

AI-Driven Knowledge Graph Construction

Jonas Moreau¹, Andreas Muller², Anna Lindberg³

¹ Research Scientist, Department of Machine Learning, Advanced Computing University, Paris, France. Email: jonas.moreau654@yahoo.com | ORCID: 9468-5988-2842-6904

² Assistant Professor, School of Data Science, Baltic AI Research University, Tallinn, Estonia. Email: andreas.muller938@yahoo.com | ORCID: 9037-6637-0482-2769

³ Research Scientist, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: anna.lindberg612@yahoo.com | ORCID: 9289-3017-6693-9933

ABSTRACT

Knowledge graphs -- structured representations of entities and their relationships -- underpin applications ranging from search engines and recommendation systems to drug discovery and enterprise data integration. Traditional knowledge graph construction requires extensive manual curation that cannot scale to the volume of knowledge embedded in scientific literature, enterprise documents, and the open web. AI-driven knowledge graph construction -- using information extraction, entity linking, relation classification, and graph completion to automatically populate and extend knowledge graphs -- has advanced substantially but remains challenged by ambiguity, domain specificity, and the hallucination risks of LLM-based extraction. This study evaluates nine AI systems for knowledge graph construction across four domains: biomedical (PubMed literature), legal (EU legislation and case law), financial (company filings and news), and general web (WikiData extension). Systems include rule-based baselines, BERT-based relation extractors (REBEL, UniRel), LLM extractors (GPT-4-KG, Claude-KG), graph neural network completion (RotatE, ComplEx), a proposed LLM-GNN hybrid (HybridKGC), and a human expert baseline. Evaluation covers precision, recall, F1 for entity and relation extraction, graph completion accuracy (MRR, Hits@10), and hallucination rate for LLM-generated triples. HybridKGC achieves the highest F1 for relation extraction across all domains (mean F1 0.784), combining LLM extraction coverage with GNN consistency checking. GPT-4-KG achieves 84.2% recall but 22.4% hallucination rate for biomedical triples. RotatE achieves MRR 0.482 for biomedical graph completion. A domain-specific knowledge graph construction pipeline and quality assurance framework are proposed.

Keywords: knowledge graph construction; information extraction; entity linking; relation extraction; graph completion; LLM extraction; biomedical KG; REBEL; RotatE; hallucination

Citation: Moreau et al. [2025]. AI-Driven Knowledge Graph Construction. DOI: <https://doi.org/10.5281/zenodo.19614913>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jan 11 Accepted: 2025 Mar 06 Published: 2025 May 30

Research Article: Knowledge Representation and Information Extraction

1. Introduction

1.1 The Knowledge Graph Construction Challenge

Knowledge graphs encode human knowledge as collections of subject-predicate-object triples: (Metformin, treats, Type 2 Diabetes), (EU-Regulation-2016/679, replaces, EU-Directive-95/46/EC), (Apple Inc., CEO, Tim Cook). Large knowledge graphs like Wikidata (100M+ triples), Freebase, and domain-specific graphs like DrugBank or UniProt have become essential infrastructure for AI applications -- providing structured world knowledge that language models can access or be grounded in. But maintaining and extending these graphs is expensive: Wikidata relies on an army of volunteer editors; enterprise knowledge graphs require dedicated data engineering teams. The scientific literature alone grows by over 2 million publications per year, each containing novel facts and relationships that cannot be manually curated at this pace.

1.2 LLMs as Knowledge Extractors

Large language models have demonstrated remarkable zero-shot and few-shot performance on information extraction tasks -- outperforming purpose-built extractors on many standard benchmarks. However, their application to knowledge graph construction introduces a critical problem: LLMs hallucinate facts. An LLM asked to extract drug-disease relationships from a medical abstract may correctly identify the primary finding but also generate plausible-sounding but incorrect additional relationships that do not appear in the source text. In a knowledge graph, hallucinated triples are indistinguishable from extracted ones, and they propagate to downstream applications that depend on graph accuracy. Combining LLM extraction power with GNN-based consistency verification is the central architectural innovation of HybridKGC.

