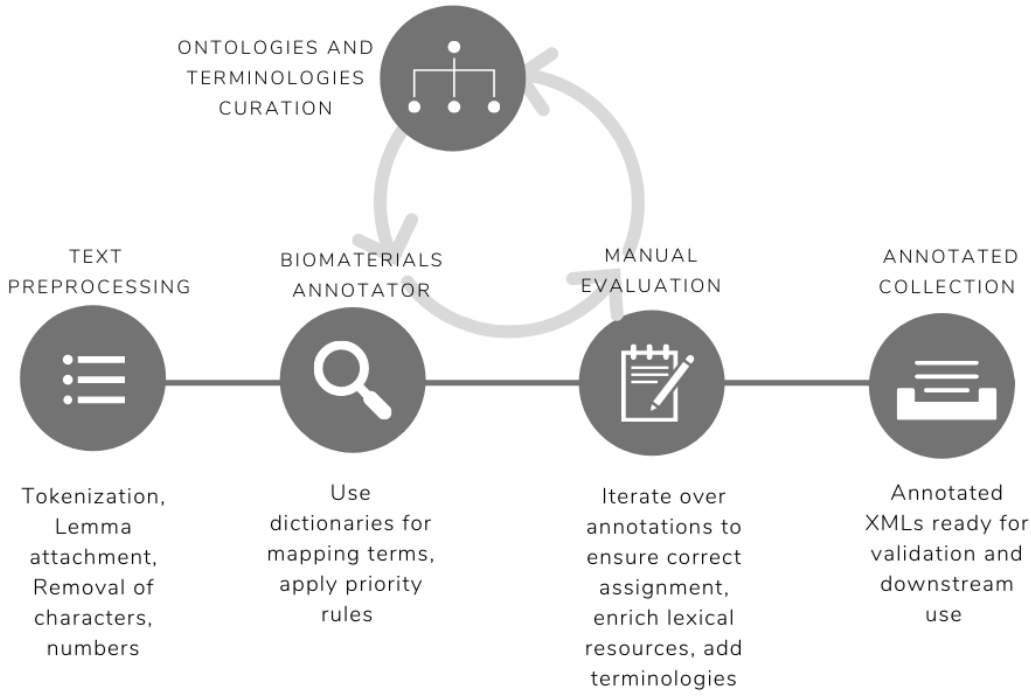

Figure 1: Overview of the workflow used in the development and validation of the Biomaterials Annotator.

lexical and semantic resources across domains, and thus to date there are no automatic biomaterials-specific NER systems to detect relevant entities of interest.

Here, we report the development of the first biomaterials-specific annotation system, designed to recognize named entities from seventeen different categories, reflecting the complexity and diversity of contemporary biomaterials research. When considering approaches for the design of the Biomaterials Annotator, i.e. lexical versus machine learning-based NER, such as CRF or RNN, it was essential to consider the number of desired annotation categories in the system (17) and the absence of an annotated corpora for text mining efforts for the majority of them. Based on these premises, it was concluded that training a model for each category was impractical. Thus, the system relies on manually curated and validated lexical resources.

To cover entities from different domains, multiple nomenclatures, vocabularies, and especially ontologies were identified and combined. To combine these resources into a single instrument, the Devices, Experimental scaffolds and Biomaterials Ontology (DEB) was used providing the logical schema and the definition of key categories (Hakimi et al., 2020).

The resulting open source-system, the Biomaterials Annotator, along with an annotated collection of biomaterials literature, are publicly available for use and further development at https://github.com/ProjectDebbie/Biomaterials_annotator.

2 Previous relevant work

Unlike general purpose NLP systems, biomedical domain-specific tools require advanced approaches to detect classes of interest such as diseases and gene names. In this area, there are several well-known and widely used systems and tools, such as Metamap (Aronson, 2001) and Pubtator (Wei et al., 2013), which were developed using different NER methodologies and approaches, e.g. gazetteers and hand-made rule-based NER; machine learning-based NER that includes Hidden Markov Model, Conditional Random Fields (CRF) and recurrent neural network (RNN); and Hybrid NER (Lee et al., 2003; McCallum and Li, 2003; Song et al., 2004; GuoDong and Jian, 2004; Zhao, 2004; Yeh et al., 2005; Campos et al., 2013; Song et al., 2018; Dang et al., 2018; Kaewphan et al., 2018; Cho and Lee, 2019). In the context of this work, generic text mining tools previously developed for the eTRANSAFE project (Pognan et al., 2021) have been adapted and further developed for the biomaterials domain.

