JAIST Repository
https://dspace.jaist.ac.jp/

Title	教師なし知識抽出と意味検索に基づくエビデンスに基づいた推論に向けて
Author(s)	VO, THIEN TRUNG
Citation	
Issue Date	2026-03
Type	Thesis or Dissertation
Text version	ETD
URL	https://hdl.handle.net/10119/20590
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 博士

JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

氏　名	Vo Thien Trung
学位の種類	博士（情報科学）
学位記番号	博情第 577 号
学位授与年月日	令和 8 年 3 月 25 日
論文題目	TOWARDS EVIDENCE-BASED REASONING VIA UNSUPERVISED KNOWLEDGE EXTRACTION AND SEMANTIC RETRIEVAL
論文審査委員	Nguyen Le Minh　JAIST　Professor
	SHIRAI Kiyoaki　JAIST　Professor
	Satoshi Tojo　Asia University　Professor
	Phan Xuan Hieu　UET, VNU-HN　Associate Professor
	Naoya Inoue　JAIST　Associate Professor

論文の内容の要旨

Large language models (LLMs) have demonstrated strong capabilities across a wide range of reasoning tasks, especially when supported by prompting or external knowledge. Among the most common ways to incorporate such knowledge is retrieval-augmented generation (RAG), which grounds model outputs in retrieved evidence.

In scientific and biomedical settings, however, users need more than a correct answer: they must be able to (i) identify which facts the system relies on, (ii) assess whether the evidence genuinely supports the conclusion, and (iii) verify the reasoning through explicit, fine-grained references to the source material. Accordingly, this thesis argues that evidence-based reasoning with LLMs requires inspectable evidence selection and verifiable reasoning traces, not only fluent answers—especially in biomedical and scientific settings where decisions must be audited. In this thesis, evidence-based reasoning means that each prediction is grounded in explicitly selected evidence and accompanied by a reasoning trace that can be checked at the appropriate granularity (sentence-level in text, path-level in graphs, and cell-level in tables).

In practice, many LLMs paired with RAG pipelines still fall short for three structural reasons. First, much of an LLM's knowledge remains implicit in its parameters, making it difficult to inspect, reuse, or maintain as explicit facts. Second, retrieval is frequently driven by lexical matching and vector similarity, which can capture topical relatedness but still miss the precise evidence required for multi-hop inference. Third, reasoning traces are often not audit-friendly—especially for tables, where verification requires localized, cell-level citations rather than generic explanations.

A common solution is to train task-specific retrievers and reasoning models. However, in biomedical and scientific applications, labeled data is often limited and expensive, and the evidence format changes across tasks, including text, graphs, and tables. As a result, supervised systems can be hard to transfer and maintain: they usually need new labels or fine-tuning when the domain or task changes. This motivates an unsupervised and explainable framework that can work with minimal annotation. We therefore organize the thesis around an unsupervised pipeline that (i) makes knowledge explicit, (ii) performs semantic retrieval, meaning unsupervised evidence selection across data formats (sentences in text, columns/cells in tables, and salient paths in graphs), and (iii) produces reasoning traces that can be directly checked. These goals align with three pillars of the dissertation: unsupervised knowledge extraction makes latent model knowledge explicit, semantic retrieval selects the right evidence precisely, and evidence-based reasoning exposes verifiable traces for auditing.

This thesis develops an unsupervised and explainable evidence-based reasoning framework that works across tabular and textual settings with minimal annotation. We first study unsupervised semantic retrieval and verifiable reasoning without external knowledge, focusing on tabular evidence. For tabular reasoning and fact verification, UCRET selects claim-relevant columns via spherical k-means and produces label-conditioned counterfactual reasoning with concise cell-level citations. To strengthen evidence selection in tables, UCRET-JS extends this idea with distribution-aware retrieval, using Jensen–Shannon divergence to compare contextual token distributions rather than relying on a single embedding.

While reasoning only from the provided input improves transparency and reduces supervision needs, it also exposes a key limitation: the evidence available in a table or a passage can be implicit, incomplete, or insufficient for multi-hop inference. To address this limitation, the thesis moves to unsupervised semantic retrieval with external knowledge, where knowledge graphs (KGs) provide additional context, enable multi-hop reasoning, and improve answer reliability. K-Bloom converts latent knowledge in pretrained language models into high-precision KG tuples using an Optimal Transport formulation, making model knowledge explicit and reusable. USCRaKe performs unsupervised text retrieval with Optimal Transport and Jensen–Shannon divergence by comparing distributions of contextual tokens rather than single-vector similarity, improving evidence selection for compositional queries. Building on the biomedical KG produced by K-Bloom, UGAT-MedQA applies unsupervised graph attention to identify salient multi-hop paths, which are verbalized into step-by-step evidence-linked reasoning for medical question answering.

Across general, biomedical, and scientific benchmarks, the results show that an unsupervised and explainable pipeline can deliver higher quality evidence selection and more verifiable reasoning without task-specific labels or fine-tuning while remaining adaptable to different evidence formats (tables, text, and graphs). Future work will focus on reducing hallucinations, making retrieval more intent- and verification-oriented, maintaining domain knowledge graphs over time, and extending unsupervised table reasoning to noisier real-world and clinical settings.

Keywords: Large Language Models, Knowledge Graphs, Optimal Transport, Unsupervised Learning, Question Answering, Table-based Fact Verification

論文審査の結果の要旨

This dissertation addresses several fundamental limitations of current large language models, namely the implicit and opaque nature of their knowledge, the limitations of vector-based retrieval for multi-hop reasoning, and the lack of transparency in explanation, particularly for table-based reasoning. The problem framing is well-articulated and targets core weaknesses that hinder trustworthy and reusable LLM-based systems. The thesis makes a strong and coherent contribution by proposing a unified, unsupervised, and explainable framework that operates consistently across text, knowledge graphs, and tables. A key strength of the work is its principled use of distribution-aware retrieval, primarily through Optimal Transport–based formulations, to move beyond single-vector similarity. The individual components—**K-Bloom**, **USCRaKe**, **UGAT-MedQA**, and **UCRET**—are technically sound and thoughtfully designed, each addressing a distinct reasoning modality while fitting naturally into a single end-to-end pipeline. In particular, the conversion of implicit model knowledge into explicit, high-precision knowledge graph tuples, the use of token-level distribution comparisons for unsupervised retrieval, and the emphasis on human-verifiable explanations for both graph- and table-based reasoning represent meaningful advances over prior work. The dissertation is methodologically rigorous, conceptually unified, and empirically validated across general, biomedical, and scientific domains. The work demonstrates a high level of originality and maturity, and the proposed framework and individual components are of a quality suitable for publication at high quality journals. The thesis makes a substantial contribution toward explainable, trustworthy, and supervision-efficient reasoning systems built on large language models. In conclusion, this is an excellent dissertation and we approve awarding a doctoral degree to Vo Thien Trung.