Optimizing Semantic Enrichment of Biomedical Content through Knowledge Sharing

Asim Abbas[1], Steve Mbouadeu[1], Avinash Bisram[1], Nadeem Iqbal[1], Fazel Keshtkar[1] and Syed Ahmad Chan Bukhari[1,*]

[1]Division Of Computer Science, Math & Science, Collins College of Professional Studies, St. John's University, Queens, NYC, USA

Abstract

Each day a vast amount of unstructured content is generated in the biomedical domain from various sources such as clinical notes, research articles and medical reports. Such content contain a sufficient amount of efficient and meaningful information that needs to be converted into actionable knowledge for secondary use. However, accessing precise biomedical content is quite challenging because of content heterogeneity, missing and imprecise metadata and unavailability of associated semantic tags required for search engine optimization. We have introduced a socio-technical semantic annotation optimization approach that enhance the semantic search of biomedical contents. The proposed approach consist of layered architecture. At First layer (Preliminary Semantic Enrichment), it annotates the biomedical contents with the ontological concepts from NCBO BioPortal. With the growing biomedical information, the suggested semantic annotations from NCBO Bioportal are not always correct. Therefore, in the second layer (Optimizing the Enriched Semantic Information), we introduce a knowledge sharing scheme through which authors/users could request for recommendations from other users to optimize the semantic enrichment process. To guage the credibility of the the human recommended, our systems records the recommender confidence score, collects community voting against previous recommendations, stores percentage of correctly suggested annotation and translates that into an index to later connect right users to get suggestions to optimize the semantic enrichment of biomedical contents. At the preliminary layer of annotation from NCBO, we analyzed the n-gram strategy for biomedical word boundary identification. We have found that NCBO recognizes biomedical terms for n-gram-1 more than for n-gram-2 to n-gram-5. Similarly, a statistical measure conducted on significant features using the Wilson score and data normalization. In contrast, the proposed methodology achieves an suitable accuracy of ≈90% for the semantic optimization approach.

Keywords

Structured data, Biomedical semantic enrichment, Annotation optimization, Recommendation,

1. Introduction

Over the last few decades, a huge volume of the digital unstructured textual content has been generated in biomedical research and practice, including various content types such as scientific papers, medical reports, and physician notes. This explosive growth in the biomedical domain has introduced several access-level challenges for researchers and practitioners. These valuable information are available in the web contents but still opaque to information retrieval and knowledge extraction search engines because of the missing machine-interpretable metadata (semantic annotations) [1]. Search engines require the metadata to properly index contents in a context-aware fashion for the precise search of biomedical literature and to foster secondary activities such as automatic integration for meta-analysis [2]. Incorporating machine-interpretable semantic annotations at the pre-publication stage (while first-time drafting) of biomedical contents and preserving them during online publishing is desirable and will be a great value addition to the broader semantic web vision [3]. However, both these processes are complex and require deep technical and/or domain knowledge. Therefore, a state-of-the-art, freely accessible biomedical semantic content authoring framework would be a game-changer.

The main components of the semantic annotation process are ontologies which are sets of machine-readable controlled vocabularies that provide the "explicit specification of a conceptualization" of a domain. Similarly semantic annotators are designed to facilitate tagging/annotating the related ontology concepts with pre-defined terminologies in a manual, automatic, or hybrid way [4]. As a result, users produce semantically richer content when compared with traditional composing processes e.g., using a word processor [5]. Owing to the significance of the semantic annotation process in biomedical informat-

4th Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2022, September 18–23 2023, Seattle, WA, USA.

*Corresponding author.