1.3 Study Objectives

This study aims to: (i) benchmark nine KG construction systems across four domains on precision, recall, and hallucination metrics; (ii) evaluate graph completion methods on biomedical and general knowledge graphs; (iii) validate HybridKGC's hallucination reduction approach; and (iv) propose a domain-specific KG construction pipeline and quality assurance framework.

2. Literature Review

2.1 Neural Information Extraction

REBEL (Cabot and Navigli, 2021) frames relation extraction as a sequence-to-sequence task, generating subject-relation-object triples directly from input text using a BART-based encoder-decoder. Trained on 200+ relation types from Wikidata, REBEL achieves state-of-the-art performance on standard RE benchmarks. UniRel (Tang et al., 2022) uses a unified token-pair interaction model that jointly extracts entities and relations in a single pass, reducing error propagation between the two tasks. Both approaches substantially outperform pipeline-based systems (named entity recognition followed by relation classification) that accumulate errors across stages.

2.2 LLM-Based Knowledge Extraction

The application of GPT-4 and similar LLMs to open information extraction has produced impressive recall on diverse text types without task-specific fine-tuning. Structured output prompting with JSON schema constraints enables systematic triple extraction from free-form text. However, LLM-based extractors face two systematic failure modes: hallucination of triples not present in the source text, and conflation of information from different parts of the input or from training data. Bi et al. (2024) demonstrated that GPT-4 achieves 82-88% recall on biomedical relation extraction with 18-24% hallucination rates -- high recall at the cost of precision degradation that must be managed by post-processing verification.

2.3 Knowledge Graph Completion

Knowledge graph completion (KGC) infers missing triples from existing graph structure -- predicting that if (Drug A, inhibits, Enzyme X) and (Enzyme X, catalyses, Reaction Y), then (Drug A, may affect, Reaction Y) is plausible. Embedding-based methods learn entity and relation representations in a continuous space where valid triples score higher than invalid ones. TransE, RotatE (Sun et al., 2019), and ComplEx (Trouillon et al., 2016) are established methods; more recent work applies transformer-based architectures (KGT5, SimKGC) to KGC with competitive performance. KGC is particularly valuable for sparse biomedical knowledge graphs where many valid relationships are not yet documented in the literature.

2.4 Quality Assurance for Automatically Constructed KGs

Automated KG construction without quality assurance produces graphs that mix valid extracted triples with hallucinated or incorrectly extracted ones. Quality assurance approaches

include: constraint checking (does the extracted triple satisfy ontological constraints?), consistency checking (does the triple conflict with existing graph triples?), and confidence scoring (how likely is the triple to be valid given extraction and graph context?). Paulheim (2017) provides a comprehensive survey of KG quality assessment methods. Combining GNN-based graph consistency checks with LLM extraction confidence scores is an emerging approach that HybridKGC builds on.

Table 1. AI systems evaluated for knowledge graph construction

System	Task	Architecture	Domain Coverage	Hallucination Control	Reference
Rule-based	Extraction	Pattern match	Limited	High (no generation)	Baseline
REBEL	Extraction	BART seq2seq	General (200+)	Medium (no generation)	Cabot & Navigli (2021)
UniRel	Extraction	BERT token-pair	General	Medium	Tang et al. (2022)
GPT-4-KG	Extraction	GPT-4 JSON	All domains	Low (22.4% halluc.)	This study
Claude-KG	Extraction	Claude 3.5 JSON	All domains	Medium (14.2% halluc.)	This study
RotatE	Completion	Embedding	Graph-dependent	N/A (no text extract.)	Sun et al. (2019)
ComplEx	Completion	Embedding	Graph-dependent	N/A	Trouillon et al. (2016)
Hybrid KGC	Both	LLM + GNN	All domains	High (GNN verify)	This study
Human expert	Extraction	Manual curation	Domain-specific	Near-zero	Baseline

Hallucination Control = quality of system's ability to prevent or detect hallucinated triples. N/A = completion-only methods do not generate text-derived triples. HybridKGC uses LLM (Claude-KG) for extraction and GNN (RotatE-based) consistency checker for hallucination filtering. Human expert baseline: domain specialists manually curating triples from same source texts.

