Lexico-syntatic Patterns for Fine-tuning an LLM

Rakesh Kandula[1], Lee Hannah[1] and Cogan Shimizu[1,*]

[1] Wright State University, Dayton, USA

Abstract

Large Language Models (LLMs) are expensive to train. However, there are techniques that can adapt LLMs more quickly and efficiently, such as fine-tuning with domain specific data. This allows foundational models to be applied to more niche use-cases in a cost efficient manner. Knowledge graphs (KGs) are excellent sources of curated data, making them an ideal source of knowledge for fine-tuning. Further, lexico-syntactic patterns (LSPs) can play an important role in representing data captured in semantic relationships in KGs as natural language text. In this paper, we discuss the use of LSPs to represent knowledge graphs (KGs) in natural language for the purposes of fine tuning. We demonstrate in our question answering use-case that fine-tuning helps in this case, but does not exceed retrieval-augmented generation approaches. We posit with larger KGs and and additional LSPs, we can achieve parity. **Poster Submission.**

Keywords

Lexico Syntactic Patterns (LSPs), Fine-tuning, Large Language Models(LLMs)

1. Introduction

In this era of Large Language Models (LLMs) (e.g., ChatGPT [1], many are looking to integrate LLMs into their research, projects, or products. LLMs are good at many common tasks, especially related to the amount of relevant training material on the Web. However, in niche cases, LLMs can struggle with domain-specific pattern extraction or multi-hop answering where reasoning needs to be done based on several entities. This frequently results in LLMs hallucinating and providing incorrect information to users. Structuring data while preserving its semantic meaning is crucial especially in fields like medicine where accuracy is important [2]. Knowledge Graphs (KGs) with their ability to structure data without losing semantic meaning and their growing popularity offer a solution [3, 4]. Much work has already been done, showing that KGs can reduce hallucinations in LLMs, making them more reliable [5].

One of the main challenges is making LLMs understand and utilize our custom data or domain-specific data. Training an LLM from scratch is time-consuming and resource-intensive. Fine-tuning [6] and Retrieval Augmented Generation (RAG) [7] [8] are two common approaches. Our approach utilizes KGs to provide factually correct, domain-specific data for fine-tuning LLMs. Specifically, we utilize *Lexico-Syntactic Patterns* (LSPs) [9].

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA
*Corresponding author.
kandula.15@wright.edu (R. Kandula); lee.hannah@wright.edu (L. Hannah); cogan.shimizu@wright.edu (C. Shimizu)
0009-0006-1064-141X (R. Kandula); 0000-0003-3332-3220 (L. Hannah); 0000-0003-4283-8701 (C. Shimizu)

2. Related Work

To our knowledge, there is little work done so far on the intersection between LSPs, KGs, and LLMs. Yet, we know that there are some related work on the mechanisms for detecting and extracting knowledge and data from unstructured natural language into ontologies [9]. Specifically, they leverage ontology design patterns (ODPs) [10] to connect domain specific knowledge to domain specific language. We intend to leverage these connections to translate knowledge and data (in a KG) that has already been captured and *translate back* to natural language for the purposes of fine-tuning.

3. Current Work

Use-Case: First, we collected data from `data.ohio.gov`, which is a State-wide data repository, related to census, public health, hospital locations, as well as administrative regions (e.g., counties) from KnowWhereGraph (KWG) [11]. After basic cleaning and alignment to KWG entities, we materialized a KG locally, we call the OhioKG.
Motivating Example. LSPs, as we use them, are defined for node-edge-node constructions in the schema diagram (or, more specifically, for each distinct type-predicate-type triple in the KG). For example, the triple `Columbus -> locatedIn -> Ohio` is converted to natural language via the LSP *qqThe city <E1> is <r> in the state <E2>"*, which is then resolved as *"The city Columbus is locatedIn in the state Ohio."* LSPs enhance communication and we posit that they can help LLMs better understand and convey information, leading to more accurate information retrieval.

Briefly, we report some example LSPs that we utilized in the fine-tuning over the OhioKG.

Example 1: Node-Edge-Node Constructions

kl-ont:MarijuanaDispensary	kl-ont:hasBusinessName	kl-ont:Organization
kl-ont:Organization	kl-ont:hasName	xsd:string

Example 1: Instance Data (NMCD – Nectar Medical Cannabis Dispensary)

kl-res:MMD.0700164	kl-ont:hasBusinessName	kl-res:NMCD
kl-res:NMCD	kl-ont:hasName	"NMCD"xsd:string

Example 1: Lexico-syntactic Pattern *"The Marijuana Dispensary <E1> <r> <E2>, Business name <E2> <r> <E3>"*

Thus, we construct: *The Marijuana Dispensary 'MMD.0700164' hasBusinessName 'Nectar_Medical_Cannabis_Dispensary'. Business name 'Nectar_Medical_Cannabis_Dispensary' hasName 'Nectar Medical Cannabis Dispensary'*

Example 2: Node-Edge-Node Constructions

kl-ont:MarijuanaDispensary	kl-ont:hasBusinessName	kl-ont:Organization
kl-ont:Organization	kl-ont:hasName	xsd:string
kl-ont:MarijuanaDispensary	kl-ont:hasAddress	xsd:string

Example 2: Instance Data (NMCD – Nectar Medical Cannabis Dispensary with Address)

kl-res:MMD.0700164	kl-ont:hasBusinessName	kl-res:NMCD
kl-res:NMCD	kl-ont:hasName	”NMCD”xsd:string
kl-res:MMD.0700164	kl-ont:hasAddress	”21100 Saint Clair Ave”xsd:string

Example 2: Lexico-syntactic Pattern *”The Marijuana Dispensary <E3> is located at address <E4>”*

Thus, we construct: *The Marijuana Dispensary ‘Nectar Medical Cannabis Dispensary‘ is located at address ‘21100 Saint Clair Ave‘*, which is a more information dense natural language statement.

