CE-KG: Citation-enhanced Knowledge Graph via Citation Sentiment Fusion and Evidence Tracing

Yalan Huang[1], Xuemei Yang[1], Bin Zhang[1] and Xiaoli Tang[1,*]

[1]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

Abstract

This study proposes a knowledge graph (KG) construction method integrating citation information to address credibility tracing of knowledge claims. By fine-tuning large language models (LLMs), SPO triples are precisely extracted from the "Conclusion" sections of literature abstracts. A dual-level fusion mechanism is developed based on sentiment analysis (Positive/Neutral/Negative) of citation contexts. This approach injects academic evaluation attributes at the paper level and links them to SPO triples. An innovative Neo4j-MySQL heterogeneous storage architecture is designed, enabling fine-grained evidence tracing from KG relationships to citation contexts through uniform identifiers. The constructed KG simultaneously provides knowledge claims and supports credibility tracing from the academic community perspective.

Keywords

Knowledge Graph, Citation Sentiment Analysis, Evidence Tracing, Credibility Verification

1. Introduction

Biomedical literature serves as a critical carrier of scientific research achievements, documenting major breakthroughs, knowledge discoveries, and direct empirical evidence from clinical studies. Current studies typically construct domain knowledge graphs by extracting entity relationships from titles and abstracts[1, 2, 3]. However, this approach exhibits two key limitations: First, while abstracts summarize academic knowledge production processes, the Conclusion section contains investigated knowledge claims, arguments, and assertions[4], where other sections may interfere with core claims. Second, identical entity-relationship pairs from different sources have unequal credibility. Notably, citations are vital pathways for disseminating knowledge claims and establishing credibility[5]. The sentiment polarity (e.g., Positive/Neutral/Negative) in citation contexts reflects citing authors' academic attitudes toward cited claims[6], serving as a key credibility source.

To overcome these limitations and leverage citation sentiment, this study proposes a KG framework integrating evidence from citation information (including sentiment and context). Core innovations include:

1. Structured Evidence Screening: SPO triples are strictly extracted from "Conclusion" sections of scientific abstracts, ensuring precise reflection of research claims while minimizing information noise.

2. Citation Information Fusion: Deep learning models parse sentiment polarity in original citation contexts to quantify academic evaluations between papers. Citation sentiments, contexts, and SPO triples are fused to enable credibility tracing. These mechanisms aim to transform static KGs into tools supporting knowledge claim credibility tracing.

These mechanisms transform static KGs into tools supporting knowledge claim credibility tracing. The code and data used in our Knowledge Graph all available at https://github.com/wang-lch/CE-KG.

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

*Corresponding author.

huang.yalan@imicams.ac.cn (Y. Huang); yang.xuemei@imicams.ac.cn (X. Yang); zhang.bin@imicams.ac.cn (B. Zhang); tang.xiaoli@imicams.ac.cn (X. Tang)

0009-0000-1638-7797 (Y. Huang); 0000-0002-2927-4166 (X. Yang); 0009-0008-5212-3375 (B. Zhang); 0000-0001-6946-3482 (X. Tang)

