Evaluation of Named Entity Recognition Systems to Improve Ontology Concept Annotation for Biomedical Knowledge Graphs

Sanya B. Taneja[1], Marcin P. Joachimiak[2], Harshad Hegde[2], William Baumgartner Jr.[3], J. Harry Caufield[2], Tiffany J. Callahan[4], Christopher J. Mungall[2], Richard D. Boyce[1]

[1]University of Pittsburgh, [2]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, [3]University of Colorado Anschutz Medical Campus, [4]Columbia University

1 INTRODUCTION

A biomedical knowledge graph (KG) consists of nodes and edges, where nodes are biomedical entities such as genes, and edges represent relations between the nodes. Biomedical KGs can be constructed from the combination of ontologies, scientific literature, human curated data, and information present in databases. Historically, large biomedical KGs created from free text or structured, non-standardized data have been high throughput and scalable, which makes validation largely intractable for humans (for example, KGs derived from automated processes like KG-COVID-19[1]). Named Entity Recognition (NER) tools are commonly used to map entities to standardized identifiers in reference knowledge like ontologies and databases for integration in KGs. NER tools facilitate inclusion of more data sources in KGs as well as accelerate human curation. However, they are known to be imperfect and the errors they introduce can systematically affect downstream applications. Formal ways to evaluate the effects of NER on biomedical KG construction and downstream applications of KGs are not readily available. In this study, we evaluate NER systems for KG development, with a focus on using them to map entities to biomedical ontologies.

2 METHODS

Our initial dataset consists of biomedical entities extracted from full texts of PubMed-indexed articles related to green tea (*Camellia sinensis*). All entities were either subjects or objects from subject-relation-object triples that were integrated into a novel biomedical KG to discover hypotheses for natural product-drug interactions. We applied two NER tools - BioPortal Annotator[2] and the OntoRunNER OGER++ wrapper[3] - to map the entities to the same 13 Open Biological and Biomedical (OBO) Foundry ontologies. We then manually reviewed the mappings to evaluate the performance of the NER tools and applied summary statistics.

3 RESULTS AND DISCUSSION

Table 1 reports the results of mapping 810 entities to biomedical ontologies using the BioPortal Annotator and the OntoRunNER OGER++ wrapper, along with the average candidate matches for the entities produced by both tools. While the OntoRunNER wrapper produces a larger set of candidate matches for the entities requiring more manual review, it performs better than the BioPortal Annotator for ontology concept annotation. For example, for the entity 'cytochromes b5', BioPortal Annotator produced incorrect mappings 'Cytochrome' (GO_0045155) and 'Cytochromes' (CHEBI_4056) but the OntoRunNER wrapper mapped to 4 candidates for 'cytochrome b5' (CHEBI_38553, GO_0009464, GO_0005489, PR_000006078). Through NER methods, we can incorporate the latest knowledge from free texts or structured, non-standardized data before it has been formally distributed in curated knowledge resources. We hypothesize that accurately mapping entities in the KG ultimately influences the accuracy of entailed inferences. To this end, we are currently measuring the differential effects on KGs and their downstream applications with graph analysis, embedding similarity, more NER tools, and data sources.

Table 1: Results of NER for 810 biomedical entities

	BIOPORTAL ANNOTATOR	ONTORUNNER
Mapped Entities	729 (90%)	735 (91%)
Avg. Candidate Matches	1.89	4.72
Entities Manually Reviewed	494	735
Correct Mappings from Manual Review	296 (60%)	559 (76%)

REFERENCES

Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, et al. KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. Patterns. 2021 Jan;2(1):100155.

Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011 Jul;39

ontoRunNER [Internet]. Monarch Initiative; 2021. Available from: https://github.com/monarch-initiative/ontorunner