Whilst there are a handful of ontologies in the biomaterials domain (such the nanoparticle ontology NPO (Thomas et al., 2011) and the Bone and Cartilage Tissue Engineering Ontology BCTEO (Viti et al., 2014)), to the best of the our knowledge, the DEB ontology (Hakimi et al., 2020) is the only one that is tailored to link and curate concepts for Biomaterials NER. Therefore, it specifically covers different categories related to the biomaterials domain.

3 Methodology overview

To develop the annotation tool, the workflow in Figure 1 was followed. Various corpora of abstracts were used during the development, covering the general biomaterials literature. These corpora included a collection of manually curated abstracts of the biomedical polymer polydioxanone (Fuenteslópez et al 2021, manuscript in preparation, GitHub repository: https://github.com/ProjectDebbie/polydioxanone_project) and a previously published biomaterials gold standard collection (Hakimi et al., 2020), comprising a total of 1222 abstracts. Corpora were passed through four steps, each described in detail below. The first step was a text preprocessing component (section 3.1). This was followed by concept recognition (section 3.2), initially using the MeSH controlled vocabulary and the DEB ontology. Then, the annotations were evaluated by two domain experts, errors were flagged up and additional lexical resources were added through keyword searches. Concept recognition, manual evaluation and curation of lexical resources were performed in an iterative manner during the development phase (section 3.3) over 1000 abstracts. Once the development phase was completed, validation by domain experts was performed on 199 independent abstracts which were not used during the development process (section 4.1). The resulting annotated collection of biomaterials abstracts was published as open source.

3.1 Text preprocessing

To prepare the text for concept recognition, several Natural Language Processing (NLP) steps were performed, namely: tokenization, sentence splitting, part-of-speech tagging and morphological analysis (Figure 2.A). We developed the Standard NLP preprocessing component which

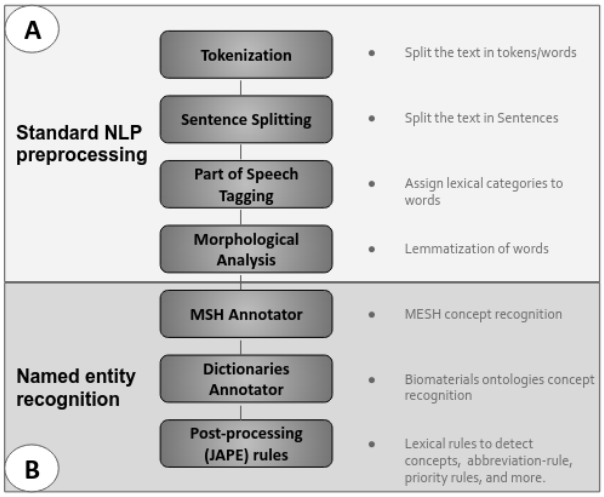

Figure 2: Overview of the components of the Biomaterials Annotator; including the standard preprocessing steps (A) and the biomaterials named entity recognition steps (B).

includes the steps previously outlined. This component is written in JAVA and it uses the Stanford CoreNLP Natural Language Processing open source toolkit. The use of the Stanford CoreNLP API benefits greatly from the provision of a set of stable, robust, high quality linguistic analysis components, which can be easily invoked for common scenarios (Manning et al., 2014). The Standard NLP preprocessing component is available at https://gitlab.bsc.es/inb/text-mining/generic-tools/nlp-standard-preprocessing.

3.2 Concept recognition

Here, we developed NER components to detect relevant entities related to the biomaterials domain based on the DEB ontology in conjunction with other open relevant resources, such as the National Cancer Institute Thesaurus (NCIT) and (CHEBI). A comprehensive description of the resources included in this work is described in section 3.3 and Appendix B. Lexical resources were transformed into gazetteers to be used in the NER process (Figure 2.B). Internally, the NER process was divided into three main steps; the MSH Annotator, which annotates relevant categories from the MeSH terminology; the Dictionary Annotator, which annotated predefined categories from the relevant dictionaries; and the Post-processing step in which specific rules were executed. These include entity recognition based on lexical rules and the removal of false positives, among other tasks.

The MSH Annotator is available at https://github.com/ProjectDebbie/debbie_umls_annotations; and the Dictionary Annotator and Post-processing rules are available at https://github.com/ProjectDebbie/DEBBIE_dictionaries_annotations. These components are instances of the nlp-gate-generic-component (https://gitlab.bsc.es/inb/text-mining/generic-tools/nlp-gate-generic-component), a generic component developed in JAVA by our team that uses the General Architecture of Text Engineering (GATE) software (Cunningham et al., 2013) and can be parametrized with gazetteers and specific handmade JAPE (Java Annotation Patterns Engine) rules. Using the Biomaterials Annotator, every recognised entity is labelled with one of the categories (Figure 3.A-B).