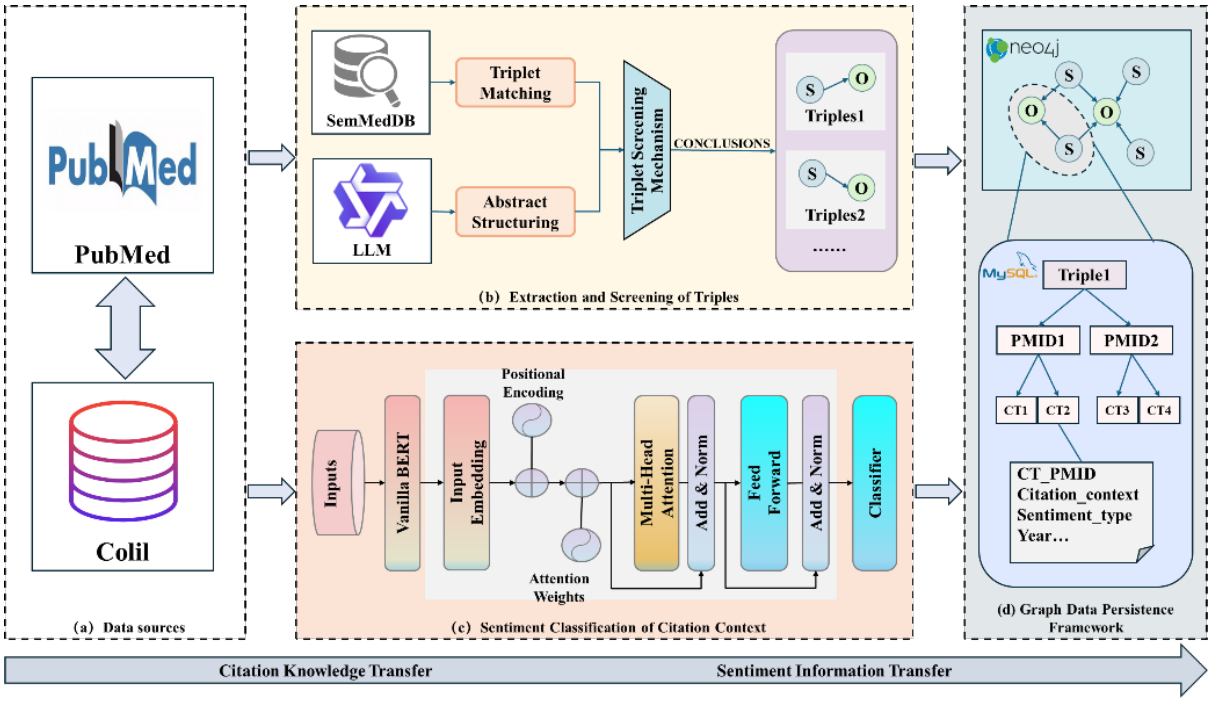

Figure 1: Citation-enhanced Knowledge Graph Construction Process

2. Methodology

2.1. Data Collection and Processing

This study establishes a multimodal biomedical knowledge base integrating three heterogeneous data sources: Metadata of 9,201 clinical research papers on breast cancer drug therapy (including trials and guidelines) published from 2000-2024 were retrieved from PubMed; 83,510 medical SPO triples were extracted from the SemMedDB[7] using PMIDs; 127,927 valid citation contexts were collected via the Colil[8] database API. To enhance extraction reliability, the fine-tuned LLM Qwen3-0.6B[9] was adopted for abstract structuring. This model automatically classifies content into Background, Objective, Methods, Results, and Conclusion sections. Training utilized 100,000 biomedical abstracts with structured labels, with optimization parameters set to learning rate 1e-5, batch size 4, and random seed 42. After three iterations, the model achieved a 0.94 F1-score for Conclusion sections on the test set. SPO triples were strictly extracted from "Conclusion" sections to focus on core research claims and minimize noise. This process yielded 14,223 deduplicated triples.

2.2. Citation-enhanced Knowledge Graph Construction

An evidence-oriented filtering strategy was adopted, using only Conclusion-derived SPO triples to ensure claims reflect empirical findings. For confidence quantification, the Dict-Senti-BERT[10] model identified sentiment polarity (Positive/Neutral/Negative) in citation contexts. Sentiments were injected through dual-level fusion: Paper-level calculations determined Positive/Negative sentiment proportions; SPO Triples-level integration embedded source literature and sentiment distributions as provenance attributes. This enables knowledge claim tracing and evidence queries. Model training employed optimization parameters of learning rate 5e-6, batch size 32, and weight decay 0.01 over 50 epochs. After training, the model achieved an overall F1-score of 0.93 and an accuracy of 0.94 on the held-out test set, demonstrating stable performance in conclusion-oriented triple extraction and sentiment-enhanced integration

