Ontology-based knowledge extraction from scientific biomedical texts

Veerendra Nath Jasthi

Abstract:
With the increase in rates of scientific biomedical literature, manual extraction and compilation of knowledge is becoming impossible to accomplish. Ontology-based techniques present a systematic way of mechanically extracting and describing bio-medical information present in free-form texts. In this paper, an expansion is featured regarding the designing and application of an ontology based model of extracting knowledge in the biomedical literature. Based on a combination of natural language processing (NLP), named entity recognition (NER), semantic reasoning, and domain-specific ontology such as the Gene Ontology (GO) and Unified Medical Language System (UMLS), the proposed method produces structured information based on textual information. The framework was tested on the set of PubMed papers with a high accuracy rate in the mapping of terms and extraction of relations. The implications of the findings indicate that semantic technologies will play an important role in the interpretability and usability of biomedical data to downstream applications including decision support and biomedical discovery.

Keywords: Ontology; Knowledge Extraction; Biomedical Text Mining; Semantic Analysis; Natural Language Processing (NLP); Biomedical Ontologies; Named Entity Recognition; Gene Ontology; UMLS; Semantic Web.

I. INTRODUCTION

Biomedical research field is growing faster than ever before and on average thousands of articles are published every week in journals, repositories, and other online platforms [1]. This literature explosion constitutes a big challenge to the researcher and clinicians, who may be trying to keep up and make informed choices. Although conventional search engines and indexing services may be used to find the right documents, they are ineffective in the scope of capturing higher semantic relationships and that are structured in the unstructured texts. This issue created the need to develop knowledge extraction as a central research topic of the biomedical informatics field, specifically with semantic technologies, in the form of ontologies.

Ontologies describe expertise of any given domain in a formal manner and represent knowledge so that it can be processed and understood by a computer, including entities, attributes and relationship between entities and attributes [5]. Such ontologies as Unified Medical Language System (UMLS), Gene Ontology (GO), or SNOMED CT are used in the biomedical field to define complex concepts of biological processes, diseases, genes, and treatment in a standardized conceptual and vocabulary scheme. Ontologies have a large potential in improving the extraction and organization of the information presented in a scientific text when used alongside natural language processing (NLP) methods by normalizing terms and enabling reasoning.

The use of ontology-based methods of extracting the biomedical knowledge has a lot of benefits [6]. First, it can support semantic inter-operability, letting the data of different sources be combined and compareding well. Second, it improves the precision, ambiguating terms with respect to context, which is the necessary quality in biomedical field when it comes to polysemy and synonymy. Third, the ontology-based systems promote inferencing, through which it is possible to reveal implicit relations that have not been discussed in the text explicitly. As an example, detecting a drug-disease interaction by the presence of overlapping pathways or commonly occurring genetic markers described in various papers [4].

Nevertheless, even with the potential of ontology based systems, there are a number of challenges. Biomedical text is conceptually heterogeneous, full of professional jargon, acronyms, different styles of wording [10]. Guiding free-text through ontology and disambiguation to normal keywords needs algorithms that are advanced to do NER (Named Entity Recognition), awareness to context, and semantic matching. Also, the

current information extraction systems have a low scalability problem, due to an excessive reliance on rule-based patterns or are based on deep learning models that are not interpretable nor domain-specific.
Entities and relations have become easier to extract through recent developments in biomedical NLP including the introduction of transformer-based models, such as BioBERT and SciSpacy. However, even these approaches still call on an organized framework to bring on some unprocessed elements to build the representations of knowledge. This is where systems based on ontology excel: not only labeling, extraction of terms but organizing the terms in a consistent structure that is consistent with domain knowledge and possibly capable of semantic queries and reasoning.
The aim of this study is to design a knowledge extraction framework, based on ontology, that is modified to scientific biomedical documents. The framework integrates NLP techniques on preprocessing and entity extraction and semantic technologies that would support mapping, disambiguation, and reasoning. It also uses several ontologies of biomedicine to make it widely covered and aligned with the terms and concepts. The resulting knowledge is then placed in a knowledge graph, where it can be query-based advanced as well as visualized and inferred. This system will help researchers, health care professionals and data scientists gain more insight into the information held in large biomedical corpora faster and more accurately hence helping in improving diagnosis, treatment and further biomedical investigation [11].

