Improving the Evaluation of NLP Approaches for Scientific Text Annotation with Ontology Embedding-Based Semantic Similarity Metrics

Pratik Devkota
Informatics and Analytics
University of North Carolina at Greensboro
North Carolina, USA
p_devkota@uncg.edu

Somya Mohanty
Artificial Intelligence Group
United Health
USA
smohanty@unitedhealthgroup.com

Prashanti Manda
Informatics and Analytics
University of North Carolina at Greensboro
USA
p_manda@uncg.edu

Abstract

Ontologies are widely used to represent data in a variety of scientific domains, including biology, physics, and geography. Ontology curation and annotation, the process of reading scientific text and associating words and phrases with appropriate ontology concepts, is essential for this representation.

Natural language processing (NLP) techniques powered by deep learning have recently become prominent in the task of ontology annotation. However, traditional metrics of accuracy, such as precision and recall, cannot be used to evaluate the accuracy of these methods, as their output is ontology concepts, not independent entities.

Semantic similarity metrics offer the capability of estimating partial accuracy and have been used in recent work to evaluate NLP methods for ontology annotation. Here, we present robust semantic similarity metrics created through the use of ontology embeddings. We tested our metrics using gold standard data pertaining to evolutionary biology created by scientists in the Phenoscape project and show that they outperform traditional semantic similarity metrics, offering a more robust and accurate assessment of NLP approaches designed for ontology annotation.

1 Introduction

As several fields of science entered the data-intensive era, ontologies grew increasingly popular for consistent, machine-readable representation of scientific data (Stevens et al., 2000; Grimm, 2009; Consortium, 2006).

Ontologies proved to be critical in life sciences and particularly in biology to power large comparative analyses (Sahoo et al., 2006; Manda et al., 2015; McENTIRE, 2002). The Gene Ontology (GO) was created in 2003 (Smith et al., 2003) as a collaborative effort to facilitate meaningful descriptions of genes in a variety of organisms (Consortium, 2006). Ontologies enable standardization of biology by providing a common vocabulary for describing biological entities and their relationships. This helps to reduce ambiguity and improve communication between researchers. They enable data integration and interoperability (Consortium, 2006). Ontologies can be used to link data from different sources, even if they are stored in different formats. This allows researchers to combine data from multiple experiments and databases to get a more complete picture of biological phenomena. Ontologies were built for supporting knowledge discovery and hypothesis generation (Ultsch and Lötsch, 2014). Ontologies can be used to reason about biological data and identify new patterns and relationships (Dahdul et al., 2015). This can help researchers to generate new hypotheses and develop new insights into biological systems. Finally, ontologies facilitate the development of new tools and applications to support biological research, such as tools for data annotation, knowledge discovery, and computational modeling (Manda et al., 2015).

While ontologies provide the necessary structure and concepts, the real benefits of ontologies can be reaped only when knowledge in scientific literature is represented using these ontologies through annotation (Devkota et al., 2023). One of the use cases for ontologies is gene annotation - Ontologies are used to annotate genes with information about their function, structure, and expression patterns (Consortium, 2012). This information can be used to better understand the role of genes in biological processes and diseases.

Creating annotations using ontologies started as a manual task undertaken by expert level human curators (Dahdul et al., 2015). Manual annotation of literature using ontologies is the process of identifying and tagging relevant biological entities and relationships in text with ontology terms. This is

a time-consuming and labor-intensive process, but it is essential for creating high-quality, machine-readable annotations that can be used to support a wide range of biological research tasks.

The manual annotation process typically involves the following steps:

- Read the literature. The annotator carefully reads the literature to identify relevant biological entities and relationships.
- Identify the appropriate ontologies. The annotator selects the ontologies that are most relevant to the topic of the literature.
- Annotate the text. The annotator tags the text with ontology terms to identify and describe the biological entities and relationships.
- Review and curate the annotations. The annotator reviews the annotations to ensure that they are accurate and consistent.

These GO annotations are the drivers for several crucial applications in biology such as gene function discovery, genome annotation, comparative genomics, functional genomics, and systems biology (Consortium, 2006; You et al., 2018; Manda et al., 2020). The ongoing generation of GO annotations as new literature is published is important to be able to conduct comparisons between species and to utilize rich genomic data for answering complex biological questions.

