Time to kick-start text mining for biomaterials

Osnat Hakimi[1,2], **Martin Krallinger**[2], **and Maria-Pau Ginebra**[1,3]

[1]Universitat Politècnica de Catalunya, Department of Materials Science and Engineering, Barcelona, Spain
[2]Barcelona Supercomputing Center, Barcelona, Spain
[3]Barcelona Institute of Technology, Institute for Bioengineering of Catalonia, Barcelona, Spain

This is a post-peer-review, pre-copyedit version of an article published in Nature Reviews Materials. The final authenticated version is available online at: https://doi.org/10.1038/s41578-020-0215-z

Rapidly expanding biomaterials data is challenging to organize. Text mining systems are powerful tools that automatically extract and integrate information in large textual collections. As text mining leaps forward by leveraging deep-learning approaches, it is time to address the most pressing biomaterials information and data processing needs.

Designing a new implant or engineering a tissue is a complex venture. It involves intricate processes, such as cell–material interactions or *in-vivo* biomaterial degradation, and the generation of vast, heterogeneous and complex data. The majority of this data is published in the form of research articles or patents, in which information is presented in free text or figures, and is therefore 'unstructured', making it a challenging input for computational processing. In the absence of readily accessible biomaterials raw data repositories (bar a few exceptions such as cBiT [cbit.bmt.tue.nl/biomaterial/browse]), tasks such as comparing the performance of similar polymers or selecting a biomaterial for a specific medical application rely on manual sifting through research articles, patents, FDA reports and conference abstracts. The increasing volume of published results makes any exhaustive synthesis a daunting manual task, potentially incomplete and possibly biased.

Emerging computational tools, in particular artificial intelligence and machine-learning-based systems, are becoming essential for coping with this information overload. Among these, text mining systems automatically extract information from text documents into structured data sets. These systems serve to organize and link data and facilitate its analysis. As already demonstrated in biology, materials and healthcare research, text mining tools can be used not only to improve retrieval and search strategies, but also to classify, structure and group information. More advanced applications include generating predictive models[1] and discovering new knowledge. For example, a recent study on inorganic materials captured structure-property information from a large collection of abstracts, leading to the discovery of candidate thermoelectric materials not previously reported[2].

The potential for knowledge discovery in the field of biomaterials is vast. Scaffolds and implants have attributes, such as surface properties, chemical composition and hierarchical organization, that are thought to modulate cell and tissue reactions. Text mining methods could potentially capture implicit links between biomaterial attributes and biological effects, generating valuable data sets. The application of advanced analysis to such sets could unveil previously unconsidered associations between attributes and biological responses, aiding design and the discovery of new candidate biomaterials. For example, combining evidence from different tissues, disease models and materials, one could search for spatial attributes acting in synergy to affect the progression of biological processes such as inflammation or angiogenesis. With the advent of additive manufacturing, which allows for great control over scaffolds' architecture, such information could lead to innovative data-driven scaffolds and implants.

However, the application of text and data mining tools to biomaterials requires addressing several domain-specific challenges first. These include the highly heterogeneous nature of the data and the multidisciplinary and rapidly evolving language used in biomaterials publications.

Stumbling blocks

Limited biomaterials lexical resources

Text mining is useful for retrieving relevant documents from large collections and transforming them into well-structured information by automated processing. This transformation underpins the construction of large-scale knowledge repositories about entities (such as implants), their attributes (such as structural or chemical properties) and their relations. These repositories, known as knowledge bases, store structured data in a manner that allows efficient retrieval, exploration and analysis of information. To extract meaningful entities and attributes into repositories, most automated text mining systems include modules that recognize concepts (or terms) using predefined categories.

Simple examples of biomaterials terms are 'cartilage', 'PLGA' (poly(lactic-co-glycolic acid)) and 'electrospinning', which should be labeled (or 'annotated') with their respective categories 'tissue', 'polymer' and 'manufacturing technique'. Many concept recognition systems make use of term lists, which organize sets of terms together with their definitions, synonyms, abbreviations and unique concept identifiers, thus enabling the recognition of concepts directly in text, a process referred to as named entity recognition.

