Enhancing Biomedical Knowledge Representation Through Knowledge Graphs

Sebastian Chalarca[1], Asim Abbas[1], Mutahira Khalid[2], Fazel Keshtkar[1], Syed Ahmad Chan Bukhari[1]

[1]Division of Computer Science, Mathematics and Science, Collins College of Professional Studies, St. John's University, Utopia,Parkway, Queens, 8000, NY, USA
[2]Technische Informationsbibliothek (TIB) Hannover, Lower Saxony, Germany
sebastian.chalarca19@my.stjohns.edu, abbasa@stjohns.edu, Mutahira.Khalid@tib.eu, keshtkaf@stjohns.edu, bukharis@stjohns.edu

Abstract

There is a plethora of information related to the biomedical domain on the internet. Unfortunately, retrieving this information online is challenging because of insufficient semantic metadata embedded within the web documents that help search engines interpret the biomedical information. Semantic annotators have partially bridged this gap, yet these tools frequently need to catch up in accuracy, speed, and the ability to dynamically represent knowledge. We initially developed "Semantically," a biomedical semantic content authoring platform to streamline and enhance biomedical annotations through a social-technical approach. Even so, the current system stores data in a relational schema, which lacks machine-readable content that allows search engines to parse the meanings to annotation recommendations. There is still the need for the amalgamation and contextually rich representation of annotation recommendation information to enhance navigation and exploration of data. Therefore, we propose a knowledge graph-based recommendation system with an nlp-enhanced search query to provide an environment for easy and quick access to optimal recommendations in a machine-readable knowledge graph format. We obtain results for the knowledge graph through an evaluation survey that substantiates the efficacy of our knowledge graph-based recommendation system, highlighting its role in advancing dynamic knowledge representation and semantic annotation in the biomedical domain. A demo is available at SebC750/Semantically at Knowledge_Graph_Branch (github.com)

Keywords: Knowledge Graph, Biomedical, Annotations, Knowledge Representation, Ontologies, Social-Technical

Introduction and Literature Review

In 2017, it was reported that every day, 2.5 million users visit PubMed while performing 3 million searches daily for biomedical content (Fiorini et al. 2017). These findings demonstrate a broader trend in these numbers: the biomedical domain has seen an incredible surge in texts and documents related to medical data. This deluge of biomedical information has led to a significant increase in the demand for quicker and more accurate data. However, many texts need contextual meta-data that give exact meanings to critical terms, making them more challenging to search when lacking such semantic meta-data. Search engines often need help finding biomedical texts related to a medical concept when many relevant texts need to include semantic meta-data that otherwise allows for more precise data retrieval. Thus, a semantic biomedical content authoring system is necessary for providing domain experts with the ability to embed machine-interpretable meta-data into biomedical documents before publishing to enhance search results. Such a system would greatly benefit the biomedical community by providing context-rich information that results in more precise data retrieval from knowledge-based search engines.

The preponderance of unstructured biomedical texts in the biomedical domain has been a subject of extensive research due to the importance of semantic biomedical annotation for achieving precise information retrieval and ease of accessing relevant biomedical search results. Other researchers have written extensively on developing semantic annotation programs. There are different annotation systems, each with a different purpose and focus. We can distinguish these systems into two main types: 1) Biomedical Semantic annotators and 2) Non-biomedical semantic annotators (Abbas et al. 2023a). The biomedical semantic annotator category is made up of two sub-categories: general-purpose annotators, which cover all biomedical domains using technologies like Natural Language Processing, ontologies, and sometimes machine learning to facilitate term-to-concept matching (Jovanovic et al. 2014). Conversely, subdomain-specific annotators focus on particular areas, such as neuroscience or pharmacology, offering targeted annotation services.