However, manual curation of scientific literature by human curators soon became an infeasible practice because it was tedious, slow, and unscalable to the rapid pace of scientific publishing (Dahdul et al., 2015). The bioinformatics community turned to text mining and natural language processing as a means to automate the process of ontology-based annotation. Ontology-based annotation is the process of reading (by a human or a machine) scientific literature and associating words in the text to appropriate ontology concepts (see Figure 1) (Cui et al., 2015).

The initial foray into automated annotation using natural language processing relied on lexical analysis and standard machine learning approaches (Devkota et al., 2022a, 2023). More recently, deep learning approaches have shown promise with text related applications (Lample et al., 2016; Boguslav et al., 2021; Casteleiro et al., 2018; Manda et al., 2020; Devkota et al., 2023, 2022a).

The past few years have witnessed an increasing focus on automated annotation of scientific literature and the development of sophisticated NLP approaches (Manda et al., 2018, 2020; Devkota et al., 2022b, 2023, 2022a). With this, a second and equally important problem came to light - robust and accurate methods for evaluating the success of these NLP methods (Dahdul et al., 2018). Evaluating the performance of NLP systems that are ontology-based is different from traditional NLP systems because of the possibility of partial accuracy.

Traditional information retrieval systems only consider whether the target information is retrieved (success) or not (failure). In contrast, ontology-based information retrieval systems allow for the possibility of partial success. This means that the system can still be successful even if it does not retrieve the exact target information, as long as it retrieves something that is semantically similar (Devkota et al., 2022a).

Ontology-based information retrieval systems are evaluated using three possibilities:

- Accurate retrieval: The system retrieves the exact target information.
- Inaccurate retrieval: The system retrieves information that is not semantically similar to the target information.
- Partially accurate retrieval: The system retrieves information that is semantically similar to the target information, but not identical.

The goal of our NLP systems is to maximize accurate retrieval rates and minimize inaccurate retrieval rates. In cases where complete accuracy is not achieved, the method aims to maximize partial accuracy.

Here is an example to illustrate the difference between traditional information retrieval and ontology-based information retrieval:

Imagine that you are searching for information about the concept of “dog”. A traditional information retrieval system would only consider whether it retrieves information that explicitly mentions the word “dog”. In contrast, an ontology-based information retrieval system would also consider retrieving information about other concepts that are semantically similar to “dog”, such as “canine” or “mammal”.

If an ontology-based information retrieval system retrieves information about “canine” when you