†These authors contributed equally.

abbasa@stjohns.edu (A. Abbas); steve.mbouadeu19@stjohns.edu (S. Mbouadeu); avinash.bisram19@stjohns.edu (A. Bisram); iqbaln@stjohns.edu (N. Iqbal); keshtkaf@stjohns.edu (F. Keshtkar); bukharis@stjohns.edu (S. A. C. Bukhari)

https://www.linkedin.com/in/asim-abbas-b2891ab8/ (A. Abbas); https://bukharilab.org (S. Mbouadeu); https://bukharilab.org (A. Bisram); https://bukharilab.org (N. Iqbal); https://bukharilab.org (F. Keshtkar); https://bukharilab.org (S. A. C. Bukhari)

0000-0001-6374-0397 (A. Abbas); 0000-0002-6517-5261 (S. A. C. Bukhari)

ics research and retrieval, the scientific community has invested considerable resources in the development of semantic annotators. Whereas the biomedical annotators predominantly use term-to-concept matching with or without machine learning-based methods [5]. Likewise, biomedical annotators such as NOBLE Coder [6], ConceptMapper [7], Neji [8], and Open Biomedical Annotator [9] use machine learning and annotate text with an acceptable processing speed. However, they lack a strong disambiguation capacity, i.e., the ability to identify the correct biomedical concept for a given piece of text among several candidate concepts. Whereas NCBO Annotator [10] and MGrep services are quite slow, RysannMd annotator claims to balance speed and accuracy in the annotation process. However, on the flip side its knowledge base is limited to certain ontologies available in UMLS (Unified Medical Language System) and does not provide full coverage of all biomedical sub-domains [11]. Other than the technical challenges as stated above, one of the main reasons why semantic authoring is still in infancy and researchers have not been able to achieve the desired objectives is because researchers did not realize the importance of original content creator (author) involvement and heavily focused on technology sophistication where systems interactions were limited to the technical persons. Typically, only the author knows why they used a particular term to explain a concept. Third-party developers are naturally not privy to such tacit knowledge. Researchers and practitioners face access-level issues due to the dissonance between those who authored the original work and those who added semantic annotations and published it. The majority of the authors lack of technical and/or domain knowledge, and there is a steep learning curve that necessitates substantial time to develop critical skills that are not the primary job of the majority of the authors. To overcome the aforementioned challenges, we propose a semantic annotation optimization approach that adopts a knowledge-sharing strategy and presents a framework through which users can seek and provide suggestions to optimize the annotation quality. Our systems keep track of the recommender confidence score, gather community feedback regarding prior recommendations, store the percentage of correctly suggested annotations, and translate that into an index to later connect the appropriate users to receive suggestions to optimize the semantic enrichment of biomedical contents. The rest of the paper is organized as follows. The proposed methodology section covers implementation details of catering preliminary semantic annotation, semantic annotation optimization and example scenarios in an annotation optimization environment. Subsequently, the result and discussion include the dataset utilized, methodology for evaluation, and results achieved at system-level. The conclusion section summarizes the systems' working and future plans.

2. Proposed Methodology

This section presents the biomedical semantic annotation recommendation and optimization processes. We developed a system through which users access a biomedical content authoring interface analogous to the MS Word editor to type or import biomedical contents for their semantic enrichment. The system generates the first layer of semantic annotation utilizing the NCBO Bioportal API [10] Figure 1(a). However, the correctness of the acquired semantic annotation varies as one annotation is available to multiple ontologies. Furthermore, the linguistic mapping mechanism of the Bioportal recommender often ignores the sentence and paragraph-level context. Therefore, the suggested annotations might be correct at the content level. However, they may be entirely incorrectly contextually in a particular setting. Only the original author knows in which context they used a specific concept. Therefore a state-of-the-art knowledge sharing approach is designated as it provides a system that allows the author to query peers for more specific semantic annotation against the biomedical term to optimize the annotation quality. In the following sections, we explain 1) Preliminary Semantic Enrichment, 2) Optimizing Semantic Enrichment, and an Example Scenario in an annotation optimization environment Figure 1. Additionally, in an Example Scenario below, we categorize the role as author who posted a query, $E_i = e_1, e_2, e_3 .... e_n$ represents responder or expert, and $U_i = u_1, u_2, u_3 .... u_n$ is community users.

