SciBERT-based Semantification of Bioassays in the Open Research Knowledge Graph

Marco Anteghini[1,2][0000−0003−2794−3853] ⋆, Jennifer
D'Souza[3][0000−0002−6616−9509], Vitor A.P. Martins dos
Santos[1,2][0000−0002−2352−9017], and Sören Auer[3][0000−0002−0698−2864]

[1] Lifeglimmer GmbH, Markelstr. 38, 12163 Berlin, Germany
[2] Wageningen University & Research, Laboratory of Systems & Synthetic Biology,
Stippeneng 4, 6708 WE, Wageningen, The Netherlands
{anteghini,vds}@lifeglimmer.com
[3] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{jennifer.dsouza,soeren.auer}@tib.eu

Abstract. As a novel contribution to the problem of *semantifying biological assays*, in this paper, we propose a neural-network-based approach to automatically semantify, thereby structure, unstructured bioassay text descriptions. Experimental evaluations, to this end, show promise as the neural-based semantification significantly outperforms a naive frequency-based baseline approach. Specifically, the neural method attains 72% $F1$ versus 47% $F1$ from the frequency-based method.

The work in this paper aligns with the present cutting-edge trend of the scholarly knowledge digitalization impetus which aim to convert the long-standing document-based format of scholarly content into knowledge graphs (KG). To this end, our selected data domain of bioassays are a prime candidate for structuring into KGs.

Keywords: Open Science Graphs · Bioassays · Machine Learning

1 Introduction

Biological assays are defined as standard biochemical test procedures used to determine the concentration or potency of a stimulus (physical, chemical, or biological) by its effect on living cells or tissues [3,4].

In the context of the current Covid-19 pandemic, bioassays are critical, for example, for vaccine development. They reveal the functional and biologically relevant immunological responses that correlate with vaccine efficacy. However, massive volumes of bioassays are being produced and researchers are inundated with this information. Apart from their sheer quantity, bioassay diversity presents enormous challenges to organizing, standardizing, and integrating the data with the goal to maximize their scientific and ultimately their public health impact as the screening results are carried forward into drug development programs.

Against this broad societal application setting, we present a solution as a step in the easier knowledge acquisition of bioassays for researchers: *the neural-based automated structuring of unstructured, non-standardized bioassays based on the standardized BioAssay Ontology (BAO) [7]*. Bioassays, until their recent semantification in an expert-annotated dataset [2,5,6] based on the BAO, were published in the form of unstructured text. Integrating their semantified counterpart in a KG facilitates their advanced computational processing. E.g., bioassays can be easily compared across their key properties, viz. Target, Perturbagen, Participants, and Detection Technology, captured as KG nodes and links. Nonetheless, the fine-grained semantification of bioassays as a manual task is a costly and time-intensive endeavor. Their automated semantification not only alleviates the costly manual task, but potentially makes it possible to rapidly semantify this data in large volumes. Herein, we present our novel SciBERT-based [1] neural BAO [7] bioassay semantification system.

2   Method

For automated bioassay semantification, we carry out the supervised machine learning of semantic statements (i.e., subject-predicate-object triples) based on the BioAssay Ontology (BAO) [7] for a given unstructured bioassay description. The code for our method is publicly available at: https://github.com/MarcoAnteghini/SciBERT-bioassays_ORKG.

2.1   Dataset

Our dataset for learning comprises an expert manually annotated collection of 983 semantified bioasssays [5,6]. In the data, each assay has between 5 and 92 semantic statements at an average of 53. To better reflect the data, we show example annotations in Table 1 for a selected bioassay.

HAS ASSAY FORMAT $\rightarrow$ BIOCHEMICAL FORMAT
HAS ASSAY FORMAT $\rightarrow$ PROTEIN FORMAT
HAS ASSAY FORMAT $\rightarrow$ SINGLE PROTEIN FORMAT
ASSAY MEASUREMENT TYPE $\rightarrow$ ENDPOINT ASSAY

Table 1: Four example semantic statement annotations (from 50 total) for PubChem Assay ID 346. Note, these statements are triples with subject "bioassay."

2.2   Problem Formulation

The dataset can be formalized as follows. Let b be a bioassay from the assays dataset B. Each b_i is annotated with an annotation sequence as_i such that $as_i \in S$, where S is a set of all possible semantic statements seen in the training dataset. Specifically,

$as_i = \{s_1, s_2, s_3, ..., s_k\}$, such that s_x is a semantic statement $\in S$; as_i has k different statements. In general, annotation sequences are of varying lengths. The dataset we use has $|S| = 1756$ unique statements (after filtering for non-informative ones).