One hurdle for the robust recognition of biomaterials concepts relates to the limited number of available biomaterials-specific lexical and semantic resources. Currently, biomaterial assets are few, relatively small and limited in scope (Table 1). In addition, there is a lack of textual data manually labeled by experts (known as annotated corpora). This represents a major bottleneck to machine-learning-based natural language processing systems, which often rely on annotated corpora for building predictive language models (which learn to automatically recognize concepts in text).

Knowledge representation of implants

Central to biomaterials research are scaffolds and implants, and much research is concerned with the design, manufacture and evaluation of 3D objects. Automatically finding such entities in text is far from trivial. These complex manufactured objects have no formal naming convention, and are frequently referred to with subjective emphasis on any of their attributes. A research article may describe the same object as a '3D-printed polymer scaffold' or a 'degradable patch for tendon repair'. The absence of a coherent object naming makes data integration and comparisons of biomaterials and their performance a demanding task. An additional challenge is the distinction between multiple objects evaluated comparatively within the same study, which is common in biomaterials research. Coupled to that is the task of defining, normalizing and representing attributes such as size, shape, porosity or surface topography. These attributes often hold the key to data pooling and interpretation. For example, otherwise identical scaffolds with different pore sizes may elicit different biological responses. Discovering attributes in text documents and linking them to the correct implant or scaffold is a challenging task.

A tower of Babel of multidisciplinary language

The intrinsic interdisciplinary nature of biomaterials research, which combines procedures and terms from medicine, chemistry, biology and engineering, is another challenge. The correct identification of concepts requires the careful selection, integration and linking of information from multiple disciplines. Useful resources include databases and ontologies of chemicals, diseases, adverse events or cellular processes (Table 1). Regrettably, the combination of lexical and semantic resources from multiple disciplines often leads to noisy, over-annotated texts with ambiguous labels. Some terms, such as abbreviations and acronyms, can have multiple meanings depending on the context. For example, PLA, commonly used to abbreviate aliphatic polyester poly-lactic-acid, is also used to refer to percutaneous local ablation and left arterial pressure in the unified medical language system (UMLS [nlm.nih.gov/research/umls/index.html]). Prioritizing the most suitable categories is a demanding task, requiring domain expertise. Moreover, relevant semantic assets (such as vocabularies and ontologies) are scattered across different research silos, each with their own hubs and formats, increasing the burden of finding and combining resources.

The way forward

eveloping terminologies

A first priority should be the expansion and improvement of semantic resources. Biomaterials research could greatly benefit from the introduction of controlled vocabularies, such as the gene ontology (GO [http://geneontology.org/]) or the systematized nomenclature of medicine – clinical terms (SNOMED-CT [snomed.org]), two mature resources that are cardinal assets for biological and clinical data analysis, respectively.

Controlled vocabularies are organized collections of concepts and of the formal relations between them. Importantly, they require the development of a naming convention, building upon a minimal community consensus. The global medical device nomenclature (GMDN [gmdnagency.org]), for example, is a naming system developed to exchange medical device information and support patient safety, but it does not cover experimental biomaterials, and is not open source. To enable open research, we should support the creation and expansion of open biomaterials semantic resources. The type of assets developed should be prioritized based on the most pressing needs of the domain, and involve community engagement. Initiatives such as the devices, experimental scaffolds and biomaterials ontology (DEB[3] https://github.com/ProjectDebbie/Ontology_DEB)[3] include mechanisms for community participation in the expansion of the ontology. The European Materials Modeling Council [https://emmc.eu/] is also backing the development of open semantic resources for material manufacturing and characterization, but to ensure biomaterials are covered in these ontologies, it is up to biomaterials experts to get involved.

Learning from related disciplines

Text mining tools for biological[4], chemical[5] and to some extent materials[6] data extraction are already in use, but need to be assessed in terms of their performance in the biomaterials domain, where they could be adapted and/or guide the design of biomaterials-specific tools. Building a catalogue of tools and understanding the strength and drawbacks of each when applied to biomaterials will accelerate the development of specialized resources. Examples of relevant tools are summarized in Table 1, and include concept annotation tools such as TaggerOne [ncbi.nlm.nih.gov/research/bionlp/tools/taggerone/], ChemSpot [informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/chemspot] and MetaMap [metamap.nlm.nih.gov/].

Text-mining systems could also be used to classify documents by employing supervised text classifiers or relevance-feedback-based information retrieval systems. This is particularly useful given that much of the discipline's open literature is scattered across resources.