2.1. Preliminary Semantic Enrichment

A biomedical annotator is an essential component of semantic annotation or enrichment [12]. Available biomedical annotators use publicly available biomedical ontologies, such as Bioportal [10] and UMLS [4], to help the biomedical community researcher to structure and annotate their data with ontology concepts for better information retrieval and indexing. However, the semantic annotation and enhancement process is tedious and requires expert curators. With our developed systems, we automate the semantic annotations assignment process. For that, we utilized the NCBO Bioportal web-service resources [10] that analyze the raw textual content and tag them with relevant biomedical ontology concepts. By pressing the "Annotate" button, users can generate a preliminary level of annotations without the need for any technical knowledge. Initially, authors can either import pre-existing content from research papers, clinical notes, and biomedical reports or start typing directly in the semantic text editor see Figure 1(a). Our systems accept the user's free text and feeds it forward as input to a concept recognition engine. The engine identifies relevant ontologies, acronyms, definitions, and ontology

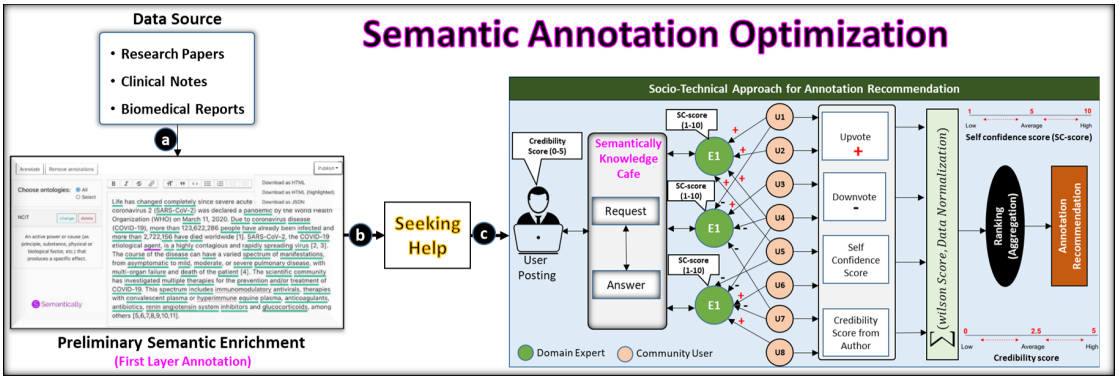

Figure 1: Proposed M'ethodology of Biomedical Content Semantic Optimization

links for individual terminologies that are best matched based on the context by following the string matching approach. This semantic information is displayed in our system's annotation panel for human interpretation and understanding Figure 1(a). Authors may alter the generated semantic information based on their knowledge and experience, such as choosing appropriate ontology from the list, selecting suitable acronyms, removing semantic information or annotating for explicit terminology, etc. Users without a technical background may easily navigate a simplified interface, while more sophisticated users may utilize advanced options to control the semantic annotation and authoring process further.

2.2. Seeking Annotation Recommendation

Subsequently at initial level semantic annotation, the author is enabled to approach and get recommendation from peer review for a correct and high quality annotation through seeking help module Figure 1(b). The authors are required to select the biomedical term from the preliminary annotation interface for correct annotation by peer review. Additionally the author is facilitated with an interface to smoothly query with available options such as a drop down menu of recommended queries for an author. Similarly author can explain their query and provide evidence and links to better convey their query to the expert E_i or peer review. Finally when the author submit their query, it is posted on "Semantically Knowledge Cafe" forum for peer review response and a notification is send to the community users as shown in Figure 1(c).

