Universidad de La Habana
Facultad de Matemática y Computación

Enfoques Zero-Shot para la Extracción de Conocimiento a partir de Lenguaje Natural

Autor: **Rolando Sánchez Ramos**

Tutor: **Dr. Alejandro Piad Morffis**

Trabajo de Diploma
presentado en opción al título de
Licenciado en Ciencias de la Computación

Noviembre de 2023

github.com/rolysr/nl2ql

Agradecimientos

Opinión del tutor

Dr. Alejandro Piad Morffis
Facultad de Matemática y Computación
Universidad de la Habana
Noviembre, 2023

Resumen

Esta tesis se centra en abordar la complejidad inherente a la consulta de bases de datos en forma de grafo, como Neo4J. Estas bases de datos a menudo requieren un conocimiento especializado en lenguajes de consulta, lo que limita su accesibilidad a un grupo reducido de usuarios con habilidades técnicas avanzadas. Para superar esta limitación, proponemos la aplicación del aprendizaje Zero-Shot, un enfoque innovador en el procesamiento del lenguaje natural. En esta investigación, se lleva a cabo un experimento basado en el modelo <variable>para traducir consultas de lenguaje natural a código *Cypher*. La evaluación se realiza utilizando el conjunto de datos de evaluación <variable>, que abarca una amplia variedad de ejemplos de consultas. Los resultados obtenidos, <variable>, establecen un punto de referencia esencial para el uso de modelos de lenguaje en la traducción de lenguaje natural a código *Cypher*.

Abstract

This thesis focuses on addressing the inherent complexity of querying graph databases like Neo4J. Such databases often require specialized knowledge of query languages, limiting accessibility to a small group of users with advanced technical skills. To overcome this limitation, we propose the application of Zero-Shot learning, an innovative approach in natural language processing. In this research, an experiment is conducted based on the <variable>model to translate natural language queries into *Cypher* code. Evaluation is carried out using the <variable>evaluation dataset, which encompasses a wide variety of query examples. The obtained results, <variable>, establish a crucial benchmark for the use of language models in translating natural language to *Cypher* code.

Índice general

Índice de figuras

Introducción

En la época actual, asistimos a un constante aumento en la producción de información en diversos formatos: visual, auditivo y textual, que abarca todos los ámbitos de la sociedad [4]. De manera particular, resulta sumamente intrigante la información generada a través del ingenio creativo y la investigación humana. Estos tipos de datos se almacenan debido a su relevancia y a la necesidad de acceder a ellos en el futuro, pudiendo optar por una organización estructurada o no. Sorprendentemente, solo alrededor del 20% de la información a nivel mundial se encuentra estructurada [1].

Las bases de conocimiento constituyen un tipo particular de bases de datos diseñadas para la administración del saber. Estas bases brindan los medios para recolectar, organizar y recuperar digitalmente un conjunto de conocimientos, ideas, conceptos o datos [2]. La ventaja fundamental de mantener la información de manera estructurada radica en su facilidad para ser consultada, ampliada y modificada. Debido a su utilidad y prevalencia, la recuperación de información a través de consultas en bases de conocimiento se ha convertido en una tarea esencial.

Es esencial que la información almacenada en bases de conocimiento adopte un formato adecuado para permitir búsquedas ágiles y precisas. Entre los formatos más comunes se encuentran los modelos de Entidad-Relación y el modelo Relacional. A pesar de ser enfoques más antiguos, el modelo Relacional (BDR) sigue siendo el más ampliamente utilizado en la actualidad [3]. No obstante, en ocasiones, las características específicas del problema demandan un formato más expresivo, y es en este punto donde las bases de datos orientadas a grafos (BDOG) [5] entran en juego.

Las BDOG han ganado progresivamente popularidad como una manera efectiva de almacenar información en los últimos años. Estas bases tienen la capacidad de modelar una diversidad de situaciones del mundo real al tiempo que mantienen un alto nivel de simplicidad y legibilidad para

los seres humanos. Las BDOG presentan numerosas ventajas en comparación con las bases de datos relacionales. Esto incluye un mejor rendimiento, permitiendo el manejo más rápido y eficaz de grandes volúmenes de datos relacionados; flexibilidad, ya que la teoría de grafos en la que se basan las BDOG permite abordar diversos problemas y encontrar soluciones óptimas; y escalabilidad, ya que las bases de datos orientadas a grafos permiten una escalabilidad eficaz al facilitar la incorporación de nuevos nodos y relaciones entre ellos. Ejemplo de un sistema de gestión de BDOG es *Neo4J* [**?**], a través del cual es posible construir instancias de este tipo de base de datos e interactuar con las mismas a través del lenguaje de programación *Cypher* [**?**], el cual posee una sintaxis declarativa similar a SQL [**?**].

Por otro lado, el avance en la comprensión del lenguaje natural se ha visto potenciado con el surgimiento de los grandes modelos de lenguajes (LLMs) [**?**] como GPT-4 [**?**] o LLaMA-2 [**?**], los cuales presentan una serie de habilidades emergentes como elaboración de resúmenes de textos, generación de código, razonamiento lógico, traducción lingüística entre otras [**?**]. Dichas herramientas constituyen modelos de *Machine Learning* entrenados con un gran volumen de datos, lo cual es posible gracias al número de parámetros con los que estos son configurados [**?**].