The nlp-gate-generic-component was configured to use the GATE Flexible gazetteer, allowing to capture the words present in the text as well as their morphological root value (lemma). This ensures that inflected forms of a word (i.e. plural, singular, -ing forms, tense) can be recognised and analysed as a single item. In addition, the dictionaries used in the Biomaterials Annotator include preferred synonyms, providing the possibility to map terms semantically to a specific primary concept. Thus, the Biomaterials Annotator performs semantic mapping of the annotations by, not only recognizing the category of an entity, but also linking it to the appropriate entry in a well-established resource (Jovanović and Bagheri, 2017). For example, the terms: *"canine"*, *"dogs"* and *"dog"* were all annotated under the 'Species' category; and inside the features of each annotation the preferred term is *"dog"*. This enables the retrieval of all the corresponding terms using the single search term *'dog'*.

To complete the annotation process, the annotator executes JAPE rules for post-processing functions, such as the removal of false positives and the addition of information to each annotation. Added information includes the ontology source, the ontology term id, the lemma and the preferred synonym (Figure 3.C-D). In addition, JAPE rules were run to identify entities using lexical constraints and address the concept recognition of abbreviations. Rule-based entities recognition can use part-of-speech of concepts, as an example; in the case of 'Cell' category, there is a lexical rule defined to detect concepts:
(Token.pos == "JJ" | Token.pos == "NN") Token.root == "cell"
The inclusion of this rule enables the detection of Cell-type concepts that are not present in the dictionaries; e.g. *"neuronal cells"*, *"cancer cell"* and *"osteogenic cells"*. The discovery of such rules is a continuous work; future Biomaterials Annotator versions will improve the lexical rules included to detect relevant concepts.

Another key problem to address is the recognition of abbreviation concepts; to achieve this problem we developed a post-processing rule based on a modified version of Schwartz's algorithm (Schwartz and Hearst, 2003). First, we detect an abbreviation candidate given a text pattern (regex=*"(?:[a-z]*[A-Z][a-z]*)2,"*); subsequently, the Schwartz's algorithm is applied to detect whether there is a definition that matches the abbreviation candidate in the sentence; in such case, if the definition has an entity class assigned to it, we annotate the abbreviation with the same class. As an example, in the following sentence: *"We investigated the potential of human bone marrow derived Mesenchymal stem cells (MSCs) for neuronal differentiation in vitro...."*; the expression *'Mesenchymal stem cells'* is annotated under the 'Cell' category. But the *'MSCs'* abbreviation is not; moreover in the rest of the text the abbreviation is used instead of its long form. The abbreviation-rule detects *'MSCs'* as an abbreviation of *'Mesenchymal stem cells'* and assigns the 'Cell' category to all the *'MSCs'* mentions in the text.

3.3 Terminologies and ontologies curation and manual evaluation

One of the main hurdles to biomaterials concept recognition is the interdisciplinary nature of the domain, with scientific texts containing concepts from various fields such as biology, chemistry, engineering and medicine. A key objective of the Biomaterials Annotator was to identify and combine lexical resources from the different domains in order to cover as many relevant biomaterials concepts as possible. Resources were identified using a manual, bottom-up approach, with cyclic re-iteration, as shown in Figure 1. As a starting point, abstracts were annotated with the automated NER approach described in section 3.2 using the DEB ontology. After each annotation round, manual evaluation was performed by