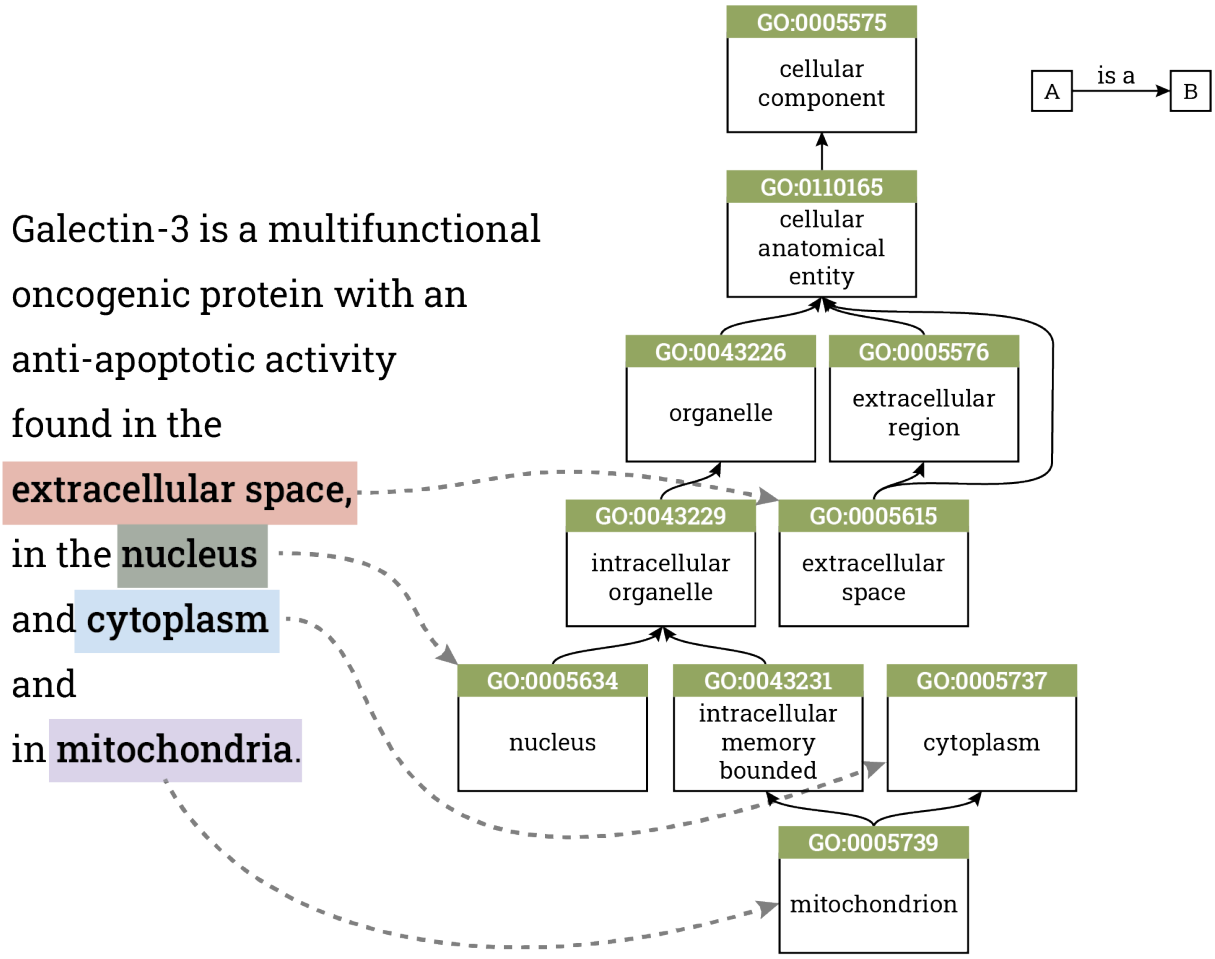

Figure 1: Illustration of ontology-based annotation. Appropriate words/phrases in scientific text are identified (shown via highlighted text) and are annotated to suitable concepts from an ontology.

are searching for information about "dog", this would be considered a partial success. The system has not retrieved the exact target information, but it has retrieved something that is semantically similar.

Ontology-based NLP systems need evaluation metrics that can account for accurate, inaccurate, and partial success (Dahdul et al., 2018). The answer to this problem can be found in semantic similarity metrics designed to estimate similarity (exact or partial) between ontology concepts (Pesquita et al., 2009). Several papers published in the recent past describing NLP techniques for automated annotation of literature have used semantic similarity metrics for evaluating accuracy (Manda et al., 2018, 2020; Devkota et al., 2022b,a, 2023). However, these existing semantic similarity measures fall short in the area of robustness since they have been shown to be quite susceptible to noise in the data (Manda and Vision, 2018).

In addition, in the case of ontology-based annotation, there is a great deal of information embedded in the ontology hierarchy and semantics that needs to be leveraged for accurate annotation (Devkota et al., 2023). Here, we present an improved semantic similarity metric for the evaluation of NLP approaches for automated ontology annotation. Our metric is based on graph embeddings created from ontologies (which are directed acyclic graphs). We compared our metric to existing metrics applied on a gold-standard dataset.

Graph embeddings are a type of machine learning technique that can be used to represent graphs in a low-dimensional vector space. This makes it possible to use machine learning algorithms to learn from and make predictions on graph data (Cai et al., 2018).

Graph embeddings are particularly useful for

NLP tasks, as many NLP problems can be represented as and are applied on graphs (Wang et al., 2014). Graph embeddings can be used to improve word representations, improve text classification, and for machine translation. Overall, graph embeddings are a powerful tool that can be used to improve the performance of a wide range of NLP tasks.

While we use data and background from the domain of biology, the methods presented here have wide scientific appeal and are generalizable. Ontologies are used in Chemistry, Physics, Astronomy, Geoscience, Materials science, Medicine, Psychology, and Social science (Strömert et al., 2022; Spencer, 1982; Hachem et al., 2011; Lesteven et al., 2007). Hence, the need for automated annotation methods and effective semantic similarity metrics for evaluation exist in all these domains.