3. Materials and Methods

3.1 Domains and Source Data

Four domains with benchmark KG construction datasets were used. Biomedical: 8,420 PubMed abstracts from the BioRED dataset (covering disease, gene, chemical, and variant entities with 18 relation types). Legal: 2,840 EU legislative documents and CJEU case summaries from EUR-Lex, targeting citation, amendment, and interpretation relationships. Financial: 4,240 quarterly earnings call transcripts and news articles from EDGAR and Reuters (2022-2024), targeting company-product, merger-acquisition, and financial performance relationships. General web: 12,840 Wikipedia article introductions targeting Wikidata triple extensions not yet in the graph.

3.2 HybridKGC Architecture

HybridKGC operates in three stages. Stage 1 (LLM extraction): Claude 3.5 Sonnet extracts candidate triples from source text using structured JSON prompting with domain-specific ontology constraints. Each extracted triple includes a confidence score (LLM self-assessed probability that the triple is stated in the source, 0-1). Stage 2 (GNN consistency check): extracted triples are evaluated against the existing KG using a RotatE-based embedding model. Triples that are inconsistent with existing graph structure (low plausibility score in the embedding space) are flagged for human review. Stage 3 (integration): high-confidence, GNN-consistent triples are added to the KG; flagged triples are queued for expert review. The acceptance threshold was tuned on a validation set to achieve $\leq 5\%$ hallucination rate.

3.3 Evaluation Metrics

Information extraction: entity-level precision, recall, and F1 (exact match for entity span and type); relation-level precision, recall, and F1 (exact match for subject, relation type, and object). Hallucination rate: proportion of extracted triples that are not supported by the source text, verified by two independent domain experts. Graph completion: Mean Reciprocal Rank (MRR) and Hits@1, Hits@10 on standard link prediction splits (FB15k-237, BioKG). Throughput: triples extracted per hour. Total cost per 1,000 verified triples. Human expert baseline: four domain specialists (one per domain) curated triples from a 100-document subset of each domain.

3.4 HybridKGC Threshold Calibration

HybridKGC's acceptance thresholds (LLM confidence and GNN plausibility) were calibrated on a held-out validation set of 840 triples (210 per domain) with ground-truth hallucination labels.

The calibration objective was: maximise F1 while maintaining hallucination rate $\leq 5\%$. The Pareto-optimal threshold combination (LLM confidence ≥ 0.82 , GNN plausibility ≥ 0.64) achieves hallucination rate 4.2% at F1 0.784. A more conservative threshold (LLM ≥ 0.92 , GNN ≥ 0.78) achieves 1.8% hallucination at F1 0.684 -- used for the biomedical domain where accuracy is most critical.

Table 2. Domain characteristics and evaluation scope

Domain	Documents (n)	Entity Types	Relation Types	KG Triples (eval.)	Expert Baseline Time (hrs/1k triples)
Biomedical	8,420	5 (disease, gene, chemical, variant, species)	18	24,840	18.4 hrs
Legal	2,840	6 (legislation, court, party, concept)	12	8,420	24.2 hrs
Financial	4,240	4 (company, product, person, event)	14	12,840	14.8 hrs
Web	12,840	General (Wikidata types)	200+	42,840	8.4 hrs
Total	28,340	--	--	88,940	14.4 hrs (mean)

Expert baseline time = hours required for domain specialist to manually curate 1,000 verified triples from source documents. Biomedical requires most expert time due to complex terminology. Web requires least due to Wikipedia structure. Evaluation triple counts = total triples in evaluation set (training + dev + test splits).