Usualmente, para el uso de los LLMs basta con ofrecerles como dato de entrada un texto (*prompt*), el cual describe la tarea que se espera que estos realicen. Además, son muchas las técnicas existentes para elaborar una entrada de calidad, esto con el objetivo de que la respuesta por parte de dicho modelo de lenguage ofrezca resultados alentadores al respecto, lo cual se conoce como *prompt engineering* [**?**]. Una técnica bastánte común es *Zero-Shot Learning* (ZSL) [**?**], la cual consiste en describirle a un LLM un procedimiento a realizar sin ofrecer de antemano ejemplos de cómo resolverlo, como por ejemplo, en tareas relacionadas con la generación de código, donde algunos de estos son capaces de generar algoritmos expresados en un lenguaje de programación formal a partir de una sentencia o consulta en lenguaje natural sin recibir como entrada del usuario algunos ejemplos de código, o especificaciones de cómo funciona el lenguaje objetivo a generar [**?**] [**?**].

En lo que respecta a la comprensión del lenguaje natural y su uso en consultas a bases de conocimiento, existen diversas vías llevadas a cabo y con resultados diversos, donde se hacen análisis sintácticos y semánticos sobre la consulta, muchas veces asistidos por diccionarios o mapas sobre la base de conocimiento en cuestión. Se usan modelos de paráfrasis como técnica de aumento de datos y finalmente Transformers [**?**] o incluso LLMs para llevar de la consulta ya curada al lenguaje de consulta formal o a un

lenguaje intermedio capaz de expresar a esta a alto nivel [**?**] [**?**] [**?**] [**?**].

Por las razones anteriormente expuestas, resulta interesante la investigación sobre la tarea de generación de código de consulta formal a partir de una sentencia en lenguaje natural mediante el uso de LLMs, especialmente el diseño e implementación un experimento capaz de demostrar las capacidades reales de estos para dicho acometido, lo cual designará la importancia de continuar el estudio de dichas herramientas con el objetivo de mejorar los sistemas de extracción de conocimientos en BDOG.

Problemática

Para utilizar el lenguaje de consulta formal *Cypher* se requiere de conocimientos básicos de programación, lo cual consume cierto tiempo y esfuerzo. Esto tiene como consecuencia que, solo aquellas personas con experiencia en el uso de lenguajes de programación puedan hacer uso de la mayoría de los sistemas de almacenamiento de datos desarrollados con esta tecnología y teniendo en cuenta la necesidad de poseer un conocimiento del dominio sobre el cual está construida la base de datos a consultar. Por lo tanto, llevar a cabo una mejora en las herramientas orientadas a democratizar dicho proceso permitiría hacer más rápido y eficiente dicho proceso de consulta en cuanto a tiempo y recursos computacionales. Debido a dicha situación, se propone una experimentación basada en un LLM capaz de traducir una consulta en lenguaje natural a un código en *Cypher*, donde a su vez se verifique la efectividad de este a partir de enfoques basados en ZSL, los cuales intuitivamente pueden ofrecer como resultado una cota inferior para la efectividad de sistemas desarrollados en base a dichos algoritmos de aprendizaje. Además, actualmente la implementación de sistemas de generación de código de consulta formal está principalmente orientada al lenguaje SQL, mientras que para el lenguaje *Cypher*, no existen suficiente estudios recientes que avalen la calidad de tales herramientas para dicho caso de uso.

Objetivos

Dadas las ideas anteriores, los objetivos principales del trabajo consistirá en diseñar e implementar una estrategia experimental capaz de evaluar la capacidad mínima de los LLMs para la consulta en lenguaje natural a bases de conocimiento estructuradas con independencia del dominio, para lo cual se empleará un enfoque basado en ZSL.

Para lograr los objetivos generales se trazaron los siguientes objetivos específicos:

1. Estudiar el estado del arte de los modelos de Aprendizaje Automático capaces de hacer predicciones de tipo texto-a-texto.
2. Analizar el trabajo de tesis sobre este tema anteriormente desarrollado en la facultad.
3. Implementar un modelo de Aprendizaje Automático capaz de convertir una consulta en lenguaje natural humano a un lenguaje formal que permita obtener datos a partir de una base de conocimiento.
4. Explorar las capacidades de enfoques Zero-Shot para la traducción de lenguaje natural al lenguaje Cypher.
5. Mejorar el sistema de evaluación de resultados permitiendo que el conjunto de datos de prueba y evaluación sea lo más realista posible y con una mayor complejidad.

Organización de la tesis

[Hablar sobre la estructuracion del documento]

Capítulo 1

Preliminares

Conclusiones y Recomendaciones

Bibliografía

[1] Amazon. What is structured data? https://aws.amazon.com/es/what-is/structured-data/, 2023. (Citado en la página 9).

[2] Web Archive. Creación de una base de conocimiento. https://web.archive.org/web/20180315025341/http://es.ccm.net/faq/2158-organizacion-crear-una-base-de-conocimientos, 2023. (Citado en la página 9).

[3] SAP insights. ¿qué es el modelado de datos? https://www.sap.com/latinamerica/products/technology-platform/datasphere/what-is-data-modeling.html, 2023. (Citado en la página 9).

[4] Fundación MAPFRE. ¿cuánta información se genera y almacena en el mundo? https://www.fundacionmapfre.org/blog/cuanta-informacion-se-genera-y-almacena-en-el-mundo/, 2023. (Citado en la página 9).

[5] Wikipedia. Graph database. https://en.wikipedia.org/wiki/Graph_database, 2023. (Citado en la página 9).