In the supervised task, the input data instance corresponds to a pair $(b, s; c)$ where $c \in \{true, false\}$ is the classification label. Thus, specifically, our semantification problem is formulated as a *binary classification task*. (b, s) is *true* if $s \in b$'s annotation sequence (as), else *false*. Where *false* instances are formed by pairing b with any other label not in the annotation sequence as of b. As an aggregate, the semantification of each bioassay is a multi-label, multi-class classification problem which we have broken up into binary classification decisions.

Intuitively, our task formulation is meaningful because it emulates the way the human expert annotates the data. Basically, the expert, from their memory of all semantic statements S, simply assigns s to a given b if they deem it as *true*; irrelevant statements are not considered, thus implicitly deemed *false*.

2.3 SciBERT-based Machine Learning

Our machine learning system is the state-of-the-art, bidirectional transformer-based SciBERT [1], pre-trained on millions of scientific articles. In each data instance $(b, s; c)$, the classifier input representation for the pair 'b, s' is the standard SciBERT format, treating them as sentence pairs separated by the special [SEP] token; the special classification token ([CLS]) remains the first token of every instance. Its final hidden state is used as the aggregate sequence representation for classification tasks fed into a linear classification layer.

3 Experiments

3.1 Experimental Setup

For robust evaluations, we perform 3-fold cross validation (2:1 train-test split). In each fold experiment, training data contains roughly 655 bioassays and the remaining 328 bioassays are used for testing, where the test assays are unique across the folds. Standard precision (P), recall (R), and f-score $(F1)$ metrics are used. We refer the reader to the SciBERT paper [1] for hyperparameter details. Finally, we have an additional parameter: *false instances per bioassay*. They are varied between 100 to 300, in increments of 10, to obtain an optimal model.

3.2 Results and Discussion

Our results are depicted in Tables 2 and 3. And we examine the **RQ***: can advanced neural technologies be leveraged to automatically semantify bioassays?* We find that the cumulative obtainable $F1$ by the SciBERT classifier out-of-the-box is 0.72 (bold in Table 3)—significantly higher than 0.47 from a naive frequency-based semantification approach. Furthermore, the difference of the neural approach from the frequency method is clearly evident in the hit-and-miss illustration in Fig 1. The top thin neck of the curve in Fig 1(a) indicates that the neural approach, for most bioassays, had faster *true* semantic statement hits among its top-scoring predictions. Thus, answering **RQ**, neural technologies can indeed perform reliable semantification of bioassays. They are also practically efficient, since, given the 1756 unique statements considered as labels, each test assay is semantified at a rate of 4 seconds.

$false$ labels	P	R	$F1$
100	0.517	0.968	0.674
...	...	...	...
160	0.549	0.931	0.688
170	**0.600**	**0.939**	**0.729**
180	0.573	0.945	0.711
...	...	...	...
300	0.471	0.674	0.551

Table 2: Bioassay semantification results from five training optimization with different $false$ classification instances (full table in appendix)

test set	P	R	$F1$
1st fold	0.600	0.939	0.729
2nd fold	0.573	0.956	0.713
3rd fold	0.589	0.936	0.719
$Avg.$	**0.588**	**0.944**	**0.720**

Table 3: Automatic bioassay semantification results from 3-fold cross validation with the optimal number of $false$ classification labels (170).

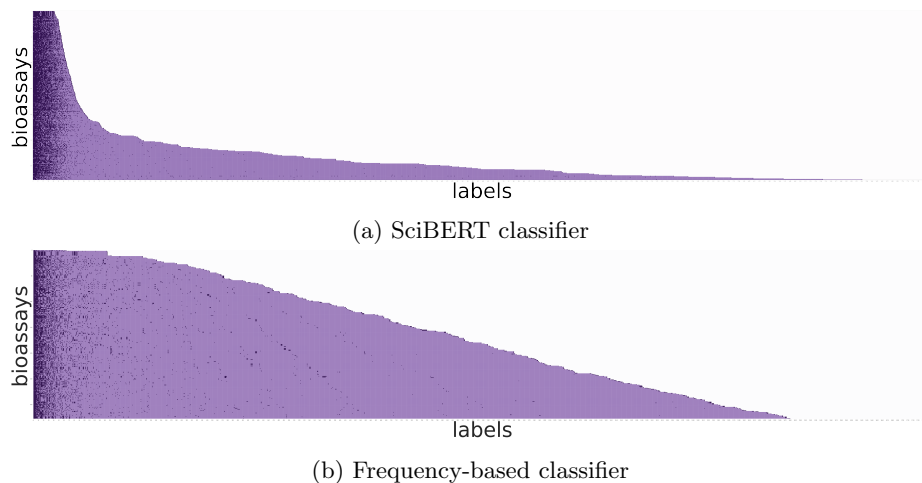(a) SciBERT classifier

(b) Frequency-based classifier

Fig. 1: Hit-and-miss Plots for semantifying bioassays by SciBERT vs. a naive frequency-based approach. Black dot is a hit; purple dot is a miss. For each assay, after all the *true* statements are predicted, the remaining dots are white.