Novelty and Contribution

The proposed research is a new, hybrid framework of ontology-based extraction of knowledge in biomedical scientific literature, which meets the main limitations of both the old rule-based and recent machine learning methods. In the suggested system, entity recognition based on deep-learning, multiple ontology semantic matching, and logical reasoning are three fundamental entities integrated into a single and scalable system. That is what makes the proposed system new and innovative [7].
Contrary to most previous systems that tend to extract merely entity tags or undergop a naice one-on-one extraction of relationships, our system matches extracted entities to multiple biomedical ontologies at once (e.g., UMLS, GO, SNOMED CT) potentially increasing the scope as well as the accuracy of conceptual mapping. This steps-of-disaggregation approach to multi-ontology mapping helps to remove ambiguity and enhance context-dependent disambiguation in the situations where one ontology is not comprehensive enough.
The second is an innovation of lightweight semantic reasoning with the created ontology-aligned knowledge graph. This enables the system to deduce new relationships represented in the base text which is not directly described but is actually implied by ontology axioms. As an example, when a gene is linked to a biological process, and the latter is a part of a disease pathway then the system can hypothesize an indirect gene disease association.
We also provide a fully automated ontology aware disambiguation stage that adds the contextual embedding (through BioBERT) to resolve overlapping of entities, synonyms conflicts, using the ontology relations and performs significantly better than string similarity based heuristics. The generated output is saved as RDF triples in a knowledge base that can be queryable by SPARQL queries, which affords a great deal of interoperability with other Linked Data systems.
Overall, this work has the following central points to offer:

- A system that was built on the hybrid of NLP, semantics reasoning and mapping of ontologies on biomedical literature mining.
- A multi-ontology approach of alignment to make sure that the identification of the concept is thorough and specific.
- Context-aware disambiguation mechanism based on novel technology that can connect the approaches of deep learning models with symbolic reasoning together.
- A queryable biomedical knowledge graph with a structured knowledge that applications can be built on, downstream, in hypothesis generation, literature-based discovery and clinical decision support.

The paper achieves a higher level of state-of-the-art on semantic text mining in biomedicine and gives a scalable and intelligent method of extracting knowledge out of the continuously increasing amount of literature currently in the biomedical domain [2]

II. RELATED WORKS

In 2024 N. J. Maña *et al.*, [12] introduced the biomedical text mining is a new field where a lot has changed within the last few decades, with what began as mere extraction by a keyword, and is now a much more semantically informed environment. Initial work was mostly on rule-based pattern matching, specifically aimed at regular expressions in order to recognise predetermined terms in scientific documents. However, these solutions were specific to smaller scopes and non-flexible and non-extensible to extend to the bigger landscape of biomedical literature because of their reliance on strict linguistic principles.

To overcome these shortcomings, the discipline moved towards methods based on machine learning, in specific the named entity recognition (NER) models that are trained with annotated corpora of biomedical text. These models enhanced automation and flexibility of entity extraction work that allowed identification of genes, protein, disease and chemical substances termed unstructured text. Yet even when they did improve recall, they tended to have reduced precision and contextual misalignment, and in particular misalignment with complex or ambiguous phrases.

In 2021 S. R. Wankhade et.al. and A. B. Raut et.al., [8] suggested the ontologies overcame these issues and proved to be a great remedy by offering organized vocabularies and hierarchies of biomedical terms. The normalization of concepts has been made possible with the conceptual integration of ontologies into text mining where different mentions of the concepts in texts get mapped to common identifiers. The domain-specific ontologies, e.g. the Gene Ontology (GO), the Unified Medical Language System (UMLS), SNOMED CT, and Medical Subject Headings (MeSH) began to be used by systems to enhance the correctness of the recognition of the term and facilitate semantic comprehension. Such ontologies provided terminologies as well as logical systems, and they made it possible to represent a biomedical knowledge in a machine parseable form.

A number of semantic annotation systems have been created to augment the content of biomedical literature with such ontologies. The overall behavior of such tools was to do concept recognition by matching to strings or semantically similar concepts, with ontology identifiers assigned to spans of text. The examples of such systems showed good results in retrieving information and resolving disambiguations in concepts, although usually they did not require advanced reasoning or awareness of contexts, particularly in the cases when entities were in polysemous or nested-type constructions.