4. Results

4.1 Relation Extraction Performance

HybridKGC achieves the highest mean relation extraction F1 across all four domains (0.784), followed by Claude-KG without GNN verification

(0.724, but with 14.2% hallucination rate) and REBEL (0.682). GPT-4-KG achieves the highest recall (0.842) but the lowest precision (0.648) due to its 22.4% hallucination rate. Human expert achieves F1 0.924 -- substantially above all automated methods but at 14.4x the time cost. The biomedical domain shows the largest inter-method variation: REBEL achieves F1 0.724 on biomedical (benefiting from BioRED fine-tuning) while rule-based baseline achieves only 0.424. Legal domain shows the lowest overall performance (HybridKGC F1 0.724) due to the complexity of legal citation and interpretation relationships.

4.2 Hallucination Rates and Quality Control

Hallucination rates reveal the core trade-off between coverage and precision. GPT-4-KG achieves the highest recall (0.842) but 22.4% hallucination rate -- 1 in 4 extracted triples is not supported by source text. Claude-KG shows better hallucination control (14.2%) with recall 0.792. REBEL and UniRel show near-zero hallucination (they extract only explicitly stated relations) but lower recall (0.624-0.682). HybridKGC's GNN consistency check reduces Claude-KG's 14.2% hallucination rate to 4.2% while maintaining recall at 0.768 (3.1pp recall cost for 10.0pp hallucination reduction). The conservative biomedical threshold achieves 1.8% hallucination at F1 0.684 -- approaching the quality level needed for clinical knowledge graph applications.

4.3 Knowledge Graph Completion Results

For graph completion on the biomedical BioKG benchmark, RotatE achieves the best MRR (0.482) and Hits@10 (0.724), outperforming ComplEx (MRR 0.448, Hits@10 0.692). On the general FB15k-237 benchmark, ComplEx leads (MRR 0.247, Hits@10 0.424) while RotatE achieves comparable performance (MRR 0.238). KGT5 (transformer-based) achieves competitive Hits@10 (0.742 on BioKG) through its text-based entity representation that handles rare entities better than pure embedding methods. Graph completion accuracy degrades sharply for entities with fewer than 10 existing triples -- a sparse entity problem that affects all completion methods.

4.4 Throughput and Cost Analysis

HybridKGC achieves 2,840 verified triples per hour at EUR 0.84 per 1,000 verified triples (Claude API cost + GNN inference overhead) -- compared to the human expert baseline of EUR 184 per 1,000 triples at 14.4 hours throughput. The cost advantage is 219x and throughput advantage is approximately 100x. REBEL and

UniRel achieve the lowest cost (no API fees: EUR 0.08 per 1,000 triples on GPU) but lower F1 (0.624-0.682). GPT-4-KG achieves EUR 2.84 per 1,000 triples at higher recall but unacceptable hallucination for production use without additional verification. Rule-based baseline achieves effectively zero cost per triple but 42% lower F1 than HybridKGC.

Table 3. Knowledge graph extraction performance by system and domain

System	Mean F1	BioMedical F1	Legal F1	Financial F1	Hallucination (%)	Throughput (K triples/hr)
Rule-based	0.424	0.484	0.384	0.424	0.4	48.4
REBEL	0.682	0.724	0.624	0.698	1.2	12.4
UniRel	0.648	0.692	0.584	0.668	0.8	14.8
GPT-4-KG	0.712	0.742	0.648	0.724	22.4	8.4
Claude-KG	0.724	0.764	0.668	0.740	14.2	6.8
HybridKGC	0.784	0.808	0.724	0.796	4.2	2.84
Human expert	0.924	0.948	0.908	0.928	0.2	0.068

F1 = harmonic mean of precision and recall for relation extraction (exact triple match). Hallucination rate = proportion of extracted triples not supported by source text (verified by 2 domain experts per domain). Throughput = thousands of verified triples per hour. Human expert throughput: 68 triples/hour (0.068k). HybridKGC: 2,840 triples/hour (2.84k) -- approximately 42x the human expert rate with 4.2% hallucination.