However, a common failure among these types of biomedical semantic annotators is that they need to maintain an acceptable level of, most importantly, both speed and accuracy. For semantic annotators who utilize machine learning to train their annotation algorithms, such as NOBLE Coder (Tsyetlin et al. 2009), Concept Mapper (Jovanovic et al. 2017), and Neji (Campos et al. 2013),

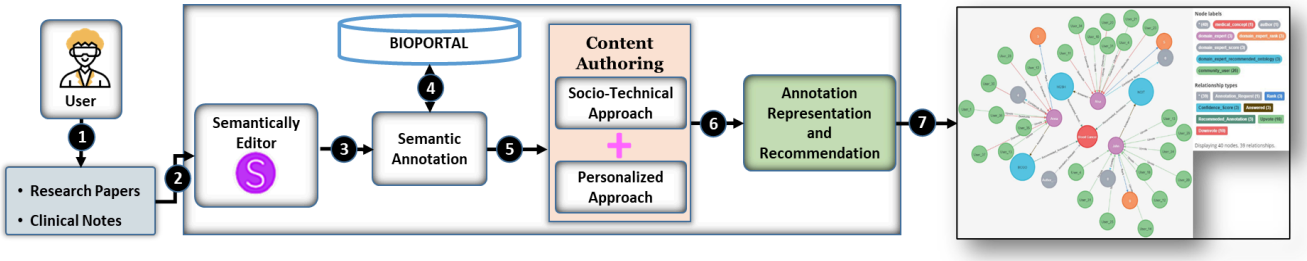

Figure 1. Overview of the Semantically biomedical content authoring workflow with a Knowledge Graph based Recommendation System Enhancement

they are still incapable of achieving strong lexical disambiguation when given terms that vary in meaning and ontologies. RysannMD attempts to address the lopsided imbalance of speed and accuracy of biomedical semantic annotators but is severely hampered due to the UMLS knowledge base, which only supports some biomedical domains (Cuzzola et al. 2017).

Initially, we tackled these shortcomings by introducing "Semantically," a system aimed at improving biomedical semantic annotation through a social-technical and personalized approach. Despite enhancements by the social-technical and personalized approach, the system's reliance on MySQL's tabular data structure limited its ability to contextualize relationships between entities like forum posts, replies, and ontologies. The demand for dynamic knowledge representation is evident—a system that presents information clearly and precisely to users, reflecting the most up-to-date information. Integrating a knowledge graph within Semantically is critical to address this need. Therefore, we are implementing a knowledge graph-based system, leveraging neo4j to dynamically represent data, making it easier for users to discover and understand the most suitable annotation recommendations. This shift not only updates our knowledge base but also allows for a more intuitive information exploration for domain experts who need up-to-date knowledge of optimal recommendations.

Proposed Methodology

Initially, we proposed "Semantically," a free tool designed to automate NLP tasks for biomedical annotation, which enhances the biomedical semantic annotation process through a social-technical and personalized recommendation approach. However, its reliance on relational schema storage, which fails to make data relationships machine-readable as well as isolation of information within separate forum posts, limits access to practical annotation recommendations for domain experts. To overcome this, we have integrated a knowledge graph-based recommendation system that offers a dynamic visual representation of annotations, enabling immediate access to top recommendations from various forum posts and improving the precision of data interpretation. We discuss the system's workflow in greater detail in the subsequent sections, as shown in Figure 1.

Semantic Annotation

In the initial semantic annotation process, users employ biomedical annotation systems to identify appropriate ontologies for key medical terms from knowledge bases like BioPortal, a task that is typically manual and tedious. This manual metadata embedding requires users to have in-depth knowledge of the relevant biomedical subdomain. Semantically streamlines this process by integrating the NCBO web service resource REST API (Jonquet et al. 2009) to analyze terms and retrieve their ontologies automatically. Users start by writing or importing documents into the Semantically editor. Upon clicking the annotation button, the system tokenizes the document's terms and fetches their ontologies from BioPortal. After this initial automatic annotation, domain experts can review the results, adjusting or deleting annotations as necessary.