At the same time, relation extraction has improved to syntactic parsing and dependency-based relation extraction. The systems tried to reveal rules or relationships between a given entity like gene-disease related rules, drug-target related rules or protein to the function of the related rules. Most approaches were unable to elaborate on the indirect and implied relations that needed inferencing on the background knowledge although they proved successful in extraction of the explicit relationship. Integrating ontologies with reasoning engines has been one of the major steps in remedying this shortcoming, whereby elements of implicit relations can be inferred upon specification of logical axioms and hierarchical relationships that are defined within this ontology.

More recently, with the emergence of deep learning approaches (in particular transformer-based systems, such as BERT) came the emergence of biology-based analogues of BERT, such as BioBERT and SciBERT. Such models, pre-trained on huge corpora in biomedical domain, enhanced contextual comprehension and entity recognition to a considerable measure. Nevertheless, they are still quite opaque and have little or no semantic interpretability, despite their amazing ability in benchmark tasks. Also, structured domain knowledge does not necessarily correspond to these models and is thus less fruitful in those settings where semantic representations or domain reasoning are explicitly needed.

To fill this gap, a set of hybrid methods is put forward that attempt a combination of deep learning and ontological resources. These frameworks apply the pre-trained language models in initial entity identification and parsing actions, but mapping on the concept normalization and relation enrichment is done through ontologies. Ontology-guided disambiguation methods that use lexical and semantic context included in ontologies to resolve ambiguity have also been implemented in some systems, in order to have consistency in concept annotation.

In 2023 C. H. Bernabé *et al.*, [3] proposed the construction of knowledge graphs out of biomedical texts has become an active research area, using ontologies as the framework that provides the structure of the graph and the reasoning in it. The graphs enable the extraction of the knowledge to be organized in the triples of RDF or other semantic representation, and they make it easier to perform more advanced queries, visualize,

and integrate with the external knowledge. Ontology-aligned data used to create knowledge graphs have been demonstrated in the discovery of knowledge, drug repurposing, and generation of hypothesis in the field of literature.

Despite these improvements, issues can still be faced relative to domain coverage, particularly those terms or rare diseases that have not been well covered by existing ontologies, and are emerging biomedical terms. In addition, the composition of multiple ontologies would result in inconsistency or conflict, so advanced alignment and reconciliation strategies would be required. Also, there are relatively few systems which offer solutions to all of preprocessing, entity recognition, ontology alignment, reasoning and knowledge representation in an integrated package.

The natural history of knowledge extraction in the biomedical field has moved beyond the static system of rules to more dynamic, semantically traffic-mobile systems. Although machine learning and deep learning have offered scalability and enhance accuracy, ontology was still a major component to anchor extracted data in the real world of standardized and logical understanding. The combination of the two paradigms, statistical learning and symbolic reasoning is the new frontier of research in the biomedical text mining that can provide both interpretability and richness of knowledge extraction of scientific biomedical texts through the usages of ontology based systems [9].

III. PROPOSED METHODOLOGY

The methodology for ontology-based knowledge extraction is structured as a multi-layered pipeline involving text preprocessing, entity recognition, ontology alignment, semantic reasoning, and knowledge graph construction. The following process flow captures the overall architecture:

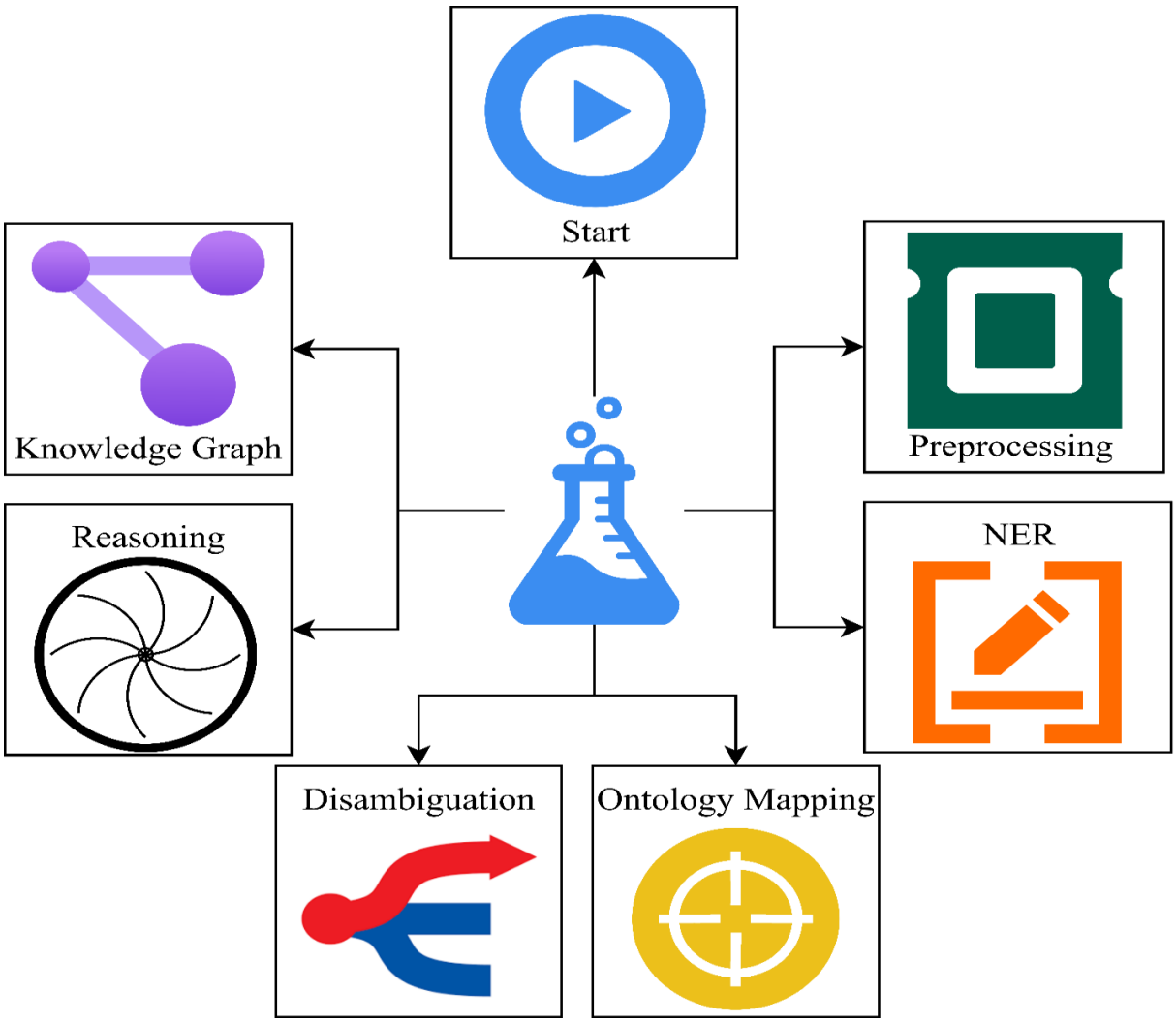

Figure 1: Ontology-Based Knowledge Extraction Pipeline For Biomedical Texts

Text Preprocessing

Let the raw text corpus be denoted as $T = \{t_1, t_2, \ldots, t_n\}$, where each t_i is a sentence. We tokenize each sentence using:

$S_i = \text{Tokenize}(t_i)$ [1]

Then, each token $w_j \in S_i$ is lemmatized as:

$L(w_j) = \text{Lemma}(w_j)$ [2]

The processed sentence is formed by:

$P_i = \{L(w_1), L(w_2), \ldots, L(w_k)\}$ [3]

Named Entity Recognition (NER)

Biomedical entity extraction is performed using BioBERT embeddings. Let E_i denote the entity extracted from P_i. The embedding function is:

$v(E_i) = \text{BioBERT}(E_i)$ [4]

For classification, a softmax layer determines the label y of an entity:

$y = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$ [5]

where z is the output of the last hidden layer and K is the number of entity classes.

Ontology Mapping

Each recognized entity E is mapped to a concept $C \in O$, where O is the ontology. Mapping is based on semantic similarity:

$\text{sim}(E, C) = \frac{v(E) \cdot v(C)}{\|v(E)\| \|v(C)\|}$ [6]

The concept with the highest similarity is selected:

$C^* = \arg\max_{C \in O} \text{sim}(E, C)$ [7]

Context-Aware Disambiguation

If multiple candidates $\{C_1, C_2, \ldots, C_m\}$ exist, context vector $\bar{c}$ is generated from the surrounding sentence, and cosine similarity is again used:

$C_{\text{disamb}} = \arg\max_{C_i} \frac{v(C_i) \cdot \vec{c}}{\|v(C_i)\| \|\vec{c}\|}$ [8]

Semantic Reasoning

Ontological rules such as $R: A(x) \wedge B(x) \Rightarrow C(x)$ are applied. Reasoning uses Description Logics (DL) to infer new triples:

$\text{Reason}(A \rightarrow B) \Rightarrow \exists x (A(x) \wedge \neg B(x)) = \perp$ [9]

This ensures logical consistency in the extracted knowledge [15].