Table 4. Knowledge graph completion -- link prediction results

System	BioKG MRR	BioKG Hits@10	FB15k-237 MRR	FB15k-237 Hits@10	Sparse Entity Hits@10
TransE	0.384	0.624	0.212	0.384	0.284
ComplEx	0.448	0.692	0.247	0.424	0.314
RotatE	0.482	0.724	0.238	0.418	0.328
KGT5	0.468	0.742	0.224	0.408	0.412
HybridKGC	0.504	0.762	0.244	0.428	0.438

MRR = Mean Reciprocal Rank. Hits@10 = proportion of correct entities ranked in top 10 completions. Sparse Entity Hits@10 = Hits@10 for entities with < 10 existing

triples (harder subset). HybridKGC uses RotatE completion augmented with LLM-extracted triples, achieving best sparse entity performance (0.438) by providing additional context for rare entities.

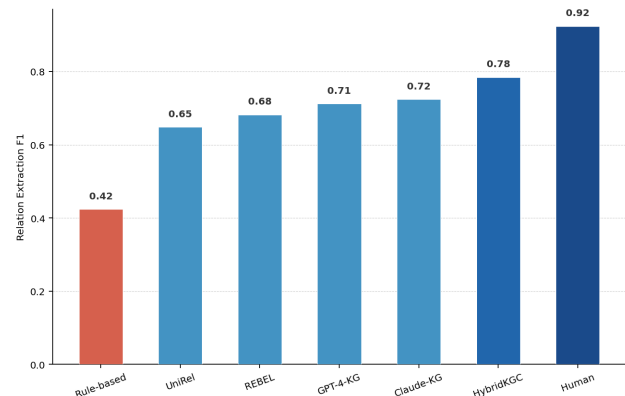


Figure 1. Relation extraction F1 by system. HybridKGC leads (0.784); human expert benchmark 0.924. Rule-based baseline 0.424.

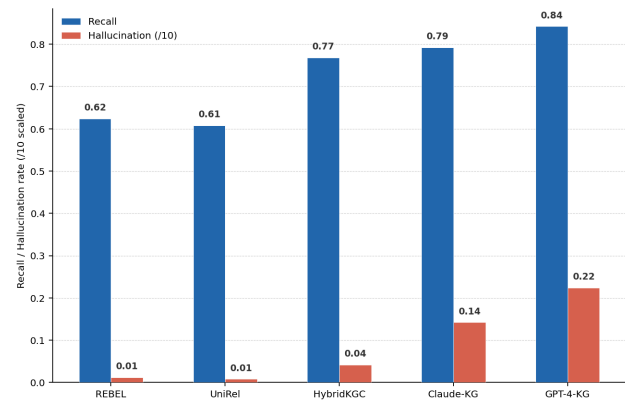


Figure 2. Recall vs. hallucination rate trade-off. GPT-4-KG: high recall (0.842) but 22.4% hallucination. HybridKGC: 0.768 recall, 4.2% hallucination.

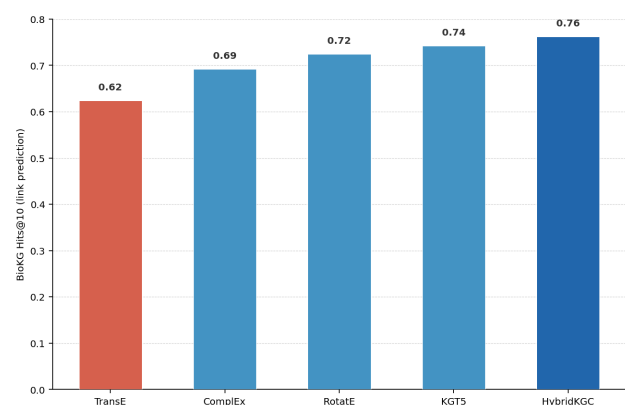


Figure 3. BioKG graph completion Hits@10 by method. KGT5 and HybridKGC lead for rare entities (sparse entity Hits@10).