3.1. Preliminary Results

OhioKG & LSPs. OhioKG consists of different entities over 1M triples, but specifically for the marijuana dispensaries, we only have ≈ 2K triples, for which We have developed 12 Q&A LSPs.
Computational Environment. These preliminary experiments were conducted using Google Colab Pro[1] where we fine-tuned LLAMA2 7B [12]. At least 22GB of VRAM is required to perform this experiment, increasing proportionally with the size of KG and LLM. Fine tuning for only dispensary data took only 10 minutes including all LSPs.
Results. In our preliminary experiments using the LSPs to generate rich natural language from KG fragments, we have only seen marginal improvement. Initially, the LLM was not able to answer any questions (indeed, in some cases, and notably ChatGPT, it refused to answer due to US Federal government’s stance on marijuana). However, after fine-tuning, we were able to get answers in the correct format (i.e., mimicking the LSPs in responses), but it is not currently serving *factual* data. We suspect that the smaller size of OhioKG and the limited LSP library for OhioKG, we have not generated enough data for the LLM to be sufficiently fine-tuned.

4. Next Steps

This experiment has demonstrated some acceptable improvement in performance, but still leaves much room for improvement. In our next steps, we intend to explore how increasing the size of the KG, as well as the number of available LSPs, which will increase the amount of factually correct data available for fine-tuning. Indeed, we suspect that we can construct MODL-like libraries for KG fragments or ODPs for a reusable resource [13].

[1]https://colab.research.google.com/

References

[1] P. P. Ray, Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems 3 (2023) 121–154. URL: https://www.sciencedirect.com/science/article/pii/S266734522300024X. doi:https://doi.org/10.1016/j.iotcps.2023.04.003.

[2] K. Soman, P. W. Rose, J. H. Morris, R. E. Akbas, B. Smith, B. Peetoom, C. Villouta-Reyes, G. Cerono, Y. Shi, A. Rizk-Jackson, et al., Biomedical knowledge graph-enhanced prompt generation for large language models, arXiv preprint arXiv:2311.17330 (2023).

[3] N. F. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, J. Taylor, Industry-scale knowledge graphs: lessons and challenges, Commun. ACM 62 (2019) 36–43. URL: https://doi.org/10.1145/3331166. doi:10.1145/3331166.

[4] P. Hitzler, Semantic Web: A review of the field, Communications of the ACM (2021). To appear.

[5] G. Agrawal, T. Kumarage, Z. Alghami, H. Liu, Can knowledge graphs reduce hallucinations in llms?: A survey, arXiv preprint arXiv:2311.07914 (2023).

[6] S. Dernbach, K. Agarwal, A. Zuniga, M. Henry, S. Choudhury, Glam: Fine-tuning large language models for domain knowledge graph alignment via neighborhood partitioning and generative subgraph encoding, in: Proceedings of the AAAI Symposium Series, volume 3, 2024, pp. 82–89.

[7] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, Z. Li, Retrieval-augmented generation with knowledge graphs for customer service question answering, arXiv preprint arXiv:2404.17723 (2024).

[8] O. Ovadia, M. Brief, M. Mishaeli, O. Elisha, Fine-tuning or retrieval? comparing knowledge injection in llms, arXiv preprint arXiv:2312.05934 (2023).

[9] D. Maynard, A. Funk, W. Peters, Using lexico-syntactic ontology design patterns for ontology creation and population, in: Proceedings of the 2009 International Conference on Ontology Patterns - Volume 516, WOP'09, CEUR-WS.org, Aachen, DEU, 2009, p. 39–52.

[10] A. Gangemi, V. Presutti, Ontology design patterns, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, International Handbooks on Information Systems, Springer, 2009, pp. 221–243. URL: https://doi.org/10.1007/978-3-540-92673-3_10. doi:10.1007/978-3-540-92673-3_10.

[11] K. Janowicz, P. Hitzler, W. Li, D. Rehberger, M. Schildhauer, R. Zhu, C. Shimizu, C. K. Fisher, L. Cai, G. Mai, J. Zalewski, L. Zhou, S. Stephen, S. G. Estrecha, B. D. Mecum, A. Lopez-Carr, A. Schroeder, D. Smith, D. J. Wright, S. Wang, Y. Tian, Z. Liu, M. Shi, A. D'Onofrio, Z. Gu, K. Currier, Know, know where, knowwheregraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence, AI Mag. 43 (2022) 30–39. URL: https://doi.org/10.1609/aimag.v43i1.19120. doi:10.1609/aimag.v43i1.19120.

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[13] C. Shimizu, Q. Hirt, P. Hitzler, MODL: A modular ontology design library, in: K. Janowicz, A. A. Krisnadhi, M. P. Villalón, K. Hammar, C. Shimizu (Eds.), Proceedings of the 10th

Workshop on Ontology Design and Patterns (WOP 2019) co-located with 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 27, 2019, volume 2459 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 47–58.