Personal Qualifying Question

Joshua Cook

Georgia Institute of Technology

CS 6460

Evaluation Framework for Knowledge Extraction and Ontology Development

Prompt: *What does the literature say about evaluation frameworks for assessing the accuracy and reliability of various knowledge extraction methods, particularly those integrating large language models with canonical NLP techniques? How can you assess the quality of the ontologies that you're developing with these various tools?*

Background

I am the cofounder of an educational technology startup, and as such I am focused on developing automated knowledge extraction and ontological mapping solutions as core technology to our product. Our current research and development efforts center on leveraging large language models and advanced natural language processing techniques to efficiently extract, organize, and utilize knowledge from unstructured textual data, with the goal of augmenting the creation of knowledge bases for internal training and information management. In developing this tech, we are focused on three central ideas.

Ontology is a formal, explicit specification of a shared conceptualization, as defined by Guarino and Giaretta (1995). It provides a structured representation of knowledge within a specific domain. **Ontology mapping** is the process of finding correspondences between concepts, properties, and relationships in different ontologies, as described by Matentzoglu et al. (2022). **Knowledge extraction** is an approach to information extraction that leverages recent advances in large language models to populate complex knowledge schemas from unstructured text, as explained by Caufield et al. (2024) in the context of their SPIRES method. It involves automated processes for identifying and structuring relevant information from raw data sources.

Recent advancements in programmatic ontology development have significantly enhanced the efficiency and standardization of knowledge representation. Tools

like LinkML, which simplifies the production of FAIR ontology-ready data (Moxon et al., 2021), and the Ontology Development Kit (ODK), which provides standardized and customizable workflows (Matentzoglu et al., 2022), have streamlined the process of creating and maintaining ontologies. The integration of large language models (LLMs) into natural language processing techniques has revolutionized knowledge extraction and ontology development. Researchers have demonstrated the potential of using models like GPT-3 for in-context learning in relation extraction tasks (Wan et al., 2023) and have developed frameworks for leveraging ChatGPT in zero-shot information extraction by decomposing complex tasks into multi-turn question-answering problems (Wei et al., 2023).

The current state of knowledge extraction methods showcases innovative approaches that harness the power of LLMs. For instance, the SPIRES method (Caufield et al., 2024) employs LLMs for zero-shot learning to automatically populate knowledge bases using predefined schemas. Another approach, described by Wang et al. (2023), utilizes a three-step process involving prompt construction, text generation, and entity labeling to extract information effectively. The integration of LLMs with traditional NLP techniques has led to paradigms like LLMs4OL (what a pun!), which automates ontology learning tasks such as term typing and relation extraction using zero-shot prompting (Babaei Giglou et al., 2023). Researchers have explored combining traditional curation methods with AI-driven approaches to create and update ontologies, as demonstrated in the development of an AI ontology (Joachimiak et al., 2024).

It is difficult to convey just how rapidly the field of knowledge extraction is progressing. We are at a very exciting time as LLMs are capable of doing the sort of ontology curation that only a few years ago required deep subject matter expertise. While it is easy to get carried away with excitement, it is also critical that we develop evaluation methodologies to assess the accuracy and reliability of these techniques. Jackson et al. (2021) describe a demand for robust frameworks to evaluate the effectiveness of various knowledge extraction methods, particularly those that integrate large language models with canonical NLP approaches. The importance of adhering to FAIR (Findable, Accessible, Interoperable, and Reusable) principles in these evaluation processes is emphasized by Jonquet & Grau (2024). This is toward ensuring that the resulting ontologies and knowledge bases are not only accurate but also widely usable and interoperable across different systems and domains.