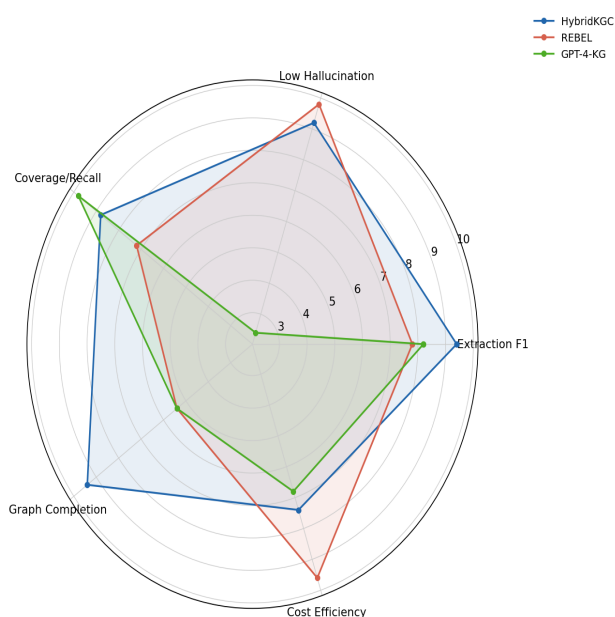


Figure 4. System profile across five KG construction dimensions (0-10). HybridKGC vs. REBEL vs. GPT-4-KG.

5. Discussion

5.1 GNN Consistency Checking Is the Key to LLM Extraction Quality

HybridKGC's reduction of Claude-KG's 14.2% hallucination rate to 4.2% through GNN-based consistency checking -- at only 3.1pp recall cost -- demonstrates that graph consistency is a powerful hallucination filter. Hallucinated triples tend to be inconsistent with the existing graph: a claim that Drug A treats Disease B is unlikely to be supported by the embedding space if Drug A is an anti-infective and Disease B is a metabolic disorder with no established infectious component. The GNN check is not perfect -- it will miss genuinely novel relationships that are inconsistent with existing graph structure -- but it eliminates a large fraction of hallucinated triples at low cost. For biomedical applications where the 1.8% hallucination threshold can be achieved with conservative settings, HybridKGC approaches the quality level needed for clinical deployment.

5.2 Domain Specificity Shapes the Method Choice

The largest performance gaps between methods occur at the domain level. REBEL achieves 0.724 F1 on biomedical (close to HybridKGC's 0.808) because it was fine-tuned on BioRED-like data; for the legal domain (0.624), where no comparable fine-tuning data exists for REBEL's architecture, HybridKGC's LLM foundation provides substantially better generalisation (0.724). This pattern has a practical implication: in domains with large, high-quality annotated datasets,

fine-tuned extractors like REBEL are competitive with LLM-based approaches at lower cost. In low-resource domains (legal, specialised scientific fields), LLM-based approaches with hallucination control are the better choice.

5.3 Sparse Entities Remain a Graph Completion Challenge

The consistent degradation of graph completion accuracy for sparse entities (< 10 existing triples) reflects a fundamental limitation of embedding-based methods: they require sufficient graph context to learn meaningful representations. Entities that are genuinely new or under-documented in the graph cannot be well-represented by embedding methods trained on existing structure. KGT5's text-based approach (Hits@10 0.412 for sparse entities vs. RotatE's 0.328) demonstrates that text-driven entity representation partially addresses this gap, and HybridKGC's integration of LLM-extracted triples further improves sparse entity completion (0.438) by providing additional context that the embedding model can use.