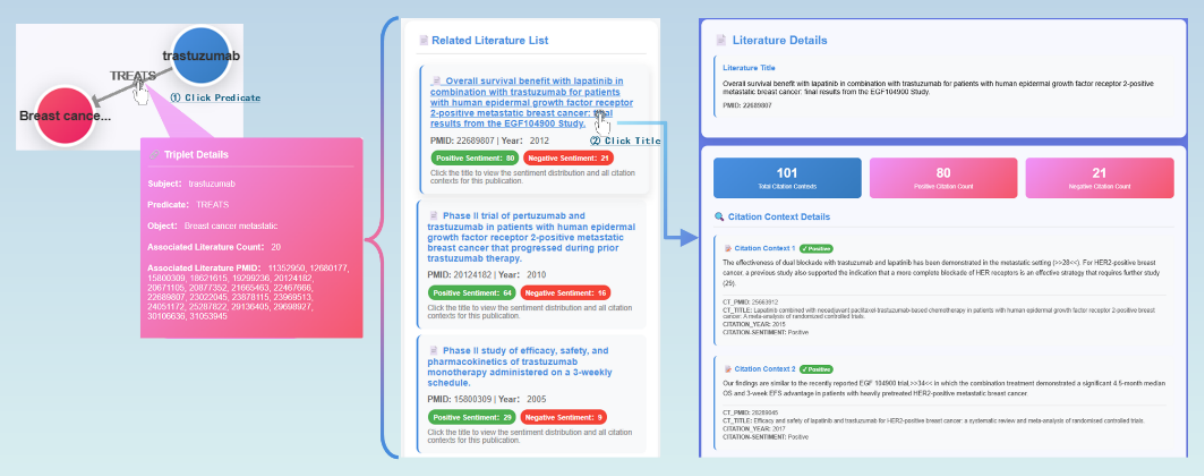

Figure 2: Citation-enhanced Knowledge Graph Visualization Solution

2.3. KG Storage and Update Mechanism

A Neo4j-MySQL heterogeneous storage architecture enables dynamic knowledge network storage and fine-grained evidence tracing. Neo4j stores normalized entities (e.g., drugs, diseases) as nodes and SPO predicates (e.g., TREATS, CAUSES) as directed relationships with embedded evidence support (PMID lists and sentiment distributions). MySQL maintains structured evidence: The Papers table records metadata (titles, structured abstracts), while the Citations table stores citation types and context snippets. A unified identifier system (Triple_ID/PMID/Citation_ID) enables cross-database mapping from KG relationships to specific citation contexts, forming an integrated solution supporting knowledge evolution and claim verification.

2.4. Knowledge Graph Visualization

This study designs a hierarchical knowledge graph (KG) visualization scheme enabling traceable analysis from knowledge claims to microscopic evidence through three progressive interfaces. The workflow begins at the semantic triple layer where clicking a predicate edge of any SPO triple triggers a detailed attribute panel. This panel displays the PMID list of associated literature. The system then dynamically renders a literature overview layer (central expansion panel). Each paper presents structured data including title, PMID, publication year, and sentiment polarity distribution (e.g., positive/negative citation counts). Finally, the citation layer comprehensively displays all citation contexts of individual papers: Total citation statistics, sentiment classification data, and sentiment-labeled citation contexts with associated metadata (source PMID, title, year) are systematically presented. This forms a complete evidence tracing chain, facilitating credibility verification of knowledge claims.

3. Conclusion and Future Work

This study proposes a knowledge graph (KG) construction method integrating citation sentiment and contextual evidence. The evidence-oriented mechanism—strictly extracting SPO triples only from abstract "Conclusion" sections—significantly improves knowledge claim accuracy. Innovatively incorporating citation information enables credibility tracing for academic knowledge discovery. The constructed KG links knowledge claims to source literature's citation data, allowing examination of academic community evaluations and facilitating access to high-credibility claims.

Future work will employ deep learning tools to extract multi-source citation data, complementing the Colil database to enrich academic evaluations. An alignment mechanism between citation contexts and SPO triples will be explored for precise evidence tracing. Additionally, applicability assessment in

other domains will be conducted.

Acknowledgments

This research was funded by the Innovation Fund for Medical Sciences of the Chinese Academy of Medical Sciences (Grant number: 2021-I2M-1-033)

Declaration on Generative AI

During the preparation of this manuscript, the authors employed ChatGPT and Grammarly to support grammar correction, spelling refinement, and overall language polishing. These tools were used to enhance clarity and professionalism in writing. The authors subsequently reviewed and revised all content, and took full responsibility for the accuracy and integrity of the final publication.

References

[1] Y. Nian, X. Hu, R. Zhang, J. Feng, J. Du, F. Li, L. Bu, Y. Zhang, Y. Chen, C. Tao, Mining on alzheimer's diseases related knowledge graph to identify potential AD-related semantic triples for drug repurposing, BMC Bioinformatics 23 (2022).

[2] J. Li, J. Gao, B. Feng, Y. Jing, PlagueKD: a knowledge graph–based plague knowledge database, Database 2022 (2022).

[3] S. Jin, H. Liang, W. Zhang, H. Li, Knowledge graph for breast cancer prevention and treatment: Literature-based data analysis study, JMIR Medical Informatics 12 (2024).

[4] X. Guo, Y. Chen, J. Du, E. Dong, Extracting and measuring uncertain biomedical knowledge from scientific statements, Journal of Data and Information Science 7 (2022) 6–30.

[5] M. J. Sarol, S. Ming, S. Radhakrishna, J. Schneider, H. Kilicoglu, Assessing citation integrity in biomedical publications: corpus annotation and NLP models, Bioinformatics 40 (2024).

[6] X. Wang, D. Zhao, Review on progress of citation sentiment identification, Information Studies: Theory & Application 47 (2023) 173–181+189.

[7] SemMedDB database details, https://ii.nlm.nih.gov/SemRep_SemMedDB_SKR/dbinfo_FML.shtml, 2025.

[8] Colil: Comments on literature in literature, https://colil.dbcls.jp/browse/papers/, 2025.

[9] Qwen, https://github.com/QwenLM, 2025.

[10] A. Daowd, M. Barrett, S. Abidi, S. S. R. Abidi, Building a knowledge graph representing causal associations between risk factors and incidence of breast cancer, in: Public Health and Informatics, IOS Press, 2021, pp. 724–728.