2.3. Optimizing the Enriched Semantic Information

To optimize the newly harvested annotations through the knowledge-sharing process, authors are required to select the existing annotation and then click to seek the help option from the panel. A pop-up appears with a drop-down of question sets that authors may ask. For example, if authors are interested to know whether a particular prelimanary annotation or ontology is correct. They can select the questions and fill in the required information. Similarly, authors can seek peers help posting a question. All the posted questions will go to the "Semantically Knowledge Cafe" forum style. The "Semantically Knowledge Cafe" is a virtual social place where people/users ask questions and seek help regarding their annotation improvement. As soon authors receive a response from the crowd, they are notified, and all suggestions start the display with the option to accept or reject. Here the authors decide to choose a particular suggestion based on social indexing. Our system calculates the social index and displays each suggestion based on its index score in descending order. To gauge the credibility of the human recommender, our systems record the recommender confidence score, collect community voting against previous recommendations, store the percentage of correctly suggested annotation, and translate that into an index to later connect the right users to get suggestions to optimize the semantic enrichment of biomedical contents. All the process information is stored in the backend knowledge base.

Consider an author is required to find correct ontology annotations from peer review for the biomedical term *"worsening shortness of breath"* as shown in Figure 2. The author posts the query on a "Semantically Knowledge Cafe" forum such as *"Which Ontology should I use for medical content 'worsening shortness of breath'"?* and receives replies from fellow users or experts E_i. We cate-

gorized users who replied as expert users as E_i with *"No of Reply-post"* and suggested the correct annotation for a required biomedical content as *"Expert Annotation"* see Figure 2. In the study three expert users participated and each expert suggested the annotation as *("RCD","UPHENO" and "NCIT")*. We also asked experts to provide their confidence score which they recorded as of (4,6 and 7) out of a scale from 1 to 10. The community users/crowd U_i at 'Semantically Knowledge Cafe' can observe the suggested recommendation and record their up and down voting about a particular suggestions. From users U_i, we recorded upvotes (9,10,11) and downvotes (9,8,7) to the expert recommended annotation. Whenever the author accepts recommended annotation from experts E_i, a credibility score is recorded. We used Wilson score confidence interval for a Bernoulli parameter to normalize and aggregate the recorded scores, see Equ. (1).

$$Wislon_{score} = \frac{\left(\hat{p} + \frac{z^2_{\frac{\alpha}{2}}}{2n} \pm z_{\frac{\alpha}{2}}\sqrt{\left[\hat{p}\,(1-\hat{p}) + \frac{z^2_{\frac{\alpha}{2}}}{4n}\right]}\right)}{\left(1 + \frac{z^2_{\frac{\alpha}{2}}}{n}\right)} \quad (1)$$

Where,

$$\hat{p} = \left(\sum_{n=1}^{N} +V\right) / (n) \quad (2)$$

$$n = \sum_{i=0}^{N}\sum_{j=0}^{M}\left(+V_i, -V_j\right) \quad (3)$$

and, $z_{\frac{\alpha}{2}}$ is the $\left(1 - \frac{\alpha}{2}\right)$ quantile of the standard normal distribution.

In Equ. (1), $\hat{p}$ is the sum of upvotes $(+V)$ of a community user's U_i to the expert E_i response for a post from an author for correct annotation divided by overall votes $(+V, -V)$ see Equ. (2). Likewise n is the sum of number of upvote and downvote $(+V, -V)$ see Equ. (3) and α is the confidence refers to the statistical confidence level: pick 0.95 to have a 95% chance that our lower bound is correct. However the z-score in this function is fixed. Likewise, a data normalization formula (see Equ. (4)) is employed on each expert E_i confidence score and author credibility score to downstream the value between 0 and 1.