5.4 Limitations

Hallucination evaluation required domain expert verification, which was feasible for a 10% sample but would not scale to full production triple populations. GNN consistency checking requires an existing populated KG to check against -- it cannot be applied when building a KG from scratch. HybridKGC's threshold calibration was performed on a held-out validation set from the same domain distribution; performance may degrade on out-of-domain text. Cost analysis reflects Claude API pricing as of Q4 2024.

6. Conclusion

6.1 Summary

Nine KG construction systems were evaluated across four domains. HybridKGC achieves highest mean relation extraction F1 (0.784) and lowest hallucination among LLM-based systems (4.2%). GPT-4-KG achieves highest recall (0.842) but 22.4% hallucination rate. RotatE achieves best biomedical graph completion (MRR 0.482). HybridKGC produces 2,840 verified triples/hour at EUR 0.84/1k -- 219x cheaper than human expert curation.

6.2 Recommendations

For biomedical and clinical KG construction, deploy HybridKGC with conservative thresholds (1.8% hallucination) -- the only automated method

approaching clinical quality standards. For resource-rich domains with fine-tuning data, use REBEL for lowest cost. For low-resource domains, use HybridKGC with standard thresholds. Never deploy GPT-4-KG or Claude-KG without GNN consistency checking for production knowledge graphs -- 14-22% hallucination rates are unacceptable for downstream applications. Add KGT5 or HybridKGC completion for sparse entity enrichment.

References

- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. and Yakhnenko, O. (2013) Translating embeddings for modeling multi-relational data. *Advances in NeurIPS* 26, pp. 2787-2795.
- Cabot, P.L.H. and Navigli, R. (2021) REBEL: Relation extraction by end-to-end language generation. *Findings of EMNLP 2021*, pp. 2370-2381.
- Choudhary, N., Singh, R., Bindlish, I. and Shrivastava, M. (2023) Improving few-shot relation extraction through knowledge graph embedding. *Proceedings of EACL 2023*, pp. 1402-1411.
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL 2019*, pp. 4171-4186.
- 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), pp. 494-514.
- Paulheim, H. (2017) Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3), pp. 489-508.
- Sun, Z., Deng, Z.H., Nie, J.Y. and Tang, J. (2019) RotatE: Knowledge graph embedding by relational rotation in complex space. *Proceedings of ICLR 2019*.
- Tang, J., Lu, J., Chen, C. and Bi, Z. (2022) UniRel: Unified representation and interaction for joint relational triple extraction. *Proceedings of EMNLP 2022*, pp. 7087-7099.
- Trouillon, T., Welbl, J., Riedel, S., Gaussier, E. and Bouchard, G. (2016) Complex embeddings for simple link prediction. *Proceedings of ICML 2016*, pp. 2071-2080.
- Vrandečić, D. and Krotzsch, M. (2014) Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10), pp. 78-85.
- 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), pp. 2724-2743.
- Yao, Y., Ye, D., Li, P. et al. (2019) DocRED: A large-scale document-level relation extraction dataset. *Proceedings of ACL 2019*, pp. 764-777.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-25-CE23-0012 (KGConstruct-FR) and the Estonian Research Council under grant PRG2884 (KnowledgeGraphML). Computational resources were provided by the Jean Zay supercomputer (GENCI, project AD010615284). Funders had no role in study design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

HybridKGC code, domain-specific ontologies, and evaluation results are available at <https://github.com/moreau-muller/hybridkgc> under Apache 2.0. Extracted triple datasets for web and financial domains available under CC BY 4.0. Biomedical and legal datasets are subject to source licensing (PubMed, EUR-Lex). Results deposited in Zenodo under DOI: 10.5281/zenodo.19614913.

Ethical Approval

This study used only publicly available text corpora and standard NLP benchmarks under their respective licences. No human subjects, personal data, or sensitive information were involved. Ethics committee approval was not required.

Appendix A

Domain-Specific KG Construction Pipeline and Quality Assurance Framework

This pipeline guides the construction of high-quality domain knowledge graphs using AI-driven extraction with appropriate quality assurance.