Literature Review

Quantitative evaluation using precision, recall, and F1 scores forms the backbone of assessing knowledge extraction and ontology mapping systems. Precision measures the proportion of correctly identified elements among all extracted elements, while recall indicates the proportion of correctly identified elements among all relevant elements in the dataset. The F1 score, the harmonic mean of precision and recall, provides a balanced measure of the system's performance. These metrics are widely applied to various knowledge extraction tasks, including entity recognition, relation extraction, and ontology alignment. However, as Matentzoglu et al. (2023) point out, these traditional metrics have limitations. They treat all mappings equally without considering their importance or difficulty, fail to account for the structure and semantics of ontologies, and cannot distinguish between slightly wrong and completely wrong mappings. Wang et al. (2023) further emphasize that these metrics fall short in capturing the semantic richness of mappings or the complexity of specific matching tasks.

Human expert involvement plays a crucial role in the evaluation and refinement of knowledge extraction and ontology development systems. Matthews et al. (2023) emphasize the importance of expert interpretation and contextualization throughout the research process, noting that "language and documentation is checked and validated at each stage by domain experts." These experts are instrumental in addressing data quality issues, particularly in real-world scenarios where data is often messy and requires cleaning and transformation. Bikeyev (2023) highlights the critical role of human experts in validating the output of Large Language Models (LLMs), ensuring that the extracted knowledge and ontological structures are accurate and meaningful within the specific domain context. Matieu (2023) further elaborates on this point, stating that human experts are essential for disambiguating unclear concepts and ensuring that the ontology accurately represents the intended meaning within the domain. The involvement of human experts also extends to the peer review process upon completion of the ontology, as noted by Matthews et al. (2023).

Domain adaptability testing is another part of evaluating knowledge extraction and ontology development systems. This process involves selecting diverse domains for testing, analyzing cross-domain performance, and identifying domain-specific challenges. Matentzoglu et al. (2022) describe a practical approach to assessing domain adaptability, using methods such as gathering known ontologies using SSSOM (Simple Standard for Sharing Ontological Mappings) and performing GitHub searches to identify additional repositories using relevant tools like the "ontology

development kit" or "ontology starter kit." This approach helps in evaluating the system's ability to handle diverse ontological structures and terminologies. Joachimiak et al. (2024) provide an example of cross-domain performance analysis in their work on the Artificial Intelligence Ontology (AIO). They conducted a Natural Language Processing (NLP) evaluation to assess the coverage and applicability of AIO within the context of practical AI research, involving lexical matching of terms from a dataset of 2,194 research publications against the term labels and synonyms defined in AIO.

Evaluation of downstream task performance is another aspect of evaluating knowledge extraction and ontology mapping systems. This involves selecting relevant downstream tasks, defining performance metrics for each task, and analyzing the real-world applicability of the extracted knowledge and ontological structures. Caufield et al. (2024), in their work on the SPIRES method, demonstrate this approach by conducting extensive experiments on Relation Extraction (RE), Named Entity Recognition (NER), and Event Extraction (EE) tasks across six datasets in both English and Chinese. They employ standard micro F1 measures and adopt two evaluation metrics: border evaluation (Rel) and strict evaluation (Rel+), providing a comprehensive view of the system's performance. Similarly, Wei et al. (2024), in their ChatIE framework, showcase the importance of cross-lingual evaluation and comparison with full-shot models. They report impressive performance on several datasets, even surpassing some full-shot models on datasets like NYT11-HRL.

Proposed Evaluation Framework

Considering this, I have developed what I call the "In-Application Evaluation Framework for Knowledge Extraction and Ontology Mapping Models" (IEF, Appendix I). The IEF combines quantitative metrics with qualitative insights, especially human-in-the-loop user input, and measurement of this input. It integrates these as part of the core application experience. The goal is the evaluation of our automated knowledge construction, but to do so in a way that surfaces the efficacy of the knowledge construction to users as they work with our product to develop internal ontologies and knowledge bases.

This is especially critical for our application which relies heavily on large language models for most aspects of knowledge extraction. We are looking to use tools like the LLM-integrated NLP techniques (Wan et al., 2023; Babaei Giglou et al., 2023) and automated ontology development tools like LinkML and ODK (Moxon et al., 2021; Matentzoglu et al., 2022). If we are going to create robust, self-updating knowledge

bases. Furthermore, if we are going to take this out of the hands of our users or at a minimum significantly augment their workflows, then we will need to engender trust. The IEF is a step in that direction.