$$z_i = (x_i - min(x))/(max(x) - min(x)) * Q \quad (4)$$

Where z_i is the i^{th} normalized value in the dataset. Where x_i is the i^{th} value in the dataset e.g the user confidence score. Similarly, $min(x)$ is the minimum value in the dataset, e.g the minimum value between 1 and 10 is 1, so the $min(x) = 1$ and $max(x)$ is the maximum value in the dataset, e.g the maximum value between 1 and 10 is 10, so the $max(x) = 10$. Consequently, a mean $\hat{x} = \frac{1}{N}\sum_{i=0}^{N} x_i$ applied on Wilson score, normalize the self confidence and author credibility score for each expert E_i that suggested annotation as *"aggregate score"* of $(0.458, 0.381, 0.518)$. Finally, $argmax(x_i)$ function is applied on the aggregate score to obtain the maximum score earned by each expert E_i annotation which is 0.518. Eventually, the high proficient and ranking annotation is recommended to the author as *"NCIT"* and *"Reply-post=3"* for the biomedical content *"worsening shortness of breath"*, see Figure 2. The same process is applied for another biomedical content, *"Acute Flaccid Myelitis"*, yet the scenario or query can be changed.

3. Results and Discussion

30 people participated in our proposed model. We recruited participants by via social media request and asking them to participate in the study. Further, We categorize participants as the most graduate-level students with computer and biological science backgrounds. Accordingly, We considered a set of 30 articles from pubmed.org [13] and randomly distributed it to the participants. Similarly, We provided a user manual of systems along with a pre-recorded video about system usage. Afterward, We asked each participant to generate queries on the "Semantically Knowledge Cafe" about the biomedical content annotation about which they like to seek social help. Collectively, our participants post 140 questions to the system. All the participants have also recorded their confidence scores between 1 and 10 from the suggestions they received as a satisfaction score. Consequently, Our system recorded 421 responses against 140 questions from expert users. Similarly, 2929 and 3149 up and down votes were also recorded against the suggestions annotations. Table 1 illustrates participants and their responses.

Table 1
Datasets utilized for experimental purpose

Title	Numbers
No of Participants	30
No of Documents	30
No of Posts	140
No of Response	421
No of Upvotes	2929
No of Downvotes	3149

3.1. Performance Measurement: Preliminary Semantic Enrichment

After catering initial level semantic information from NCBO Bioportal, we analyzed the content following the n-gram strategy. This strategy is crucial for the biomedical word or concept boundary detection process. A set of

Medical Content	No of Reply_post	Expert Annotation	Up Votes	Down Votes	Total Votes	Wilson scoring	Expert Confedence Level	Normalize confedence Score	Credibility Score	Normalize Credibility score	Agreegate (wilson , Confidence and	Final Score	Final Recommendation
Which Ontology should I used?													
worsening shortness of breath	1	RCD	9	9	18	0.291	4	0.333	4	0.75	0.458	0.518	Reply_Post =3
	2	UPHENO	10	8	18	0.337	6	0.556	2	0.25	0.381		
	3	NCIT	11	7	18	0.386	7	0.667	3	0.5	0.518		
Acute Flaccid Myelitis	1	MEDLINEPLUS	20	5	25	0.609	5	0.444	5	1	0.684	0.684	Reply_Post =1
	2	ICD10CM	10	2	12	0.552	4	0.333	4	0.75	0.545		
	3	MESH	16	3	19	0.625	2	0.111	1	0	0.245		
	4	HHEAR	4	0	4	0.511	3	0.222	1	0	0.244		

Figure 2: A statistical process of semantic annotation optimization approach.