4   Conclusion

The discovery of cures during pandemics such as Covid-19 can be greatly expedited if scientists are given intelligent information access tools, and our work toward automatically semantifying bioassays are a step in this direction. We refer the reader to the Appendix for an illustrated use case of semantified bioassays data in next-generation digital libraries.

References

1. Beltagy, I., Lo, K., Cohan, A.: Scibert: Pretrained language model for scientific text. In: EMNLP (2019)
2. Clark, A.M., Bunin, B.A., Litterman, N.K., Schürer, S.C., Visser, U.: Fast and accurate semantic annotation of bioassays exploiting a hybrid of machine learning and user confirmation. PeerJ **2**, e524 (2014)
3. Hoskins, W.M., Craig, R.: Uses of bioassay in entomology. Annual review of entomology **7**(1), 437–464 (1962)
4. Irwin, J.: Statistical method in biological assay. Nature **172**(4386), 925–926 (1953)
5. Schürer, S.C., Vempati, U., Smith, R., Southern, M., Lemmon, V.: Bioassay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. Journal of biomolecular screening **16**(4), 415–426 (2011)
6. Vempati, U.D., Przydzial, M.J., Chung, C., Abeyruwan, S., Mir, A., Sakurai, K., Visser, U., Lemmon, V.P., Schürer, S.C.: Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the bioassay ontology (bao). PloS one **7**(11), e49198 (2012)
7. Visser, U., Abeyruwan, S., Vempati, U., Smith, R.P., Lemmon, V., Schürer, S.C.: Bioassay ontology (bao): a semantic description of bioassays and high-throughput screening results. BMC bioinformatics **12**(1), 257 (2011)

SciBERT-based Semantification of Bioassays in the Open Research Knowledge Graph

Marco Anteghini

Jennifer D'Souza

Vítor Martins dos Santos

Sören Auer

LifeGlimmer GmbH

Wageningen University

TIB Leibniz Information Centre for Science and Technology, Hannover

Introduction

- Comparisons across several bioassays is a challenging problem in today's context-dependent structure.

- Semantically structured scholarly knowledge can alleviate the problem.

- We do not have automatic methods to automatically semantify the data.
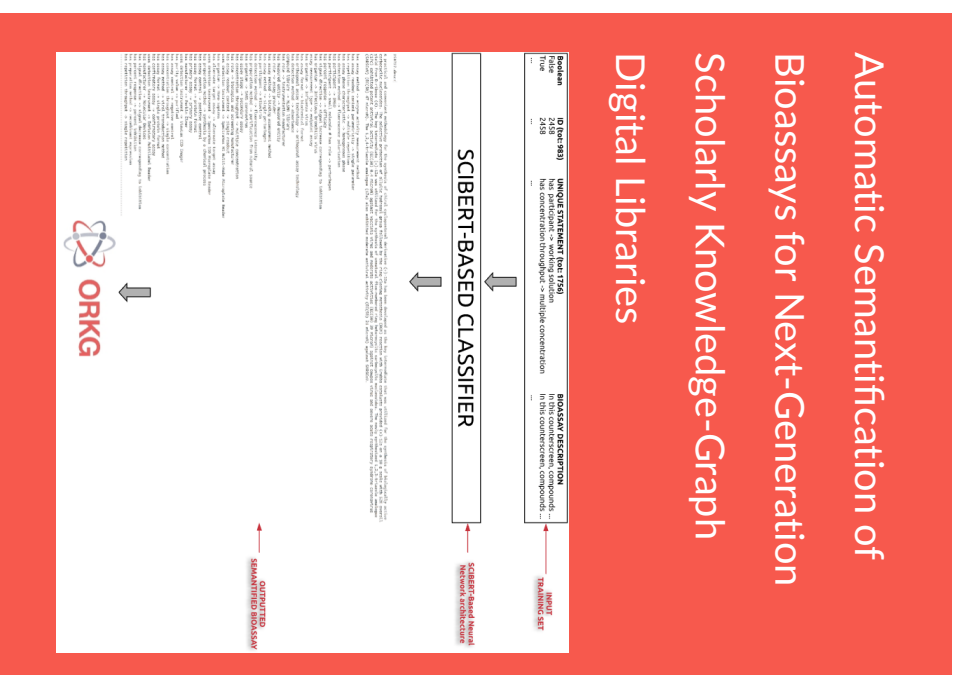
Performance