Socio-Technical Approach

Following the automated annotation of the corpora, the user can manually assess the accuracy of ontologies given for highlighted terms. However, there is still a problem with lexical disambiguation because terms often have varying meanings and, thus, have definitions from multiple ontologies. To tackle this issue, we propose a social-technical approach where the community is involved in optimizing the semantic annotation process for other users by providing trustworthy and precise feedback. We have integrated a “Semantically Knowledge Cafe” page, where users can seek trustworthy recommendations for terminologies from public forum posts (Abbas et al. 2023a). Let us imagine a scenario where a user may write a document concerning different types of cancer. Suppose the user is unsure which ontology is most suitable for the term “Blood Cancer.” In that case, they can create a social-technical question, which consists of the specific question and a description explaining their problem. For

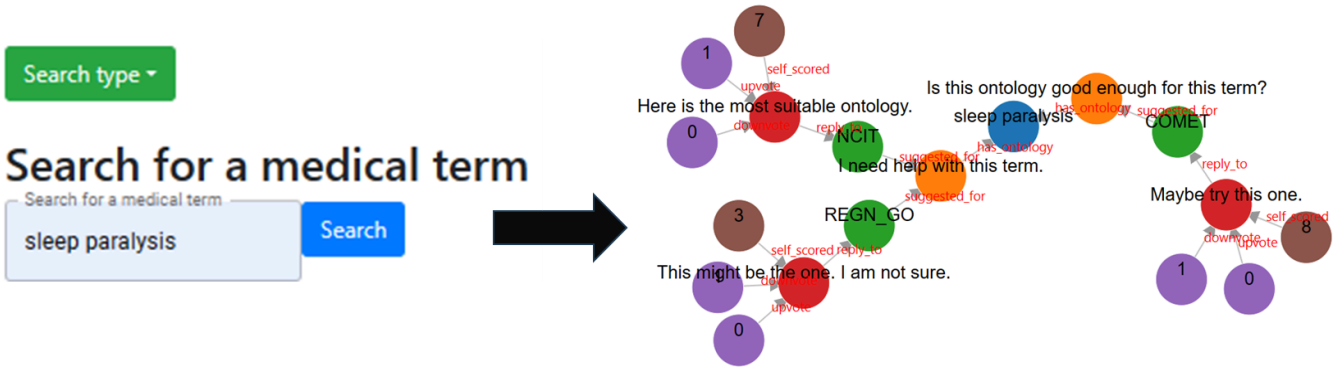

Figure 2. A visual representation of the process of using the Knowledge Graph Recommendation System interface. The diagram demonstrates a use-case scenario of the search query and the resulting knowledge graph for the given medical term input "sleep paralysis".

this scenario, the user may ask, "What is the suitable ontology for 'Blood Cancer'?". Once the question is submitted, the post will be publicly available to community members. Additionally, we suggest a personalized expert recommendation system for users seeking direct expert advice, utilizing cosine similarity to match users with the most relevant experts based on word embeddings, ensuring tailored guidance in their specific biomedical field (Abbas et al. 2023b). The integrations of the social-technical and personalized approaches scored 0.9 in precision, recall and F1 following benchmark testing (Abbas et al. 2024).

Annotation Recommendation and Representation

Using a social-technical and personalized approach, domain experts can guide users in selecting appropriate ontologies. Experts answer queries through replying by selecting an ontology and classification via the NCBO tree widget, assigning a confidence score, and explaining their recommendations. The user who posed the question receives these recommendations upon accessing their document, while other experts can vote on these responses to indicate agreement or disagreement. The user can then accept or reject the recommendations. Accepting the recommendation updates the annotation with the recommended ontology and definition, while rejecting preserves the original data.