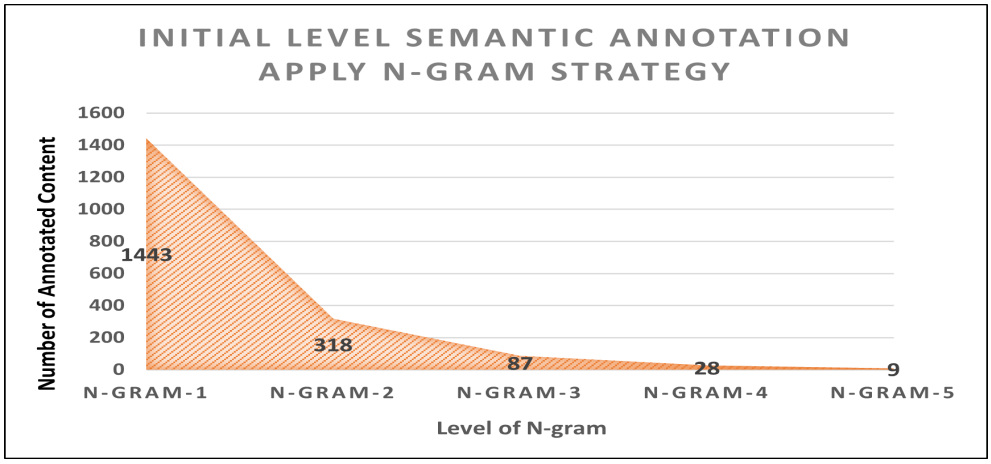

Figure 3: Initial level semantic annotation performance employ biomedical term of N-gram strategy.

30 pubmed.org [13] articles is processed at the initial level, consequently obtaining annotated biomedical terms in the range of n-gram-5. Subsequently scrutinized, we found the proposed annotation system identifies biomedical terms of n-gram-1 quantitatively higher than n-gram-2 to n-gram-5. A very few biomedical terms identify of n-gram-5 see Figure 3. However, the biomedical terms of $n - gram > 1$ deliver extra meaningful and coherent information to the user contextually. For example, "blood pressure is high", "he has coronary artery disease", and "liver function test is normal" are more meaningful terms as compared to a single term such as "pressure", "blood", "coronary", and others. As the $n - gram$ word size increases, the accuracy of composite terms decreases, as shown in Figure 3. Because the proposed system employs exact word matching to the terminology (Bioportal) approach, the primary characteristic of the exact word matching approach is that a single word matches more accurately than a combined or compound word.

3.2. Performance Measurement: Knowledge-sharing based Semantic Enrichment Optimization

A domain expert from the academia at the professor level is engaged to evaluate these results manually based on their knowledge and experience. After that, calculate the system level accuracy for semantic annotation before socio-techinical semantic annotation optimization and after socio-technical semantic annotation optimization approach Figure 4. A document's level accuracy is recorded with-out a socio-technical and with a socio-technical approach. The Figure 4 on X-axis represents the number of 30 documents processed. In contrast, the Y-axis at the left represents the level of accuracy with-out a socio-technical approach, and the Y-axis at the right represents the level of accuracy with a socio-technical approach. Consequently, scrutinizing the results of a system with a socio-technical approach performed better than with-out a socio-technical at document level. Until high accuracy of 90% has been gained by nine documents and lower accuracy of 87% has gained by three documents and maximum number of documents has gained accuracy between 87% and 90% with socio-technical approach see

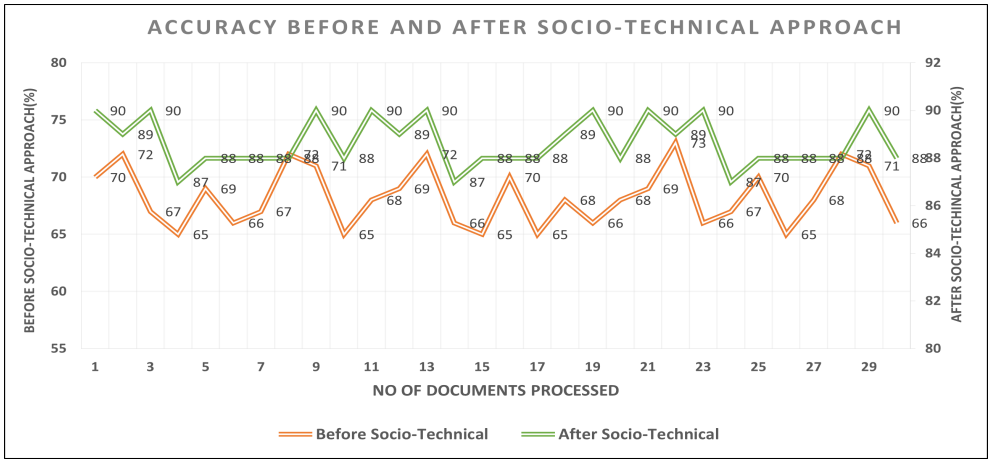

Figure 4: System level performance for socio-technical annotation recommendation.