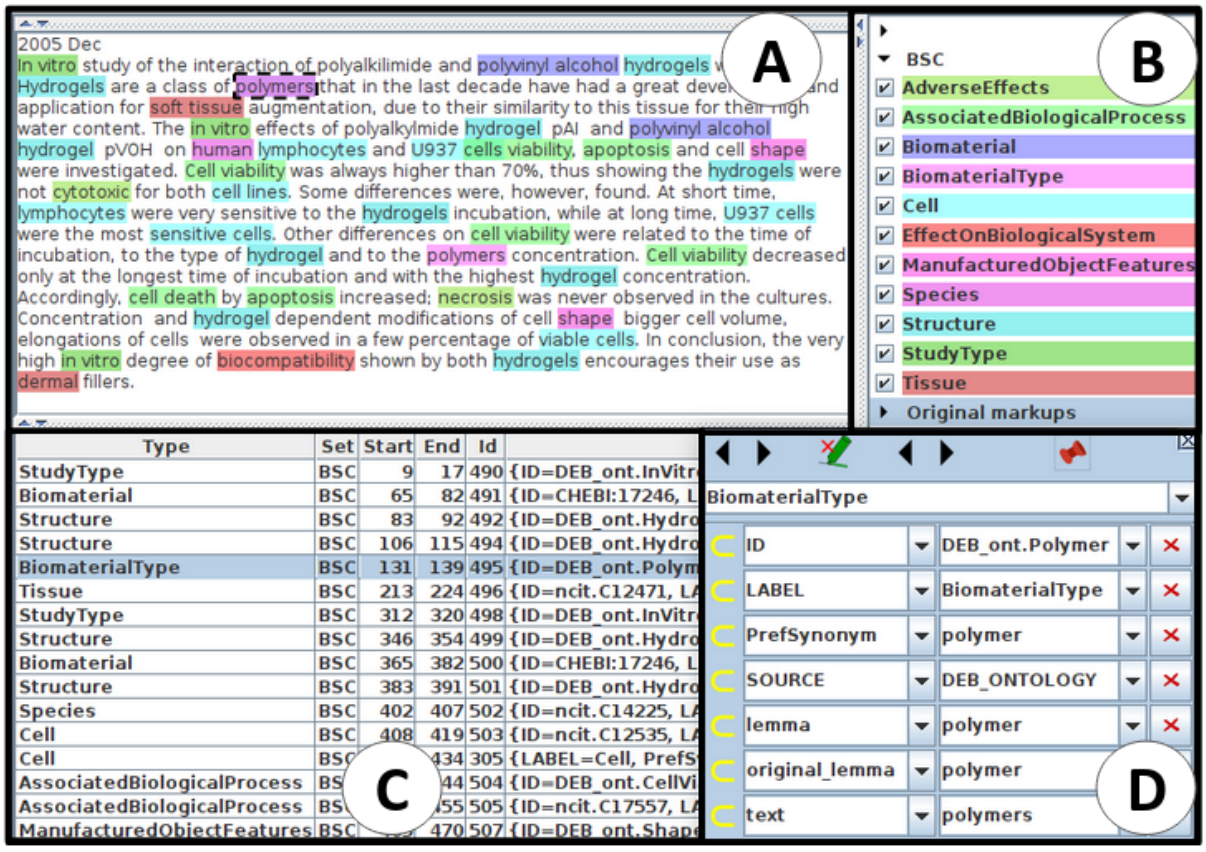

Figure 3: The appearance of an annotated abstract on GATE's user interface. A) Shows the annotated text and in B) colored labels used to tag annotations by their respective category. C) Information regarding each annotation (type, position, features), and in D) a specific example: "polymers": "BiomaterialType"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

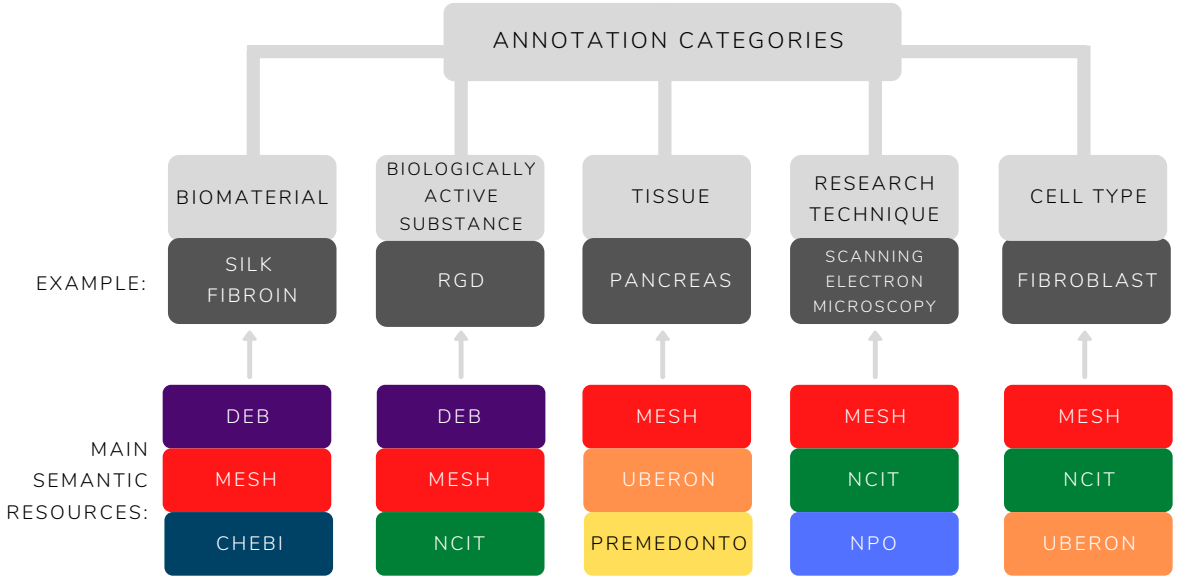

Figure 4: An illustration of part of the annotation schema (showing five out of seventeen categories), which relies on multiple semantic resources for each annotation category. Full details of all categories and resources are in the Appendix A.