Machine learning approaches

Recent progress in artificial intelligence and language technologies are promising for advancing systematic data extraction. Machine-learning-based text processing can cope with two important limitations of traditional terminology-based text processing: ambiguity and variability of language. Terminological resources, albeit useful for organizing information, can not cover all possible typographical, spelling, word order variants and synonyms for a large collection of terms. Artificial intelligence tools can obtain results of quality close to that achieved by humans when sufficiently large, high-quality manually labeled training examples are available. For Biology, chemistry and materials science, there are already openly available labeled corpora developed for information extraction tasks (Table 1). Supporting biomaterials text corpus construction initiatives through the development of labeled data sets would enable the application of machine-learning algorithms to biomaterials.

Nonetheless, due to the limited number of high-quality manually annotated corpora, data-science experts have started exploiting computational-intensive deep learning and neural networks language models to build on unlabeled data collections[7]. The rapid progress in this field could be harnessed to solve some of the key challenges faced by biomaterials text mining, such as expanding vocabularies and performing automatic biomaterial concept detection. Useful applications of language models include open ontology learning, the process of extracting and organizing concepts into a new logical ontology in an unsupervised way, and language models for named entity recognition. These include SciBERT [https://github.com/allenai/scibert] and BioBERT [https://github.com/dmis-lab/biobert], both trained on very large corpora specifically for scientific text mining. There are fantastic proof-of-concept examples of deep-learning instruments in biology and material science, for tasks ranging from biological entity labeling[8] to materials discovery[6]and property prediction[9]. Even so, their application and validation in biomaterials research will require substantial fine-tuning and domain adaptation, and labeled data sets for evaluation purposes.

Why we need structured data

In the age of information, the importance of investing in open, structured data is increasingly evident. The more accurate and complete the data is, the more valuable it is for downstream analyses and as input for modeling and simulation efforts. Moreover, resources such as knowledge bases can enable the discovery of new knowledge by combining information derived from disparate data sets and providing the infrastructure for automatic inference systems. Discovered knowledge could range from implant design features for a specific medical application to the probability of clinical adverse events for a given design. In the long term, prediction tools could reduce the burden of *in-vitro* and *in-vivo* testing. Ultimately, new algorithmic instruments should enable evidence-based materials and design features selection, leading to improved safety and performance.

Outlook

Although biomaterials text mining is still in its infancy, some key efforts can already be highlighted. Several domain-specific publicly available ontologies already exist, including the bone and cartilage tissue engineering ontology (BCTEO [bioportal.bioontology.org/ontologies/BCTEO]), the nanoparticle ontology (NPO [bioportal.bioontology.org/ontologies/NPO]), and the devices, experimental scaffolds and biomaterials ontology (DEB `https://github.com/ProjectDebbie/Ontology_DEB`).

Data extraction endeavors include the development of a comprehensive database of polymer properties using a hybrid human–computer approach[10], and an automated text-mining pipeline extracting biomaterials information from PubMed abstracts into an open-access database (DEBBIE [https://github.com/ProjectDebbie]). More are needed, and with funding bodies and policymakers already encouraging data exploitation, the time is right to kick-start additional efforts.

As a research community, if we wish to transform biomaterials research into a data-driven endeavor, we should first strive to improve data sharing in our domain. Creating biomaterials data-sharing platforms, community-endorsed terminologies and FAIR (findable, accessible, interoperable and reusable) tools will require a concerted effort, involving a collaborative interaction and the continuous engagement of end users.

But who should spearhead the development of these tools? The application of general-purpose machine-learning and natural-language processing techniques to biomaterials-specific tasks requires not only considerable tailoring, but also a good understanding of the intricate processes involved. Thus, biomaterials scientists willing to acquire and incorporate data-mining, text-mining and computational skills might be in the best position to guide such efforts. The future of data-driven biomaterials research may depend on those daring to cross discipline boundaries and face the unstructured data.