Figure 4. Similarly high accuracy of 73% is yielded by one document and low accuracy of 65% is gained by five documents and maximum number of documents gained accuracy in the range of 65% to 73% with-out socio-technical approach. Overall the proposed annotation optimization socio-technical system remains the winner by obtaining high precision for each document related to the with-out socio-technical.

3.3. Semantically Workspace: Semantic Annotation Optimization Demonstration

Initially the author is enable to import or write the biomedical content in the editor and click on annotation button to get preliminary annotation see Figure 5. The underline word with green color presented annotated term, subsequently when author select any term the underline color change to pink and *"Need Help"* option is appeared on left side panel to the author see Figure 5(a). After click on *"Need Help"* an interface is open, where author can write there query from expert for recommended annotation to explicit terminology Figure 5(b). Additionally author is facilitated with primary options for quick query. When the author click on "*Submit*" button, the query is posted on the "Semantically Knowledge Cafe" forum and a new post notification received to the community users U_i as shown in Figure 5(c). Whenever users U_i click on "Semantically Knowledge Cafe" , the new posted query is appeared as shown in Figure 5(d). Now that if user know the answer of the posted query, he/she is enable to click on "*Answer*" button to reply author post/query as shown in Figure 5(d). Subsequently reply to the post by user with record self confidence score, now the role of this user is consider as a domain expert. Similarly, a smooth interface with possible option is available to the expert for reply post. Subsequently reply by the expert to the author post with precise annotation, same while other community users U_i are enable to give up-vote or down-vote to the expert reply post shown in Figure 5(e). Finally a high quality annotation recommendation notification is generated to the author by aggregating wilson score, and expert self confidence score as shown in Figure 5(f). Whenever the author click on *"New Recommendation"* link, a high quality expert recommended annotation is pop-up see Figure 5(g). Now here author is allow either accept the recommended annotation or reject, while accepting annotation a credibility score is recorded to the author profile between 1-5, vise versa no score is recorded to the author profile. Similarly by accepting recommended annotation, initial annotation for specific terminology is replaced by recommended one and thus annotation optimization process is completed Figure 5(g).

4. Conclusion

This research advances state-of-the-art biomedical semantic research and systems, enabling various biomedical users to author context-aware content with no prior technical skills needed. An out-of-the-box socio-technical semantic annotation optimization approach is presented to automate the semantic enrichment mechanism and discover the precise semantic annotation while keeping the original content creator in the loop. The end user is facilitated with an authoring interface similar to the MS Word editor type/write biomedical contents. To cater the preliminary semantic annotation or enrichment at the content level, we utilized Bioportal endpoint APIs and automated the configuration process for au-