The decisions made in creating the IEF are grounded in the concepts and tools identified in the literature review. The integration of quantitative and qualitative evaluation methods addresses the limitations of traditional metrics (Matentzoglu et al., 2023) while incorporating crucial expert involvement (Matthews et al., 2023; Bikeyev, 2023). The emphasis on user-centric evaluation and domain adaptability testing aligns with the importance of expert interpretation and cross-domain performance analysis (Joachimiak et al., 2024). The framework's focus on downstream task performance evaluation (Caufield et al., 2024; Wei et al., 2024) ensures real-world applicability, while its alignment with FAIR principles (Jonquet & Grau, 2024) promotes the creation of findable, accessible, interoperable, and reusable ontologies and knowledge bases.

Appendices

Appendix I: In-Application Evaluation Framework for Knowledge Extraction and Ontology Mapping Models

1. Basic Model Evaluation
    a. Automated Performance Metrics
        i. Implement real-time tracking of precision, recall, and F1 scores for
            1. entity extraction
            2. relation extraction
            3. mapping accuracy
        ii. Create dashboards for quick performance overview
    b. Benchmarking Against Baseline
        i. Maintain a set of benchmark datasets
        ii. Regularly compare new model versions and prompting strategies against previous versions and simple baselines
2. User-Facing Evaluation Products
    a. In Product Human-In-The-Loop Tooling
        i. Create visualizations of extracted concepts and ontology mappings
        ii. Allow users to interactively verify, modify, or reject extractions and mappings

        iii. Collect and analyze user interactions with these visualizations
    b. Knowledge Base Expansion Dashboards
        i. Track the growth and evolution of the knowledge base over time
        ii. Evaluate the impact of knowledge base expansion on overall system performance
    c. In-app Feedback Mechanisms
        i. Implement simple feedback options (e.g., thumbs up/down) on extracted concepts and mappings
        ii. Create an easy way for users to report incorrect extractions or mappings
        iii. Track simple join, creation, and deletion of entities and relations
    d. User-Facing Extraction and Mapping Dashboards
        i. Display performance metrics for knowledge extraction and ontology mapping models vs human-in-the-loop
    e. Automated Ontology Enrichment
        i. Implement mechanisms for suggesting additions or modifications to ontologies based on extracted knowledge
        ii. Evaluate the quality and usefulness of these automated suggestions
    f. User-Driven Knowledge Refinement
        i. Implement a process for users to suggest refinements to the knowledge base and ontology mappings
        ii. Conduct follow-up interviews with users who suggest significant changes to understand their reasoning and use cases
        iii. Develop a system to evaluate and incorporate user-suggested refinements into prompting strategies
3. User-Centric Evaluation and Feedback
    a. Comprehensive User Interview Program
        i. Conduct regular, structured interviews with diverse user groups (e.g., power users, new users, different roles)
        ii. Design interview protocols to gather insights on:
            1. Accuracy and relevance of extracted knowledge
            2. Usefulness and intuitiveness of ontology mappings
            3. Workflow integration and efficiency gains
            4. Feature requests and pain points
        iii. Use a mix of open-ended questions and specific task-based inquiries
        iv. Implement a system to categorize and quantify qualitative feedback from interviews

        v. Conduct periodic longitudinal interviews with select users to track perceptions over time
        vi. Implement user journey mapping to understand how the tool fits into broader workflows
        vii. Use card sorting exercises to evaluate the intuitiveness of ontology structures
    b. Feedback Synthesis and Action Planning
        i. Regularly synthesize insights from in-app feedback, surveys, and user interviews
        ii. Conduct cross-functional meetings (including developers, product managers, and customer success teams) to review synthesized feedback
        iii. Develop action plans based on user feedback, balancing quick wins and long-term improvements
        iv. Create a feedback loop to inform users about how their input has influenced product development
        v. Implement a prioritization matrix to weigh user requests against technical feasibility and business impact
4. Company-Facing Evaluation Products
    a. Company-Facing Extraction and Mapping Dashboards
        i. Display performance metrics for knowledge extraction and ontology mapping models vs human-in-the-loop
        ii. Track model performance across different customer domains
        iii. Identify domains where the model underperforms and prioritize improvements
        iv. Visualize trends in model performance over time, correlated with major updates or changes

References

Babaei Giglou, H., D'Souza, J., & Auer, S. (2023). LLMs4OL: Large language models for ontology learning. In International Semantic Web Conference (pp. 408-427). Cham: Springer Nature Switzerland.