two domain experts. The evaluation entailed reviewing samples of 10-20 abstracts in order to flag annotation errors and highlight relevant concepts which were missed by the system. The flagged terms were used for keyword search in the Bioportal (Martínez-Romero et al., 2017) and the UMLS metathesaurus browser (Bodenreider, 2004). Through these searches, specific classes (or 'parent concepts') within relevant ontologies and UMLS 'semantic types' were identified and added to the annotation schema (part of which is shown for illustration in Figure 4). The resources identified belonged to three categories: ontologies, controlled vocabularies and nomenclatures. All the ontologies were open access and downloaded in .owl format from NCBO Bioportal (http://bioportal.bioontology.org) (Martínez-Romero et al., 2017). The controlled vocabulary (MeSH) was downloaded for use under license from the UMLS Terminology Services. The GMDN nomenclature was kindly provided in .xml format by the GMDN agency under a license. A summary of all the used resources is in Appendix B. For the resources to be used in the annotation system, relevant classes were imported into a dictionary (gazetteer) containing the following fields: the term, its label (annotation category), the ID and whenever available, a preferred synonym. The extraction of desired classes from the ontologies to dictionary format was done using an implementation of owlready2 (Lamy, 2017) and the code (named owl2dict_light) is available in an open github repository as part of the (https://github.com/ProjectDebbie/OWL2DICT). The resulting dictionaries are also available (https://github.com/ProjectDebbie/DEBBIE_dictionaries_annotations). These were in turn used by the Dictionary Annotator component for concept recognition as described above in section 3.2.

4 Results

4.1 Expert validation

To measure the efficiency of a text mining system such as the Biomaterials Annotator, it is fundamental to organize and plan a validation stage aimed at indicating the performance of the system. The Biomaterials Annotator was validated through manual verification of the validation set, an independent collection of 199 abstracts. The annotated validation set, resulting from the execution of the Biomaterials Annotator, was manually verified by 9 biomaterials experts. The validation process was performed using the GATE user interface, where annotations made by the system were presented to the biomaterials experts with the possibility of adding missing annotations, removing false annotations and editing annotations. Once the expert had finished the validation of a document, it was saved as a different validated copy.

Two strategies to indicate if two annotations agree or not were considered; a strict approach, in which the annotations agree if they have the same origin and end offset, and a more relaxed or "lenient" approach, where the annotations agree if they overlap at some point. For example, in the partial approach the biomaterials expressions "polyvinyl alcohol" and "polyvinyl" are considered to agree, which does not happen in the strict agreement.

To measure the performance of the NER system, the set validated by the experts was taken as the gold standard and the system's output as the set to be validated. Table 1 shows the recall, precision and F-score, including the strict and lenient approaches, as well as an average between them. The global scores calculated for the system are also presented, obtaining an 0.75 strict F-score, 0.79 lenient F-Score and 0.77 average F-score.

Figure 5 shows the average F-scores calculated for the different categories. Categories with an average F-score above 0.8 are considered categories in which the concepts are satisfactorily covered by the resources used (e.g. Structure, BiomaterialType and Tissue). On the other hand, there are categories with lower scores, and specifically: 'Biomaterial', 'Biologically active substance' and 'Cell'. The categories Biomaterials and Biologically active substance had significantly reduced accuracy because they 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 use in another set-up (e.g. collagen). In the latter case, the human validator will delete the 'Biomaterial' annotation. Solving this kind of ambiguities will require other strategies, such as specific lexical rules or machine learning approaches. Another factor impeding good quality annotations of Biomaterials is the lack of good quality vocabulary of medical polymers. Polymer and co-polymer naming is notoriously variable, with

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.

some named by their commercial name or abbreviation. To address these inaccuracies, future work will involve expanding relevant ontologies using tools such as Spike (Taub-Tabib et al., 2020), including additional lexical rules, and adding machine learning components.

4.2 Full system implementation and availability

A significant challenge for scientific software applications is providing facilities to share, distribute and run such systems in a simple and convenient way. Furthermore, an important concern is the possibility of replicating the results obtained during the research. In order to accomplish these requirements and follow good practices, we developed the Biomaterials Annotator using Docker as software container technology and Nextflow as the workflow manager. Through the use of Docker, all the subcomponents of the Biomaterials Annotator were individually compartmentalized; hosting their own dependencies and programs that work only inside the isolated container. In addition, the Nextflow workflow manager was used for the automated orchestration and execution of the pipeline. By using this architecture, the entire tool, or any of its individual components, can be easily installed and run in heterogeneous environments. The Biomaterials Annotator is available at https://github.com/ProjectDebbie/Biomaterials_annotator.