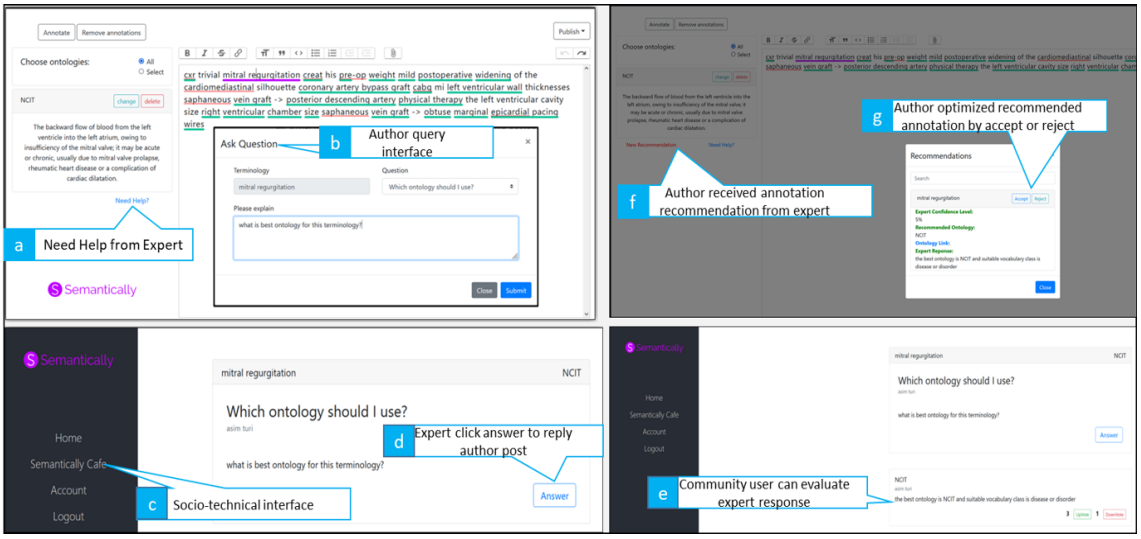

Figure 5: Semantic Annotation optimization and Enrichment Demonstration Interfaces.

thors. Similarly, the semantic annotation optimization approach is designed where the author can post their query for optimized annotation recommendation. In our future work, we plan to expand the backend knowledge graph and apply the neural graph networks. The semantic annotation optimization system is available at https://gosemantically.com.

Acknowledgments

This work is supported by the National Science Foundation grant ID: 2101350.

References

[1] A. Abbas, S. F. Mbouadeu, F. Keshtkar, J. DeBello, S. A. C. Bukhari, Biomedical scholarly article editing and sharing using holistic semantic uplifting approach, in: The International FLAIRS Conference Proceedings, volume 35, 2022.

[2] S. A. C. Bukhari, Semantic enrichment and similarity approximation for biomedical sequence images, Ph.D. thesis, University of New Brunswick (Canada), 2017.

[3] P. Warren, J. Davies, D. Brown, The semantic web-from vision to reality, ICT futures: Delivering pervasive, real-time and secure services (2008) 55–66.

[4] A. Abbas, M. Afzal, J. Hussain, S. Lee, Meaningful information extraction from unstructured clinical documents, Proc. Asia Pac. Adv. Netw 48 (2019) 42–47.

[5] K. Hasida, Semantic authoring and semantic computing, in: New Frontiers in Artificial Intelligence, Springer, 2003, pp. 137–149.

[6] E. Tseytlin, K. Mitchell, E. Legowski, J. Corrigan, G. Chavan, R. S. Jacobson, Noble–flexible concept recognition for large-scale biomedical natural language processing, BMC bioinformatics 17 (2016) 1–15.

[7] C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter, K. Verspoor, Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters, BMC bioinformatics 15 (2014) 1–29.

[8] D. Campos, S. Matos, J. L. Oliveira, A modular framework for biomedical concept recognition, BMC bioinformatics 14 (2013) 1–21.

[9] J. Jovanović, E. Bagheri, Semantic annotation in biomedicine: the current landscape, Journal of biomedical semantics 8 (2017) 1–18.

[10] C. Jonquet, N. Shah, C. Youn, C. Callendar, M.-A. Storey, M. Musen, Ncbo annotator: semantic annotation of biomedical data, in: International Semantic Web Conference, Poster and Demo session, volume 110, Washington DC, USA, 2009.

[11] J. Cuzzola, J. Jovanović, E. Bagheri, Rysannmd: a biomedical semantic annotator balancing speed and accuracy, Journal of Biomedical Informatics 71 (2017) 91–109.

[12] S. F. Mbouadeu, A. Abbas, F. Ahmed, F. Keshtkar, J. De Bello, S. A. C. Bukhari, Towards structured biomedical content authoring and publishing, in:

2022 IEEE 16th International Conference on Semantic Computing (ICSC), IEEE, 2022, pp. 175–176.
[13] PubMed, National Center for Biotechnology In- formation, 2022. https://pubmed.ncbi.nlm.nih.gov/.