3-Fold CV on a dataset of 983 bioassays

test set	P	R	$F1$
1st fold	0.600	0.939	0.729
2nd fold	0.573	0.956	0.713
3rd fold	0.589	0.936	0.719
Avg.	**0.588**	**0.944**	**0.720**

Conclusion

We propose a neural-based semantification approach for text-based bioassays to be integrated as a software module in next-generation digital libraries such as the Open Research Knowledge Graph.

Automatic Semantification of Bioassays for Next-Generation Scholarly Knowledge-Graph Digital Libraries

SCIBERT-BASED CLASSIFIER

ORKG

Our Dataset

We based our analyses on the dataset provided by Clark et al., containing 983 bioassays downloaded from PubChem. For each bioassay, we filtered the non informative labels including the ones with values as string literals remaining with 1756 labels. In our experiments, we varied between 100 to 300 false labels to obtain an optimal model.

Hit-&-Miss Plots of Semantifying Bioassays by Frequency-based Classifier (top) vs. SciBERT-based classifier (bottom)
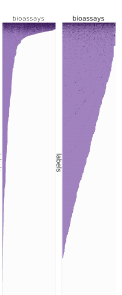
Survey Comparisons of Structured Bioassays

An automatically generated comparisons of semantified bioassays in the ORKG digital library (DL). Full graph at www.orkg.org

References

Clark, A., Bunin, B., Litterman, N., Schurer, S., Visser, U.: Fast and accurate semantic annotation of bioassays exploiting a hybrid of machine learning and user confirmation. PeerJ 2, e524 (08 2014).

LifeGlimmer

A   Unique statements (labels) distribution

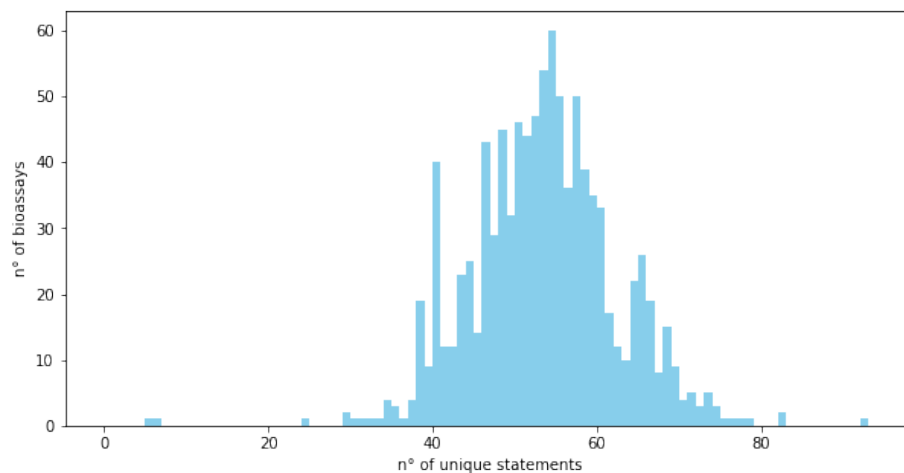Each bioassays present on average 53 labels. The distribution is visible in Figure 2

Fig. 2: Unique statements distribution

B   Snapshot of Semantified Bioassay in the Open Research Knowledge Graph

Figure 3 is an instance of integrating one semantified bioassay in the ORKG DL. This bioassay was semantified on eight semantic statements based on the BAO. Integrating machine actionable graphs of bioassays is essential for the ORKG DL to automatically compute the tabulated comparison surveys of several bioassays as shown in Figure 4 in the next section.