The Biomaterials Annotator is part of DEBBIE (Database of biomedical materials), a wider system that retrieves abstracts from pubmed, annotates using the Biomaterials Annotator and deposits them in an open access database. DEBBIE is under development and can be accessed at https://github.com/ProjectDebbie/DEBBIE_pipeline.

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.

4.3 Annotated corpus release

Another key objective was to generate the first annotated corpus with entities related to the biomaterials domain. Such a corpus will facilitate the development and evaluation of text mining models for automated extraction of biomaterials-related data from text.

The biomaterials annotated dataset consists of 1222 biomaterials abstracts describing the evaluation of biomaterials in either a laboratory or a

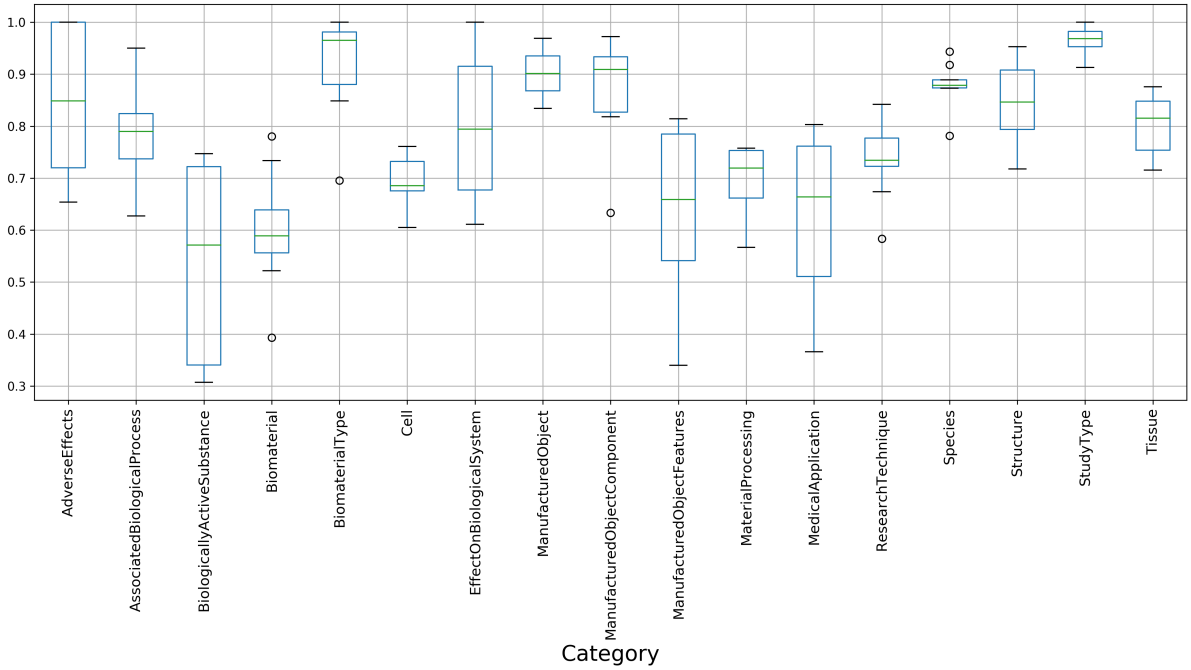

Figure 5: Average F-score of the automated annotations across categories.

clinical setting. Each abstract is individually contained as a separate file under the GATE format. Table 2 shows statistics concerning the number of concepts corresponding to the different categories, as well as the number of total entities, sentences and tokens.

The annotated biomaterials corpus is available and free for use; information to access the corpus can be found at https://github.com/ProjectDebbie/Biomaterials_annotator.

5 Conclusions and future directions

In this work we present the Biomaterials Annotator, an ontology-based NER system that identifies 17 domain specific types of concepts and delivers an annotated biomaterials corpus of 1222 MEDLINE articles available for future text mining and machine learning efforts. We have carried out a validation activity to measure the performance of the NER system, with the participation of nine biomaterials experts, obtaining a global average F-score of 0.77.

Future work in the development of the system could involve annotating relations and linking identified concepts to manufactured biomaterials objects. It may also include incorporating additional categories using controlled resources. Improvements to the system will continue in an iterative manner aiming to enhanced performance in key categories such as Biomaterials and Cells. In addition, future versions of the Biomaterials Annotator will be closely related to the DEBBIE system and include additional functionalities and features developed to achieve its main objectives.

The Biomaterials Annotator and the annotated corpus are open source and available to the community to promote future efforts in the field and contribute towards its sustainability.