Knowledge Graph Recommendation System and NLP Enhanced Search Query

Social-technical and personalized recommendation approaches have significantly enhanced the biomedical content authoring process but there is still room for improvement in this kind of system. Specifically, social-technical and personalized recommendations in "Semantically" are isolated to community posts and personalized one-to-one cooperation for improving annotations, potentially leaving users unaware of the most up-to-date and relevant ontology recommendations. Also, the system's tabular MySQL data structure fails to clearly represent the connections between users, terminologies, ontologies, and feedback (votes, confidence, and rating scores), hindering understanding of entity relationships and affecting machine interpretation. Knowledge graphs, adopted by sites like Google and Wikidata (Ringler and Paulheim 2017), utilize a triple model (subject, predicate, object). This structure is what makes Knowledge graphs particularly so useful. Having data represented in triples can contextualize the meanings of relationships defined in the graph, which is also machine-readable since this logical structure can be computable. Because of this, we can use Knowledge graphs to enhance data readability, relevance, and accuracy through knowledge representation. Knowledge graphs are helpful for many applications, from scene generation (D'Souza et al. 2023) to semantic networks for museum collections (Gawriljuk et al. 2016). Most of all, knowledge graphs significantly boost recommendation systems across various industries (Nigam et al. 2020). Leveraging this, we integrate a knowledge graph based recommendation system in "Semantically" for representing easily explorable information for annotation recommendations in a knowledge graph form.

Knowledge Graph Development, Interface and Functionality

To streamline the knowledge graph creation process, we migrated our MySQL database to Neo4j using the Neo4j ETL tool for schema mapping. This initial step involved using the ETL tool to identify node attributes and relationships based on existing database metadata. Despite the ETL's automatic schema mapping, we made manual adjustments to refine the graph's structure, such as correcting the translation of vote tables into relationships rather than distinct nodes with specific attributes.

Given Neo4j's use of the Cypher query language, which differs from MySQL's, we rewrote the backend code to accommodate Cypher queries for efficient database

management. Following the database setup, we focused on the user interface, specifically a Knowledge Graph search tab. This tab features a search bar with Named Entity Recognition (NER) capabilities and a dropdown menu for filtering user searches, medical terms, and ontologies. The search results are displayed as a color-coded graph illustrating the relationships between nodes, with labels explaining each connection, along with a table highlighting top recommendations based on various metrics like vote ratio and confidence scores. The process behind this system is shown in figure 2.

All Top results

Rank	Suggested ontology	Terminology	Ontology definition	Upvotes	Downvotes	Wilson Score	Confidence score	Rating	Score
1	NCIT	sleep paralysis	Temporary inability to speak or move while waking up or falling asleep.	1	0	0.207	7	5	0.636
2	COMET	sleep paralysis	Represents an address	0	1	0.000	8	0	0.267
3	REGN_GO	sleep paralysis	Cellular division	0	1	0.000	3	0	0.100

Figure 3. Rankings of recommendations obtained from the Knowledge Graph after user searches for the medical term “sleep paralysis”.

Suppose a user needs an ontology definition for the term "sleep paralysis". They receive a knowledge graph displaying the top 5 annotations with their scores. This graph shows recommended ontologies, the forum post suggesting each, the specific reply, and its reception (confidence score, upvotes, and downvotes). Recommendations are ranked by a total score derived from the Wilson score of upvotes and votes, normalized confidence and rating scores, and then averaged (Abbas et al. 2023a). Figure 3 illustrates the top-ranking table of recommendations for the term “sleep paralysis” sorted by the total score obtained from the aforementioned formula.

Results And Discussion

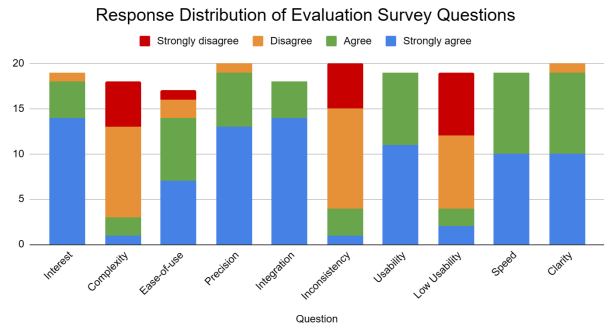

Figure 4. Results of performance metrics obtained from the evaluation survey for the Knowledge Graph based recommendation system.

To evaluate our knowledge graph-based recommendation system, we developed an evaluation protocol examining its performance in a social-technical/personalized context across several criteria: interest, ease of use, precision, integration, usability, speed, and clarity. Concerns such as inconsistency, low usability, and complexity were also addressed to ensure optimal ontology recommendations. We detailed our methodology, including score calculation formulas and the system's workflow, on a dedicated webpage accompanied by demo videos and images.