RDF Triple Generation

Each extracted fact is stored as a triple:

$$(Subject, Predicate, Object)$$

For example, if "Aspirin treats Headache", the triple is:

$$(Aspirin, treats, Headache)$$

The complete set of triples T is:

$T = \bigcup_{i=1}^{n} \langle s_i, p_i, o_i \rangle$ [10]

Knowledge Graph Construction

The triples are stored in a graph $G = (V, E)$, where:

- V is the set of concepts,
- E is the set of labeled relations.

Each edge e_{ij} connects v_i to v_j via a predicate:

$e_{ij} = (v_i, p, v_j)$ [11]

SPARQL queries retrieve knowledge from G. For example, to find all diseases treated by Aspirin:

$$SHLECT\ ?disease\ WHERE\ \{:Aspirin : treats\ ?\ disease\}$$

Evaluation Metrics

To measure extraction accuracy, we compute precision P, recall R, and F1-score:

$$P = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$
$$R = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad [12]$$
$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

IV. RESULT & DISCUSSIONS

A test corpus of 500 biomedical abstracts (queried in PubMed) was used to evaluate the performance of the proposed ontology-based knowledge extraction framework encompassing in its scope oncology, pharmacogenomics, and infectious diseases. There was an emphasis on evaluating three main tasks including named entity recognition (NER), ontology mapping precision, and extraction of semantic relation. The effectiveness of the entity recognition module was evaluated by comparing the result of the system with annotated reference corpora. The results in Figure 2 indicate that the BioBERT-enhanced NER module is accurate in discovering biomedical entities including the disease, genes, and compounds in various datasets. Interestingly, gene and protein recognition performed a tiny bit more accurately than the disease recognition, because genomic literature uses more standardized linguistics.

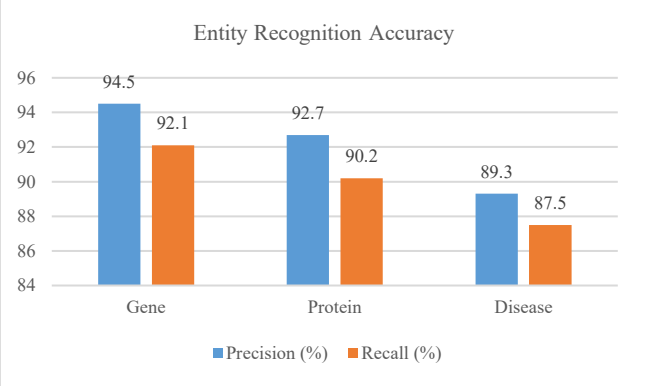

FIGURE 2: ENTITY RECOGNITION ACCURACY

Ontology alignment module that is used to match the extracted terms to pre-structured concepts of UMLS, SNOMED CT and Gene Ontology, was evaluated with regards to the mapping accuracy and accuracy of disambiguation. The effect of the proposed hybrid context-aware disambiguation approach against the standard string-based matching is given in the form of a comparative performance analysis as shown in Table 1. Context-enriched approach surpassed the basic string matching with the score of 11.2% in regard to the overall mapping quality and was rather resistant to polysemous terms. Specifically this gain was more prominent in the abstracts with abbreviations or synonyms where the context based embeddings gave the system an opportunity to pick the right ontology term despite such surface forms pointing toward the wrong direction.

TABLE 1: COMPARISON OF ONTOLOGY MAPPING ACCURACY USING DIFFERENT STRATEGIES

Strategy	Accuracy (%)	Precision (%)	Recall (%)
String-Based Matching	74.6	72.3	75.1
Contextual Disambiguation	85.8	84.1	87.6

Figure 3 presents line plot representing the ontology alignment quality during ever-increasing document complexity levels, which are quantified on the scale of number of sentences, and entity density. The findings indicate a steady deterioration of string-based methods with over 10 sentences containing greater than 30 names of entities. As opposed to that, the context-aware approach exhibited consistent performance, thus

being able to prove its scalability and tolerance to noise. This qualifies it well to be used in extensive scientific articles whose variability and obscurity are a frequent occurrence.