2 Gold standard dataset

The dataset used here was developed for the Phenoscape project funded by the United States National Science Foundation (Manda et al., 2015; Dahdul et al., 2018). Three expert scientists were asked to independently annotate the same text related to evolutionary biology. A gold standard dataset was created by consensus of the three curators. The intuition behind this experimental set-up is that the annotations created by each curator are expected to be quite similar to the gold standard if not identical. This allows us to test different semantic similarity metrics by testing how well they retrieve similarity from known biologically similar annotations (Dahdul et al., 2018). The dataset is publicly accessible from Dahdul et. al. (Dahdul et al., 2018)

3 Ontology

The ontology used in the dataset described above is UBERON (Mungall et al., 2012). UBERON is a cross-species ontology that represents anatomical structures in animals (Mungall et al., 2012). This ontology is used to annotate biological data, such as gene expression data and protein interaction data, to enable data integration and knowledge discovery (Chandak et al., 2023). It has been used to develop tools and applications to support biological research and clinical practice (Zhao et al., 2020).

3.1 Creating Ontology Embeddings

A directed graph of the UBERON ontology is constructed using the ontology OWL file, using the subClassOf schema or is_a relationship. This graph is used as input to the Node2Vec algorithm to generate 128-dimensional embeddings for all concepts in the ontology. We refer to these embeddings as Ontology Embeddings (OE).

The Node2Vec algorithm (Grover and Leskovec, 2016)implements the following two steps:

1. Use a biased random walk to generate sentences (lists of ontology IDs) from the ontology graph.

2. This list of sentences generated from the random walk constitutes the corpus that represents the ontology. The Word2Vec algorithm is applied to this corpus to learn and calculate the embedding vector for each concept identifier in the ontology. The embeddings are learned through a deep learning model after hyper parameter tuning.

3.2 Semantic similarity computations

We compared 6 similarity metrics in this study - cosine similarity, manhattan distance, dot product, minkowski distance, euclidean distance, and jaccard similarity. Jaccard similarity is a widely used similarity metric that uses the ontology hierarchy and not ontology embeddings (Pesquita et al., 2009). The rest of the metrics use ontology embeddings generated in this study.

We use annotations from the three human-curated datasets as well as the Gold Standard (GS) data from Dahdul (Dahdul et al., 2018) for our semantic similarity comparisons. For each snippet of text, semantic similarity is computed between each of the three curators' annotation and the corresponding annotation in the GS in the dataset. Note that similarity is always computed between a curator's annotation and the GS annotation for the same piece of text. These similarity scores quantify how closely aligned or similar each annotator's interpretation is to the GS annotation.

3.2.1 Random Forest Classifier to estimate robustness

The next step after computing the semantic similarity metrics is to evaluate the robustness

of the metrics computed from ontology embeddings and compare it to existing metrics.

We used a Random Forest (RF) classifier to assess the robustness of the different semantic similarity measures. The goal of this model is to predict the ground truth (GS annotation) given the three curator annotations and a semantic similarity. The model was trained on tuples that contained six items - $(C1, S_{C1,GS}, C2, S_{C2,GS}, C3, S_{C3,GS})$ where Ci is an annotation by curator i, $S_{C1,GS}$ is the similarity score between Ci and the GS annotation. The target to be predicted is the GS annotation.

Random Forest Classifiers are trained separately for each similarity metric on a total of 1266 observations. The classifier's performance is evaluated using 10-fold cross-validation strategy and the average of accuracy and f1 score from each fold is reported.

3.2.2 Super similarity metrics

After evaluating the robustness of different similarity metrics, we combined the top metrics based on the RF model accuracy. We aimed to see if combining two robust semantic similarity metrics results in a more robust "super" metric. We used the weighted sum approach, assigning a weight to each similarity metric and calculating their weighted sum to create a composite score. The super similarity score is computed as:

$$\text{super\_similarity} = \alpha * \text{similarity\_metric\_1} + (1 - \alpha) * \text{similarity\_metric\_2}$$

We chose the alpha that resulted in the highest super similarity score using a grid search.

4 Results

The curator and gold standard datasets obtained from Dahdul et al (Dahdul et al., 2018). contained 1266 UBERON annotations.