6 Acknowledgements

This project has received funding from the European Union Horizon 2020 programme under the Marie Skodowska- Curie grant agreement DEBBIE, project number: 751277. O. H. is funded through a Bosch-Aymerich fellowship. J-M.F, S.C-G and J-L.P are partly supported by INB Grant (PT17/0009/0001 - ISCIII-SGEFI / ERDF). M-P. G. acknowledges the ICREA Academia Award from Generalitat de Catalunya. J.C. is partly supported by eTRANSAFE (received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777365 and support from the European Union's Horizon 2020 research and innovation programme and EFPIA).

References

A. R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*.

David Campos, Sérgio Matos, and JoséLuís Oliveira. 2013. Current Methodologies for Biomedical Named Entity Recognition. In *Biological Knowledge Discovery Handbook*, pages 839–868. John Wiley Sons, Inc.

Hyejin Cho and Hyunju Lee. 2019. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, 20(1):735.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology*.

Thanh Hai Dang, Hoang Quynh Le, Trang M. Nguyen, and Sinh T. Vu. 2018. D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*, 34(20):3539–3546.

Zhou GuoDong and Su Jian. 2004. Exploring deep knowledge resources in biomedical name recognition. page 96.

Osnat Hakimi, Josep Luis Gelpi, Martin Krallinger, Fabio Curi, Dmitry Repchevsky, and Maria-Pau Ginebra. 2020. The Devices, Experimental Scaffolds, and Biomaterials Ontology (DEB): A Tool for Mapping, Annotation, and Analysis of Biomaterials Data. *Advanced Functional Materials*, 30(16):1909910.

Jelena Jovanović and Ebrahim Bagheri. 2017. Semantic annotation in biomedicine: The current landscape.

Suwisa Kaewphan, Kai Hakala, Niko Miekka, Tapio Salakoski, and Filip Ginter. 2018. Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modeling. *Database*, 2018(2018):96.

Jean Baptiste Lamy. 2017. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine*, 80:11–28.

Ki-Joong Lee, Young-Sook Hwang, and Hae-Chang Rim. 2003. Two-phase biomedical NE recognition based on SVMs.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marcos Martínez-Romero, Clement Jonquet, Martin J. O'Connor, John Graybeal, Alejandro Pazos, and Mark A. Musen. 2017. NCBO Ontology Recommender 2.0: An enhanced approach for biomedical ontology recommendation. *Journal of Biomedical Semantics*.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191.

François Pognan, Thomas Steger-Hartmann, Carlos Díaz, Niklas Blomberg, Frank Bringezu, Katharine Briggs, Giulia Callegaro, Salvador Capella-Gutierrez, Emilio Centeno, Javier Corvi, Philip Drew, William C. Drewe, José M. Fernández, Laura I. Furlong, Emre Guney, Jan A. Kors, Miguel Angel Mayer, Manuel Pastor, Janet Piñero, Juan Manuel Ramírez-Anguita, Francesco Ronzano, Philip Rowell, Josep Saüch-Pitarch, Alfonso Valencia, Bob van de Water, Johan van der Lei, Erik van Mulligen, and Ferran Sanz. 2021. The etransafe project on translational safety assessment through integrative knowledge management: Achievements and perspectives. *Pharmaceuticals*, 14(3).

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*.

Hye Jeong Song, Byeong Cheol Jo, Chan Young Park, Jong Dae Kim, and Yu Seop Kim. 2018. Comparison of named entity recognition methodologies in biomedical documents. *BioMedical Engineering Online*, 17(Suppl 2).

Yu Song, Eunju Kim, Gary Geunbae Lee, and Byoungkee Yi. 2004. POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004.

Hillel Taub-Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen, and Y. Goldberg. 2020. Interactive extractive search over biomedical corpora. *ArXiv*, abs/2006.04148.

Dennis G. Thomas, Rohit V. Pappu, and Nathan A. Baker. 2011. NanoParticle Ontology for cancer nanotechnology research. *Journal of Biomedical Informatics*.

Federica Viti, Silvia Scaglione, Alessandro Orro, and Luciano Milanesi. 2014. Guidelines for managing

data and processes in bone and cartilage tissue engineering. *BMC Bioinformatics*, 15(Suppl 1):S14.

Chih Hsuan Wei, Hung Yu Kao, and Zhiyong Lu. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(Web Server issue).

Alexander Yeh, Alexander Morgan, Marc Colosimo, and Lynette Hirschman. 2005. BioCreAtIvE task 1A: Gene mention finding evaluation. *BMC Bioinformatics*, 6(SUPPL.1):1–10.

Shaojun Zhao. 2004. Named entity recognition in biomedical texts using an HMM model. (Grefenstette 1994):84.