Links

SciBERT https://github.com/allenai/scibert
BioBERT https://github.com/dmis-lab/biobert
DEBBIE https://github.com/ProjectDebbie

cBiT cbit.bmt.tue.nl/biomaterial/browse
DEB [https://github.com/ProjectDebbie/Ontology$_DEB$]
$BCTEO$[$bioportal.bioontology.org/ontologies/BCTEO$]
NPO[$bioportal.bioontology.org/ontologies/NPO$]
$SNOMED-CT$[$snomed.org$]
$UMLS$[$nlm.nih.gov/research/umls/index.html$]
$GMDN$[$gmdnagency.org$]
$TaggerOne$[$ncbi.nlm.nih.gov/research/bionlp/tools/taggerone/$]
$ChemSpot$[$informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/chemspot$]
$MetaMap$[$metamap.nlm.nih.gov/$]
GO[$http://geneontology.org/$]
$EuropeanMaterialsModelingCouncil$[$https://emmc.eu/$]

Box 1: Useful tools and resources for starting biomaterials text mining

Resource type	Name	URL	Accessibility
Biomaterials repositories	The Compendium for Biomaterial Transcriptomics (cBiT)	cbit.bmt.tue.nl/biomaterial/browse	Open
General repositories with biomaterials subsets	ArXiv is an open-access archive for pre-prints in multiple fields	arxiv.org/	Open
	Core, collection of open access research papers	core.ac.uk/	Open
	The United State Patent and Trademark Office (USPTO)	bulkdata.uspto.gov/	Open
Biomaterials ontologies	Devices, Experimental scaffolds and Biomaterials Ontology (DEB)	bioportal.bioontology.org/ontologies/DEB	Open
	Bone and Cartilage Tissue Engineering Ontology (BCTEO)	bioportal.bioontology.org/ontologies/BCTEO	Open
	NanoParticle Ontology (NPO)	bioportal.bioontology.org/ontologies/NPO	Open
Ontologies from related disciplines	The European Materials Modelling Ontology (EMMO)	github.com/emmo-repo/EMMO	Open
	Chemical Entities of Biological Interest Ontology (CHEBI)	bioportal.bioontology.org/ontologies/CHEBI	Open
	Chemical Methods Ontology (CHMO)	bioportal.bioontology.org/ontologies/CHMO	Open
Terminologies from related disciplines	Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)	snomed.org/	License needed
	Unified medical language system (UMLS)	nlm.nih.gov/research/umls/index.html	License needed
	Global Medical Device Nomenclature (GMDN)	gmdnagency.org/	License needed
Concept recognition resources	PubTator Central (PTC) Biomedical Named Entity Recognition system	ncbi.nlm.nih.gov/research/pubtator/	Open
	TaggerOne identifies biomedical concepts such as diseases,chemicals	ncbi.nlm.nih.gov/research/bionlp/tools/taggerone/	Open
	ChemSpot identifies mentions of chemicals in texts	informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/chemspot	Open
	ChemicalTagger tags and parses chemistry experimental sections	github.com/BlueObelisk/chemicaltagger	Open
	AnatomyTagger is an entity mention tagger for anatomical entities	nactem.ac.uk/anatomytagger/	Open
	MetaMap maps biomedical text to the UMLS Metathesaurus	metamap.nlm.nih.gov/	Open
	ScpaCy library for advanced Natural Language Processing in Python.	spacy.io/	Open
Annotated corpora	GENIA (biology, medicine)	geniaproject.org/	Open
	Biocreative-ii-gm (biology)	github.com/openbiocorpora/biocreative-ii-gm	Open
	CHEMDNER (Chemistry)	biocreative.bioinformatics.udel.edu/resources/ biocreative-iv/chemdner-corpus/	Open
	The Materials Science Procedural Text Corpus (Materials)	github.com/olivettigroup/annotated-materials-syntheses	Open

References

1. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* (2018).
2. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* (2019).
3. Hakimi, O. *et al.* The devices, experimental scaffolds, and biomaterials ontology (deb): A tool for mapping, annotation, and analysis of biomaterials' data. *Adv. Funct. Mater.* (2020).
4. Hirschman, L. *et al.* Text mining for the biocuration workflow. *Database: The J. Biol. Databases Curation* (2012).
5. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J. & Valencia, A. Information retrieval and text mining technologies for chemistry. *Chem. Rev.* (2017).
6. Isayev, O. Text mining facilitates materials discovery. *Nature* (2019).
7. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* (2015).

8. Xuan, W. *et al.* Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics* (2019).

9. Jha, D. *et al.* Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat. Commun.* (2019).

10. Tchoua, R. B. *et al.* A hybrid human-computer approach to the extraction of scientific facts from the literature. *Procedia Comput. Sci.* (2016).