During the embedding creation step, the sentences were generated using the following hyper parameters:

i. p = 0.5
ii. q = 2.0
iii. walk_number = 100
iv. walk_length = 5
v. edge_weight = 1

The embeddings were learned using a deep learning model trained with the following hyperparameters:

i. batch_size = 50
ii. learning_rate = 1e-03
iii. output_activation = sigmoid
iv. epochs = 2
v. embedding_dimension = 128

After training, we obtained 128-dimensional normalized embedding vectors for each of the 15,539 concepts in the UBERON ontology.

1266 tuples of the form (Ci, GS) where Ci is an annotation by curator i and GS is the corresponding gold standard annotation. Jaccard similarity was computed for these 1266 annotation pairs. Subsequently, cosine, dot product, minkowski distance, euclidean distance, and manhattan distance were computed between the embeddings of the concepts in (Ci, GS).

A RF model was trained on predict the gold standard annotation based on the three curators' annotations and their similarity to the GS. Table 1 shows that.

These results clearly show that our similarity measures based on ontology embeddings substantially outperform the commonly used Jaccard similarity. Among the metrics that use ontology embeddings we found dot product to be the most robust.

Next, we combined our top performing metric (Dot product) with all the other metrics to create "super metrics" to increase robustness. Table 2 reports the combinations where the score of the"super metric" exceeded the highest scores from individual metrics (Row 1, Table 1). We see that the combination of individual embedding based metrics does provide a boost in accuracy. The highest accuracy is obtained by combining Dot product with Minkowski distance.

5 Conclusions

Our goal was to develop robust semantic similarity metrics based on ontology embeddings to evaluate the performance of NLP approaches for automated annotation of scientific literature. We tested five embedding-based metrics against a widely used traditional metric and showed that the embedding-based metrics outperformed the traditional metric. We also found that combining the embedding-based metrics further improved accuracy.

Table 1: Accuracy of a Random Forest model on predicting the gold standard annotation given individual curator annotations and semantic similarity scores

Metric	Mean accuracy	Mean F1 score
Dot product	**0.87**	**0.86**
Cosine similarity	0.83	0.82
Euclidean distance	0.83	0.82
Manhattan distance	0.83	0.82
Minkowski distance	0.82	0.81
Jaccard similarity	0.76	0.75

Table 2: Composite Similarity score using two different semantic similarity metrics

Metric 1	Metric 2	Super Similarity Score
Dot product	Minkowski	0.89
Dot product	Manhattan distance	0.88
Dot product	Euclidean distance	0.87
Dot product	Cosine similarity	0.87

These results suggest that traditional semantic similarity metrics, which are based on comparisons of ontology subsumers, can be improved by using ontology embeddings. The new, more robust, and sensitive similarity metrics will enable an accurate assessment of NLP approaches for ontology annotation.

Acknowledgements

This work is funded by a CAREER award (#1942727) from the Division of Biological Infrastructure at the National Science Foundation, USA.

References

Mayla R Boguslav, Negacy D Hailu, Michael Bada, William A Baumgartner, and Lawrence E Hunter. 2021. Concept recognition as a machine translation problem. *BMC bioinformatics*, 22(1):1–39.

Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE transactions on knowledge and data engineering*, 30(9):1616–1637.

Mercedes Arguello Casteleiro, George Demetriou, Warren Read, Maria Jesus Fernandez Prieto, Nava Maroto, Diego Maseda Fernandez, Goran Nenadic, Julie Klein, John Keane, and Robert Stevens. 2018. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. *Journal of biomedical semantics*, 9(1):13.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Gene Ontology Consortium. 2006. The gene ontology (go) project in 2006. *Nucleic acids research*, 34(suppl_1):D322–D326.

Gene Ontology Consortium. 2012. Gene ontology annotations and resources. *Nucleic acids research*, 41(D1):D530–D535.

Hong Cui, Wasila Dahdul, Alexander T Dececchi, Nizar Ibrahim, Paula Mabee, James P Balhoff, and Hariharan Gopalakrishnan. 2015. Charaparser+ eq: performance evaluation without gold standard. *Proceedings of the Association for Information Science and Technology*, 52(1):1–10.

Wasila Dahdul, T Alexander Dececchi, Nizar Ibrahim, Hilmar Lapp, and Paula Mabee. 2015. Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy. *Database*, 2015.