C   Application: Comparisons of Bioassays in ORKG

Next generation DLs target semantified scholarly knowledge. The ORKG with the semantified bioassays integrated, automatically computes their survey comparisons depending on how many of the machine-actionable assays were selected to be compared by the user. Such tools must be available to scientists to assist them in such massive knowledge ingestion scenarios to quickly grasp the scholarly knowledge highlights fostering faster progress with discoveries.
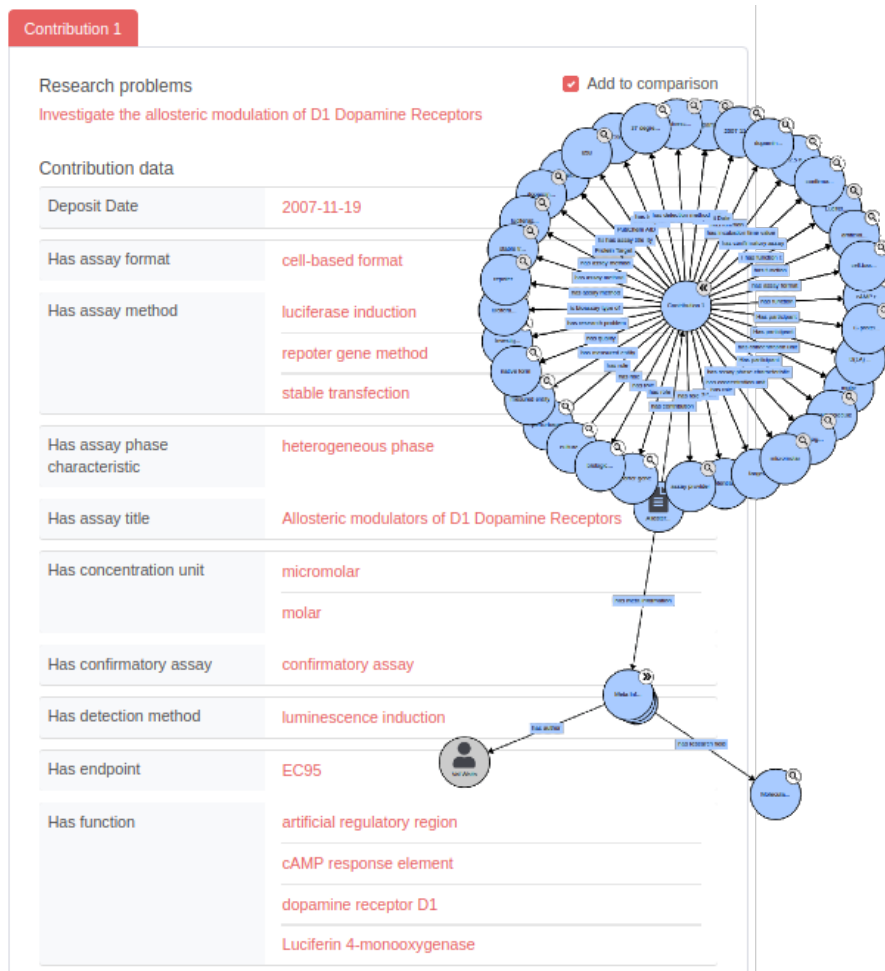
Fig. 3: An ORKG representation of a semantified Bioassay with an overlayed graph view of the assay. Accessible at: https://www.orkg.org/orkg/paper/R48146/R48147

Properties	SAR analysis of tumor necrosis factor alpha (TNF-alpha) induced IL-8 secretion in MCF-7/NOD1 cells Contribution 1	Fluorescence Cell-Based Screen to Identify Inhibitors of Calcium Ionophore-Mediated P-Selectin Induction on Platelets Contribution 1	Allosteric Modulators of D1 Receptors: Dose-dependent Assay Contribution 1
Has detection method	absorbance	flow cytometry	luminescence induction
Has endpoint	IC50	Empty	EC95
Has incubation time value	30 min	20 min	2.5 h
Has measured entity	TNF-alpha measured entity	measured entity	Allosteric modulators of D1 Dopamine Receptors Dopamine mesured entity
Has participant	Tween 20 secreted	A23187 DMSO P-selectin P-selectin glycoprotein ligand 1 small molecule	D(1A) dopamine receptor G protein coupled receptor Luciferin 4-monooxygenase small molecule
Has quality	native form	Empty	native form
Has research problem	Selectivity of hits emerging from the NOD1 and NOD2 specific assays Usage of tumor necrosis factor alpha (TNF-alpha), for identification of hits specific to TNFa-modulated pathways	Inhibitors of patelet dense granule release	Investigate the allosteric modulation of D1 Dopamine Receptors
Has role	assay provider culture medium culture serum detergent instrumentation manufacturer perturbagen potentiator target	Substrate assay provider perturbagen target	assay provider biologics and screening manufacturer culture medium perturbagen potentiator reporter gene target
Has temperature value	25 degree celsius	Empty	37 degree celsius
Is bioassay type of	Empty	localization assay	luciferase reporter gene assay
Pubchem aid	2245	2522	858

Fig. 4: Automatically generated comparisons of semantified bioassays in the ORKG digital library (DL). Full graph https://www.orkg.org/orkg/comparison?contributions=R48195,R48179,R48147