An evaluation survey employing a five-point scale (from strongly agree to strongly disagree) was distributed to 20 participants, comprising medical experts, programmers, and non-technical users, to collect feedback on the defined metrics. Responses were tallied to analyze the system's performance, with Figure 4. illustrating the distribution of responses across all performance metrics. The data suggests that the knowledge graph system excels in all positive aspects while mitigating negative ones like complexity and inconsistency. Most notably, the knowledge graph structured information was thought to be very clear, easy to interpret and quickly retrieved, demonstrating that the system successfully achieves ease of information navigation and exploration. The vast majority of users also expressed interest, highlighting a need for this enhancement. The system's high usability, integration, low complexity, and the effective consolidation of information into a single, easily accessible environment suggest a streamlined annotation recommendation process. Precision was generally received well but a few criticisms arose regarding the scoring system's perceived bias due to the equal weighting of different score types. Despite this, the results affirm a significant enhancement over existing biomedical annotation systems, indicating overall satisfactory efficiency while achieving dynamic knowledge representation.

Conclusion

This paper explores enhancing the semantic biomedical annotation system with Knowledge Graphs by incorporating a Knowledge Graph-based recommendation system into the semantic annotation process. This integration aims to enrich the representation of biomedical knowledge, significantly boosting the precision of document annotation. We highlight the advantages of layering knowledge representation with a Knowledge Graph amid the ongoing release of unstructured biomedical information. As unstructured biomedical information continues to be released on research publishing sites, there will continue to be a need to improve the semantic annotation process. This research underlines our commitment to further optimizing our methodology. The demo is accessible with instructions for local deployment at: SebC750/Semantically_at_Knowledge_Graph_Branch (github.com).

References

Abbas, Asim, et al. “Enhancing biomedical semantic annotations through a knowledge graph-based approach.” The International FLAIRS Conference Proceedings, vol. 36, 2023a.

Abbas, Asim, Steve Mbouadeu, et al. “Personalized semantic annotation recommendations on biomedical content through an expanded socio-technical approach.” *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies*, 2023b.

Abbas, Asim, Tahir Hameed, et al. “A socio-technical approach to trustworthy semantic biomedical content generation and sharing.” *Information Sciences*, vol. 666, May 2024.

Campos, David, et al. “A modular framework for biomedical concept recognition.” *BMC Bioinformatics*, vol. 14, no. 1, 24 Sept. 2013.

Cuzzola, John, et al. “RYSANNMD: A biomedical semantic annotator balancing speed and accuracy.” Journal of Biomedical Informatics, vol. 71, July 2017.

D’Souza, Jessica, et al. “Knowledge-based scene graph generation in the medical field.” 2023 IEEE International Conference.

Fiorini, Nicolas, et al. “Towards pubmed 2.0.” *eLife*, vol. 6, 30 Oct. 2017.

Gawriljuk, Gleb, et al. “A scalable approach to incrementally building knowledge graphs.” Research and Advanced Technology for Digital Libraries, 2016.

Jonquet, C. et al. “NCBO annotator: semantic annotation of biomedical data,” in International Semantic Web Conference, Poster and Demo session, 2009, vol. 110, [Online].

Jovanović, Jelena, et al. “Automated semantic tagging of textual content.” *IT Professional*, vol. 16, no. 6, Nov. 2014, pp. 38–46.

Jovanović, Jelena et al. “Semantic annotation in biomedicine: The current landscape.” Journal of Biomedical Semantics, vol. 8, no. 1, 2017.

Nigam, Vanshika Vikas, et al. “A review paper on the application of knowledge graph on various service providing platforms.” 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Jan. 2020.

Ringler, Daniel, and Heiko Paulheim. “One knowledge graph to rule them all? analyzing the differences between DBpedia, Yago, Wikidata & Co..” *KI 2017: Advances in Artificial Intelligence*, 2017, pp. 366–372

Tseytlin, Eugene et al. “Noble – flexible concept recognition for large-scale biomedical natural language processing.” BMC Bioinformatics, vol. 17, no. 1, 2016.