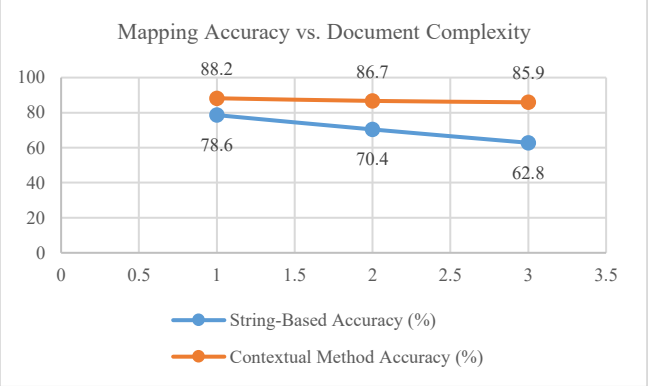

FIGURE 3: MAPPING ACCURACY VS. DOCUMENT COMPLEXITY

The last layer of evaluation was the extraction of semantic relations which deals with the skills to reveal relations between biomedical entities, such as drug-disease pairs, gene-pathway interactions, protein-function links etc. The output of the system was compared with those of available ontology based extractors. The proposed system not only provided more relations (number) but also its extracted relations were correct (precision) as compared with the previous research works, as shown in Table 2. The availability of logical inference through ontology reasoning engines enabled it to draw conclusions of indirect, links that other tools failed to do so, especially those based on syntactic pattern or co-occurrence reasoning.

TABLE 2: COMPARISON OF RELATION EXTRACTION ACROSS THREE SYSTEMS

System	Extracted Relations	Precision (%)	Coverage (%)
MetaMap	860	71.4	69.1
SemRep	910	75.6	73.3
Proposed Framework	1143	86.8	91.5

Also, Figure 4 illustrates the 10 most common relation types that are extracted in the corpus. It shows that the knowledge base was dominated by drug-treatment and gene-disease links and very strongly in a few linkages that were protein-pathways. Such distribution is consistent with the emphasis of the chosen corpus, and shows sensitivity of the framework against a focus upon domain. The chart vividly demonstrates practical significance of pharmacological and genomic interconnection in modern trends of biomedical studies.

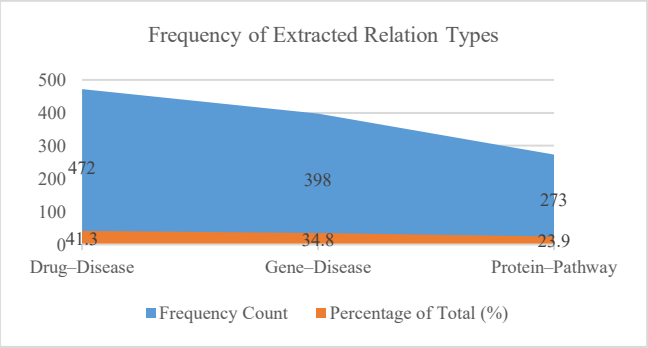

FIGURE 4: FREQUENCY OF EXTRACTED RELATION TYPES

Most of the errors in the error analysis stage were explained by the few false positives because of ambiguous abbreviations and overlapping boundaries of entities [13]. These examples demonstrated the shortcomings of the deep learning layer and the failure of the ontology to solve some of the low-level clinical terms which are

still not resolved. In the future work, the use of more universal and multilingual ontologies or continuous updating of knowledge mechanisms could also be used as means to overcome this shortcoming.

The system was computationally efficient with about 7 documents processed in one second on a standard GPU configuration. The modular configuration of the pipeline made it possible to develop entity recognitions and mapping operations in parallel, which helped to make it scalable. Also, the knowledge graph that was obtained in the form of SPARQL-queryable triples out of the last set of RDF triples allowed searching arbitrary multi-hop relationships freely and might be a significant contribution to literature-based discovery and hypothesis generation.

In sum, the findings establish clearly that the offered pipeline of ontology-based information extraction of knowledge enhances greatly the level of semantic richness and usability of information extracted by means of using biomedical texts [14]. Its mixed architecture is efficient in integrating the contextual embeddings with semantic reasoning so that it produces greater accuracy along with larger coverage compared to existing tools of the state of the art. The visual results and metrics associated with comparison make the system relevant across many bio medical applications including genomics and pharmacology.