Bikeyev, A. (2023). Synthetic ontologies: A hypothesis. Available at SSRN 4373537.

Caufield, J. H., Hegde, H., Emonet, V., Harris, N. L., Joachimiak, M. P., Matentzoglu, N., ... & Mungall, C. J. (2024). Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. Bioinformatics, 40(3), btae104.

Guarino, N. and Giaretta, P. (1995) Ontologies and Knowledge Bases. In: Towards Very Large Knowledge Bases, IOS Press, Amsterdam, 1-2.

Jackson, R., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., ... & Peters, B. (2021). OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database, 2021, baab069.

Joachimiak, M. P., Miller, M. A., Caufield, J. H., Ly, R., Harris, N. L., Tritt, A., Mungall, C. J., & Bouchard, K. E. (2024). The Artificial Intelligence Ontology: LLM-assisted construction of AI concept hierarchies. Lawrence Berkeley National Laboratory.

Jonquet, C., & Grau, N. (2024). M4. 4-Review of Semantic Artefact Catalogues and guidelines for serving FAIR semantic artefacts in EOSC.

Matentzoglu, N., Balhoff, J. P., Bello, S. M., Bizon, C., Brush, M., Callahan, T. J., ... & Mungall, C. J. (2022). A Simple Standard for Sharing Ontological Mappings (SSSOM). Database, 2022, baac035. https://doi.org/10.1093/database/baac035

Matentzoglu, N., Caufield, J. H., Hegde, H. B., Reese, J. T., Moxon, S., Kim, H., ... & Mungall, C. J. (2023). Mappergpt: Large language models for linking and mapping entities. arXiv preprint arXiv:2310.03666.

Matentzoglu, N., Goutte-Gattat, D., Tan, S. Z. K., Balhoff, J. P., Carbon, S., Caron, A. R., ... & Osumi-Sutherland, D. (2022). Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. Database, 2022, baac087.

Matthews, J., Love, P. E. D., Porter, S., & Fang, W. (2023). Curating a domain ontology for rework in construction: challenges and learnings from practice. Production Planning & Control. https://doi.org/10.1080/09537287.2023.2223566

Moxon, S. A., Solbrig, H., Unni, D. R., Jiao, D., Bruskiewich, R. M., Balhoff, J. P., ... & Mungall, C. J. (2021). The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics. ICBO, 3073, 148-151.

Procko, T. T., Elvira, T., & Ochoa, O. (2023). GPT-4: A Stochastic Parrot or Ontological Craftsman? Discovering Implicit Knowledge Structures in Large Language Models. In 2023 Fifth International Conference on Transdisciplinary AI (TransAI) (pp. 147-154). IEEE.

Qiang, Z., Wang, W., & Taylor, K. (2025). Agent-OM: Leveraging LLM Agents for Ontology Matching. Proceedings of the VLDB Endowment, 18(1). https://doi.org/10.48550/arXiv.2312.00326

Silva, J., Revoredo, K., Baião, F. A., & Lima, C. (2023). Enhancing Ontology Matching: Lexically and Syntactically Standardizing Ontologies Through Customized Lexical Analyzers. Semantic Web, IOS Press.

Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., & Kurohashi, S. (2023). GPT-RE: In-context Learning for Relation Extraction using Large Language Models. Kyoto University, Japan; Zhejiang University, China.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. arXiv preprint arXiv:2301.13294. https://arxiv.org/abs/2301.13294

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., & Han, W. (2024). ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. Retrieved from https://arxiv.org/abs/2302.10205