A Semantic resources

Table 3: List of semantic resources used by the Biomaterials Annotator

	Semantic Resource Name	Acronym	Scope and relevance	Type
1	Chemical Methods Ontology	CHMO	Methods used to collect chemical experiments data.	Ontology
2	Chemical Entities of Biological Interest	CHEBI	Compounds of biological relevance, macromolecules.	Ontology
3	The Devices, Experimental Scaffolds and Biomaterials Ontology	DEB	Biomaterials-related concepts, materials, structures, material processing.	Ontology
4	EDAM Bioimaging Ontology	EDAM-BIOIMAGING	Imaging and sample preparation techniques.	Ontology
5	Global Medical Device Nomenclature	GMDN	Full names of medical devices.	Nomenclature
6	Interlinking ontology of biological concepts	IOBC	Biological concepts including biological phenomena, diseases, molecular functions, research imaging techniques.	Ontology
7	Medical Subject Headings	MeSH	A hierarchically organized vocabulary produced by the NLM.	Controlled vocabulary
8	National Cancer Institute Thesaurus	NCIT	Vocabulary for clinical care, translational and basic research.	Controlled vocabulary
9	Nanoparticle ontology	NPO	The description, preparation, and characterization of nanomaterials.	Ontology
10	Ontology for Biomedical Investigations	OBI	Biomedical protocols, instruments, data generated, materials, analysis performed.	Ontology
11	Ontology of Nuclear Toxicity	ONTOTOXNUC	Models, chemicals, tools, research techniques and models.	Ontology
12	Precision Medicine Ontology	PREMEDONTO	Human disease terms, genomic, molecular, phenotype, related medical vocabulary.	Ontology
13	Uber Anatomy Ontology	UBERON	An integrated cross-species anatomy ontology	Ontology

B Annotation categories

Table 4: Annotation categories, their respective semantic resources and imported classes

	Annotation category	Definition	Resource and imported classes
1	Biomaterial	A non-drug raw material or substance suitable for inclusion in systems which augment or replace the function of bodily tissues or organs.	DEB: Biomaterials CHEBI: Macromolecule MeSH: Biomedical or Dental Material
2	Biomaterial Types	Classification or nature of biomaterials.	DEB: Biomaterial Type
3	Biologically Active Substance	Substance included in a manufactured object in order to impart a biological activity.	DEB: Biologically Active Substance MeSH: amino acid, peptide, protein Biologically Active Substance Pharmacologic Substance NCIT: Protein Domain
4	Manufactured Object	A physical object created by hand or machine.	DEB: Manufactured Object MeSH: Medical device GMDN: Full nomenclature
5	Manufactured Object Component	A part, region or component referred to as a distinct unit, such as a surface or a layer.	DEB: Manufactured Object Component
6	Medical Application, Disease or condition	Intended use, context, function or outcome of the manufactured object.	DEB: Medical Application MeSH: Disease or Syndrome Therapeutic or Preventive Procedure Anatomical Abnormality
7	Manufactured Object Features	Characteristics inherent or given during processing to a manufactured object or its components.	DEB: Manufactured Object Features MeSH: Chemical Viewed Structurally
8	Structure	The configuration, form or texture associated with a manufactured object or its components.	DEB: Structure
9	Associated Biological Process	A cellular or biological process that the manufactured object is designed to cause or support, or is measured to affect.	DEB: Associated Biological Process MeSH: Organ or Tissue Function Molecular Function Cell Function Biological function NCIT: Cellular Process
10	Material Processing	A planned process which results in physical changes in a specified input material.	DEB: Material Processing CHMO: Material Processing
11	Cell	The reported cell line or primary cell type.	MeSH: Cell NCIT: Cell UBERON: Bone cell, cardiocyte circulating cell connective tissue cell epithelial cell
12	Species	The species and /or breed used in the study.	MeSH: Mammal
13	Tissue	A tissue or an organ mentioned in the study as the target or test system for the biomaterial object or medical device.	MeSH: Tissue, Body Location or organ Body part, organ or organ component UBERON: Tissue PREMEDONTO:Body Part, Organ, Organ System

Table: Continued

	Annotation category	Definition	Resource and imported classes
14	Adverse Effects	An unfavourable or unintended disease, sign, or symptom (including an abnormal laboratory finding) that is temporally associated with the use of a medical device or biomaterial.	DEB: Adverse Effects MeSH: Pathologic Function
15	Research Technique	The reported laboratory technique or instrument used in an experimental study.	MeSH: Laboratory Procedure, Molecular Biology Research Technique DEB: Research Technique NCIT: Research Technique NPO: Instrument IOBC: Microscope OBI: Assay EDAM: Imaging, Sample preparation ONTOTOXNUC: Outil
16	Effect On Biological System	The effect associated with manufactured object in a specific test system (cells, tissue or organism).	DEB: Effect On Biological System
17	Study Type	The study set up, such as in vitro, in vivo, or clinical.	DEB: Study Type