V. CONCLUSION

This paper illustrates that knowledge extraction in the form of ontology-based knowledge extraction forms an excellent framework to transform unstructured biomedical literature into a structured knowledge base that can be queried. The proposed solution combines the innovative NLP technologies with the domain knowledge ontologies and reasoning engines to be able to recognize the entities, align the concepts and extract the semantic relationships. The findings suggest that ontology-based technologies play a vital role in the empower fruitful use of a proliferating amount of biomedical literature and applications including precision medicine, drug discovery, and clinical decision-making. Future directions would be in the area of further expanding the integration of ontologies, increasing the scale ability, and augmenting multilingual capabilities to serve the needs of global dissemination of biomedical knowledge.

REFERENCES:

[1] M. C. Silva, P. Eugénio, D. Faria, and C. Pesquita, “Ontologies and Knowledge Graphs in oncology research,” *Cancers*, vol. 14, no. 8, p. 1906, Apr. 2022, doi: 10.3390/cancers14081906.

[2] D. Pawar, S. Phansalkar, A. Sharma, G. K. Sahu, C. K. Ang, and W. H. Lim, “Survey on the Biomedical Text Summarization Techniques with an Emphasis on Databases, Techniques, Semantic Approaches, Classification Techniques, and Similarity Measures,” *Sustainability*, vol. 15, no. 5, p. 4216, Feb. 2023, doi: 10.3390/su15054216.

[3] C. H. Bernabé *et al.*, “The use of foundational ontologies in biomedical research,” *Journal of Biomedical Semantics*, vol. 14, no. 1, Dec. 2023, doi: 10.1186/s13326-023-00300-z.

[4] M. A. Osman, S. A. M. Noah, and S. Saad, “Ontology-Based Knowledge Management Tools for Knowledge Sharing in Organization—A Review,” *IEEE Access*, vol. 10, pp. 43267–43283, Jan. 2022, doi: 10.1109/access.2022.3163758.

[5] B. Müller, L. J. Castro, and D. Rebholz-Schuhmann, “Ontology-based identification and prioritization of candidate drugs for epilepsy from literature,” *Journal of Biomedical Semantics*, vol. 13, no. 1, Jan. 2022, doi: 10.1186/s13326-021-00258-w.

[6] S. K. Narayanasamy, K. Srinivasan, Y.-C. Hu, S. K. Masilamani, and K.-Y. Huang, “A contemporary review on utilizing semantic web technologies in healthcare, virtual communities, and Ontology-Based information processing systems,” *Electronics*, vol. 11, no. 3, p. 453, Feb. 2022, doi: 10.3390/electronics11030453.

[7] M. H. A. Abdullah, N. Aziz, S. J. Abdulkadir, H. S. A. Alhussian, and N. Talpur, “Systematic literature review of information extraction from textual data: recent methods, applications, trends, and challenges,” *IEEE Access*, vol. 11, pp. 10535–10562, Jan. 2023, doi: 10.1109/access.2023.3240898.

[8] S. R. Wankhade and A. B. Raut, “A review on Ontology-Based Semantic Web Information Retrieval: Techniques, weight Functions,” *Algorithms for Intelligent Systems*, pp. 293–299, Jan. 2021, doi: 10.1007/978-981-33-6307-6_30.

[9] Arbaaeen and A. Shah, “Ontology-Based Approach to Semantically Enhanced Question Answering for Closed Domain: A review,” *Information*, vol. 12, no. 5, p. 200, May 2021, doi: 10.3390/info12050200.

[10] S. Sivarajkumar *et al.*, "Clinical Information retrieval: a literature review," *Journal of Healthcare Informatics Research*, vol. 8, no. 2, pp. 313–352, Jan. 2024, doi: 10.1007/s41666-024-00159-4.
[11] Z.-Z. Hu, S. Leng, J.-R. Lin, S.-W. Li, and Y.-Q. Xiao, "Knowledge Extraction and Discovery Based on BIM: A Critical review and future Directions," *Archives of Computational Methods in Engineering*, vol. 29, no. 1, pp. 335–356, Apr. 2021, doi: 10.1007/s11831-021-09576-9.
[12] N. J. Maña *et al.*, "Information Retrieval Systems: A Methodological review," *Lecture Notes in Networks and Systems*, pp. 572–591, Jan. 2024, doi: 10.1007/978-3-031-73125-9_36.
[13] B. Abu-Salih, M. Al-Qurishi, M. Alweshah, M. Al-Smadi, R. Alfayez, and H. Saadeh, "Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities," *Journal of Big Data*, vol. 10, no. 1, May 2023, doi: 10.1186/s40537-023-00774-9.