Human Systems Engineering and Design (IHSED 2026), Vol. 237, 2026, 479–484
https://doi.org/10.54941/ahfe1008187

Scalable Extraction and Normalization of Biomedical Knowledge From Research Literature

Supraja Krovvidi, Finn Vos, Ramesh Jakka, Laurent Hasson, and Shloke Meresh

CapsicoHealth, USA

Abstract

The growth of biomedical literature poses significant challenges for researchers conducting systematic and scoping reviews. In fields such as digital biomarkers for the treatment of heart failure and cardiovascular disease, manually screening thousands of papers is time-consuming and not scalable. To address this problem, we developed AI-assisted tools for large-scale analysis and structured knowledge extraction from biomedical research articles, building on established knowledge graph methodologies for organizing scientific data (Bodenreider, 2004). We developed a knowledge graph generation pipeline that extracts subject-predicate-object triplets representing scientific claims from research articles, following biomedical relation extraction approaches similar to BioBERT-based systems (Ji et al., 2022). To operationalize this pipeline at scale, our implementation ingests PDF documents from local or cloud-based storage, performs page-level text extraction, and segments content into overlapping sentence-aware chunks prior to language model processing. To address redundancy caused by linguistic variation across documents, we implemented a multi-stage consolidation pipeline that extends beyond lexical normalization to include semantic and ontology-based merging. Unlike earlier approaches that relied primarily on string matching and exact deduplication, our updated system introduces a downstream normalization stage that consolidates semantically equivalent entities and relations across documents. This approach builds upon embedding-based similarity techniques such as Sentence-BERT (Hogan et al., 2021) and ontology alignment methods from UMLS-based biomedical systems (Devlin et al., 2019). The pipeline integrates lexical normalization, exact deduplication, semantic merging using embedding similarity and Jaccard similarity, and ontology-based alignment using vocabularies such as MeSH and RxNorm, reducing both lexical redundancy and conceptual duplication while producing a more compact and interoperable knowledge graph representation. We evaluated our approach on a corpus of biomedical research articles, processing thousands of extracted triplets. The normalization pipeline significantly reduced redundant entity variants and improved graph clarity while preserving semantic meaning. In conclusion, the integration of lexical, semantic, and ontology-based consolidation provides a scalable framework for cross-document knowledge synthesis, enabling more efficient systematic reviews and advanced querying across biomedical literature.

Keywords: Knowledge graphs, AI, LLMs, Normalization, Entities, Relations, Biomedical, Digital health technology, Digital biomarker

Received May 15, 2026 Revised June 24, 2026 Accepted July 7, 2026 Available online September 25, 2026

INTRODUCTION

The volume of biomedical literature has grown exponentially, making it increasingly difficult for researchers to synthesize existing knowledge efficiently, as documented in large-scale studies of scientific publication growth (Devlin et al., 2019).

Traditional methods of literature review rely heavily on manual screening and interpretation, which are time-consuming and inconsistent. While recent advances in natural language processing (NLP), particularly transformer-based models such as BERT (Himmelstein et al., 2017), have enabled automated information extraction, these methods often produce unstructured or redundant outputs that are difficult to aggregate across multiple documents.

Knowledge graphs provide a promising framework for organizing scientific information into structured, machine-readable representations, as demonstrated in prior surveys on knowledge graph systems (Bodenreider, 2004). However, constructing knowledge graphs from large-scale research introduces a key challenge: redundancy due to linguistic variation. The same scientific claim may appear across multiple documents with different wording, leading to duplication, inflated evidence counts, and reduced interpretability (Hogan et al., 2021).

In this paper, we present a system that combines AI-powered literature analysis, knowledge graph construction, and a multi-stage consolidation pipeline designed to reduce semantic redundancy across documents.

RELATED WORK

Biomedical Literature Mining

Prior work in biomedical NLP has focused on named entity recognition, relation extraction, and document classification. Systems such as BioBERT (Ji et al., 2022) have enabled large-scale extraction of biomedical relationships but often lack structured aggregation across documents.

Knowledge Graph Construction

Knowledge graphs have been widely used in biomedical domains to represent relationships between diseases, treatments, and biomarkers. However, many approaches rely on curated datasets rather than automated extraction from raw literature (Kilicoglu et al., 2012). Systems such as Hetionet (Lee et al., 2020) and SemMedDB (Reimers and Gurevych, 2019) demonstrate the effectiveness of structured biomedical graphs but depend heavily on predefined relationships.

Entity Resolution and Deduplication

Entity resolution and semantic deduplication are well-studied problems in NLP and databases. Embedding-based methods, such as Sentence-BERT (Hogan et al., 2021), have improved the ability to detect semantically equivalent entities. However, cross-document consolidation of extracted scientific claims remains an open challenge.

METHODOLOGY

Our process consisted of three overarching stages: triplet extraction, knowledge graph creation, and triplet merging.

From each document, we extracted structured representations of scientific claims in the form of (subject, predicate, object). These triplets capture relationships between disease-biomarker associations, treatment effects, and clinical outcomes. Extraction was performed using LLM-based methods, following structured output constraints like biomedical NLP systems (Ji et al., 2022). To improve scalability and preserve context, full-text articles were segmented into overlapping sentence-aware chunks prior to extraction. This ensures consistent claim extraction across long documents.

Knowledge Graph Creation

Extracted triplets were aggregated into knowledge graphs, where nodes represent entities, and edges represent relationships. Graphs were constructed across documents for global synthesis. At this stage, graphs exhibited significant redundancy due to repeated or paraphrased claims.

Triplet Merging Pipeline

Initial knowledge graphs contained duplicate triplets, paraphrased claims, and lexical variations representing identical relationships. The original pipeline primarily performed lexical normalization and exact deduplication. While effective for removing trivial redundancy, this approach did not address semantically equivalent claims expressed differently across documents. The updated system introduces a multi-stage consolidation pipeline that operates after triplet extraction and initial graph construction.