Wasila Dahdul, Prashanti Manda, Hong Cui, James P Balhoff, T Alexander Dececchi, Nizar Ibrahim, Hilmar Lapp, Todd Vision, and Paula M Mabee. 2018. Annotation of phenotypes using ontologies: a gold standard for the training and evaluation of natural language processing systems. *Database*, 2018:bay110.

Pratik Devkota, Somya Mohanty, and Prashanti Manda. 2022a. Knowledge of the ancestors: Intelligent ontology-aware annotation of biological literature using semantic similarity. *Proceedings of the International Conference on Biomedical Ontology*.

Pratik Devkota, Somya D Mohanty, and Prashanti Manda. 2022b. A gated recurrent unit based architecture for recognizing ontology concepts from biological literature. *BioData Mining*, 15(1):1–23.

Pratik Devkota, Somya D Mohanty, and Prashanti Manda. 2023. Ontology-powered boosting for improved recognition of ontology concepts from biological literature.

Stephan Grimm. 2009. Knowledge representation and ontologies. In *Scientific data mining and knowledge discovery: principles and foundations*, pages 111–137. Springer.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Sara Hachem, Thiago Teixeira, and Valérie Issarny. 2011. Ontologies for the internet of things. In *Proceedings of the 8th middleware doctoral symposium*, pages 1–6.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Soizick Lesteven, S Derriere, Pascal Dubois, F Genova, A Preite Martinez, N Hernandez, Josiane Mothe, A Napoli, and Y Toussaint. 2007. Ontologies for astronomy. In *Library and Information Services in Astronomy V*, volume 377, page 193.

Prashanti Manda, James P Balhoff, Hilmar Lapp, Paula Mabee, and Todd J Vision. 2015. Using the phenoscape knowledgebase to relate genetic perturbations to phenotypic evolution. *genesis*, 53(8):561–571.

Prashanti Manda, Lucas Beasley, and Somya Mohanty. 2018. Taking a dive: Experiments in deep learning for automatic ontology-based annotation of scientific literature. *Proceedings of the International Conference on Biomedical Ontology*.

Prashanti Manda, Saed SayedAhmed, and Somya D. Mohanty. 2020. Automated ontology-based annotation of scientific literature using deep learning. In *Proceedings of The International Workshop on Semantic Big Data*, SBD '20, New York, NY, USA. Association for Computing Machinery.

Prashanti Manda and Todd J Vision. 2018. On the statistical sensitivity of semantic similarity metrics. In *ICBO*.

ROBIN McENTIRE. 2002. Ontologies in the life sciences. *The knowledge engineering review*, 17(1):77–80.

Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13(1):1–20.

Catia Pesquita, Daniel Faria, Andre O Falcao, Phillip Lord, and Francisco M Couto. 2009. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7).

Satya S Sahoo, Christopher Thomas, Amit Sheth, William S York, and Samir Tartir. 2006. Knowledge modeling and its application in life sciences: a tale of two ontologies. In *Proceedings of the 15th international conference on World Wide Web*, pages 317–326.

Barry Smith, Jennifer Williams, and Schulze-Kremer Steffen. 2003. The ontology of the gene ontology. In *AMIA Annual Symposium Proceedings*, volume 2003, page 609. American Medical Informatics Association.

Martin E Spencer. 1982. The ontologies of social science. *Philosophy of the social sciences*, 12(2):121–141.

Robert Stevens, Carole A Goble, and Sean Bechhofer. 2000. Ontology-based knowledge representation for bioinformatics. *Briefings in bioinformatics*, 1(4):398–414.

Philip Strömert, Johannes Hunold, André Castro, Steffen Neumann, and Oliver Koepler. 2022. Ontologies4chem: the landscape of ontologies in chemistry. *Pure and Applied Chemistry*, 94(6):605–622.

Alfred Ultsch and Jörn Lötsch. 2014. Functional abstraction as a method to discover knowledge in gene ontologies. *PLoS One*, 9(2):e90191.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1591–1601.

Ronghui You, Zihan Zhang, Yi Xiong, Fengzhu Sun, Hiroshi Mamitsuka, and Shanfeng Zhu. 2018. Golabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, 34(14):2465–2473.

Mengge Zhao, James M Havrilla, Li Fang, Ying Chen, Jacqueline Peng, Cong Liu, Chao Wu, Mahdi Sarmady, Pablo Botas, Julián Isla, et al. 2020. Phen2gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR genomics and Bioinformatics*, 2(2):lqaa032.