This pipeline consists of three key steps:

1. **Lexical Normalization and Exact Deduplication**

   Entity strings are standardized through lowercasing, stop word removal, acronym resolution, and punctuation cleanup. This reduces surface-level variation. For the entity deduplication stage, triplets that are identical after normalization are removed.

2. **Semantic Consolidation**

   Semantically similar entities and relations are merged using embedding-based similarity and Jaccard similarity, extending approaches such as Sentence-BERT (Hogan et al., 2021). This allows paraphrased claims across documents to be unified.

3. **Ontology-Based Alignment**

   Entities are optionally mapped to standardized biomedical concepts using ontologies such as MeSH and RxNorm, following methods used in UMLS-based systems (Devlin et al., 2019). Unlike the original pipeline, which focused on string-level similarity, this updated approach enables cross-document concept consolidation, reducing both lexical redundancy and semantic duplication. SNOMED CT was also considered through UMLS as part of

the concept normalization strategy; however, it could not be incorporated in this implementation because of licensing restrictions.

Additionally, the system introduces improvements for scalability and robustness, including model prewarming, concurrent document processing, configurable normalization thresholds, and per-phase timing instrumentation for performance analysis. The final pipeline integrates lexical, semantic, and ontology-based merging to produce a more compact and interpretable knowledge graph.

RESULTS

Across 76 processed documents, the system produced 8,207 initial triplets, which decreased to 6,297 after entity normalization and claim-level merging, corresponding to a 23.3% reduction in triplet volume. This reduction indicates effective removal of redundant and variationally expressed entities while preserving relation-level information, resulting in a more compact and analytically usable knowledge graph.

Qualitatively, the merged graph showed improved structural clarity and reduced semantic fragmentation, with core claim meaning retained across normalized entities and consolidated assertions. Together, these findings suggest that entity normalization substantially improves graph efficiency and interpretability without compromising semantic fidelity, supporting its value as a critical post-extraction step in knowledge graph construction.

DISCUSSION

The transition from lexical to semantic merging highlights a key limitation of surface-level approaches. Merging based on lexical similarity alone is insufficient for capturing conceptual equivalence in scientific literature. The introduction of a downstream semantic consolidation stage allows the system to merge conceptually equivalent claims across documents, improving the reliability of aggregated knowledge.

Implications

Entity normalization has important downstream implications for evidence-centric knowledge graphs. By collapsing duplicate and surface-variant representations of the same entities and assertions, it prevents artificial inflation of evidence counts that can otherwise bias claim prevalence estimates. This normalization step also enables more reliable aggregation of scientific claims across documents, improving the validity of cross-study synthesis and trend analysis. In practice, the resulting graph is easier to interpret because semantically related evidence is organized under consistent entity identities, allowing researchers to trace support, compare findings, and identify contradictions with greater confidence.

CHALLENGES AND LIMITATIONS

Entity normalization in knowledge graphs remains subject to methodological limitations. Outcomes are sensitive to similarity thresholds: conservative settings can under-merge equivalent entities, while aggressive settings can over-merge distinct concepts. This is amplified by semantic ambiguity in biomedical text, where identical terms can be context-dependent, and near-synonyms are not always interchangeable. There is therefore an inherent tradeoff between graph compactness and semantic precision.

A further limitation is dependence on embedding quality and ontology coverage. If embeddings miss domain-specific distinctions, or ontologies are incomplete, normalization errors can propagate and reduce claim fidelity; aggressive merging can also suppress meaningful nuance. Practical challenges include cross-ontology mapping and weak or absent bridges between ontologies. For example, certain domain-specific concepts, such as radioligand therapy, may be incompletely represented in MeSH and RxNorm, potentially limiting normalization performance. For these reasons, human-in-the-loop validation remains necessary, especially for high-impact or borderline merge decisions.

FUTURE WORK

Future work should focus on improving both robustness and downstream utility of entity-normalized knowledge graphs. A key priority is adaptive threshold optimization, where merge criteria are dynamically calibrated by entity type, context, and evidence density rather than fixed globally. Extending the framework to temporal knowledge graphs would also enable representation of how claims evolve over time, including shifts in evidence strength and potential contradictions across publication periods.

In parallel, tighter integration with retrieval-augmented generation systems could make normalized graphs directly actionable for evidence-grounded question answering and scientific synthesis. Additional gains are likely from expanded ontology integration to improve coverage of domain-specific terminology and reduce normalization gaps. Where ontology terms are missing, a curated manual mapping layer should also be incorporated to improve normalization quality. Finally, confidence scoring for extracted relationships should be added to quantify uncertainty, support quality-aware downstream analysis, and prioritize human review for low-confidence or high-impact claims.

CONCLUSION

We present a scalable system for AI-assisted literature analysis and knowledge graph construction in biomedical research. By introducing a multi-stage consolidation pipeline that extends from lexical normalization to semantic and ontology-based merging, we address a fundamental challenge in multi-document knowledge graph construction: redundancy due to linguistic variation. This approach enables more accurate and scalable synthesis of biomedical knowledge.

REFERENCES

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 32(Database issue), D267–D270.

Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology, 66(11), 2215–2222.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171–4186.

Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S.E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife, 6, e26726.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J.E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge Graphs. ACM Computing Surveys, 54(4), 1–37.

Ji, S., Pan, S., Cambria, E., Marttinen, P., and Yu, P.S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems, 33(2), 494–514.

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., and Rindflesch, T.C. (2012). SemMedDB: A PubMed-scale repository of biomedical semantic predications. Bioinformatics, 28(23), 3158–3160.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992.

Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724–2743.