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

Estado del Arte

Con el incremento constante de la cantidad de información generada en todo el mundo, la recuperación de información se ha convertido en un aspecto de creciente relevancia tanto en el ámbito industrial como en el académico. En consecuencia, la reducción del tiempo transcurrido entre el momento en que un usuario desea acceder a la información y el momento en que efectivamente puede hacerlo ha sido objeto de un número creciente de investigaciones científicas en los últimos años. Este capítulo se dedica a la evaluación de diversas estructuras de interfaces entre el ser humano y bases de conocimiento, las cuales tienen como objetivo abordar esta problemática.

Las Interfaces de Lenguaje Natural a Bases de Datos (NLIDB, por sus siglas en inglés) representan un campo de investigación dinámico centrado en facilitar las interacciones entre humanos y computadoras con bases de datos relacionales utilizando consultas en lenguaje natural. A lo largo de las últimas décadas, el desarrollo de NLIDB ha pasado por varias fases transformadoras, impulsadas por avances tecnológicos y metodológicos, así como por una creciente demanda de una mejor accesibilidad a las bases de datos.

Las etapas iniciales del desarrollo de NLIDB se caracterizaron por sistemas específicos de dominio. Estos sistemas fueron diseñados para trabajar dentro de áreas de conocimiento bien definidas, donde se utilizaba el procesamiento de lenguaje natural controlado para garantizar la comprensión de las consultas y la interacción con la base de datos. Por ejemplo, algunos trabajos pioneros [**?**] [**?**] demostraron NLIDBs que se adaptaban a dominios específicos, lo que los hacía altamente efectivos pero limitados en alcance. Del mismo modo, en años posteriores [**?**] [**?**] se continuó la exploración del

uso de interfaces de lenguaje natural controlado dentro de dominios de conocimiento particulares.

Otro enfoque durante esta fase implicó NLIDBs basados en reglas. Algunos sistemas propuestos al respecto [**?**] dependían de reglas predefinidas para traducir consultas en lenguaje natural en declaraciones *SQL* para la recuperación de datos en la base de datos. Si bien estos sistemas ofrecían ciertas ventajas, carecían de versatilidad para manejar una amplia gama de consultas de usuarios en diferentes dominios.

A medida que avanzaba la investigación en NLIDB, hubo un cambio hacia la independencia de dominio y la flexibilidad. Los sistemas recientes han buscado reducir la dependencia del conocimiento específico del dominio y las reglas. Algunos investigadores [**?**] [**?**]) han desarrollado NLIDBs que utilizan técnicas de aprendizaje supervisado, lo que los hace más adaptables a varios dominios y entradas de usuario. Además, un avance significativo se ha producido con la integración de redes neuronales profundas en el desarrollo de NLIDBs, donde varias investigaciones [**?**] [**?**] han demostrado el potencial del aprendizaje profundo en NLIDB, aprovechando vastos repositorios de texto y código para el entrenamiento. Este enfoque ha mejorado significativamente el rendimiento de NLIDB, permitiendo un procesamiento de consultas más natural y contextual.

Para el caso específico de NLIDB con respecto a BDOGs inicialmente se desarrollaron trabajos enfocados en técnicas similares a las empleadas para *SQL* [**?**] [**?**], donde se realizaba un preprocesamiento en la consulta dada, se aprovechaba la información ofrecida por el esquema [**?**] de la base de datos a consultar y finalmente dicho conocimiento era utilizado por un modelo de aprendizaje profundo entrenado sobre un conjunto de pares de lenguaje natural y lenguaje de consulta formal como por ejemplo *Cypher*.

Recientemente, los trabajos orientados a esta área de estudio han estado enfocados en dos metodologías principales:

1. Enfoques neurosimbólicos basados en representación intermedia de la consulta dada en lenguaje natural.

2. Enfoques basados en técnicas de *prompt engineering* mediante LLMs.

1.1. Enfoques neurosimbólicos basados en representación intermedia (IR, por sus siglas en inglés) de la consulta dada en lenguaje natural

Con respecto a la recuperación de información de una base de conocimiento con este enfoque, mediante consultas hechas en lenguaje natural se han hecho varias investigaciones científicas que se pueden agrupar bajo el patrón de la Figura 1.1.

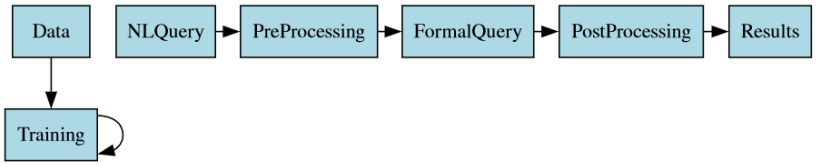

Figura 1.1: Secuencia de flujo representativa del Estado del Arte de traducción de Lenguaje Natural a Lenguaje Formal utilizando el enfoque neurosimbólico basado en (IR).

1.1.1. Conjuntos de entranamiento y evaluación

En cuanto a los datos para el entrenamiento existen dos opciones. Básicamente se puede buscar un conjunto de datos de referencia(*Benchmark*) en los que entrenar y probar el sistema, o se crea uno. La mayoría de las investigaciones existentes elijen la primera opción. Algunos de los *Benchmarks* más populares son:

- **WikiSQL**: WikiSQL [**?**] es el banco de datos más grande y más utilizado, contiene 26531 tablas y 80654 pares de consultas en lenguaje natural y lenguaje *SQL*. Las tablas se extraen de tablas *HTML* de Wikipedia. Luego, cada consulta en *SQL* se genera automáticamente para una determinada tabla bajo la restricción de que la consulta produce un conjunto de resultados no vacío.
- **Spider**: Este *Benchmark* [30] es un punto de referencia multidominio a gran escala con 200 bases de datos de 138 dominios diferentes y 10.181 pares de consultas.

- **MetaQA**: El conjunto de datos METAQA-Cypher, originalmente conocido como METAQA [**?**], contiene más de 400000 pares de preguntas y respuestas de múltiples pasos obtenidos de la base de conocimiento WikiMovies [**?**]. Mientras que investigaciones previas se han centrado principalmente en la anotación SPARQL [**?**], nuestra innovación implica reconfigurar METAQA en *Cypher*, estableciéndolo como un valioso punto de referencia para el aprendizaje de pocos ejemplos.

En el caso alternativo, se suelen usar las propias bases de conocimiento objeto de estudio para crear un conjunto de entrenamiento. Una de las técnicas empleadas para esto es Random Walk [**?**], en la que se hace un recorrido aleatorio sobre un subconjunto de las entidades y relaciones de la base de conocimiento, y se elaboran consultas artificiales que respondan a dichas entidades y relaciones [**?**].

1.1.2. Preprocesamiento

En general las investigaciones en el área realizan algún tipo de preprocesamiento a la consulta. Algunos realizan parafraseo para llevar la consulta a representaciones canónicas [**?**], en su lugar otros realizan un análisis morfológico léxico, con toquenización, lematización, eliminación de stopwords, tagueo de partes de la oración (*POS-taguing*), etc. [**?**]. También se encuentran las investigaciones que usan en esta etapa traducciones basadas en diccionarios especializados en el dominio, o de propósito general, al igual que ontologías, para "suavizar"vocablos difíciles de entender por el resto del sistema [**?**]. Además se usan técnicas como vectorización de palabras (*word2vector*), y transformación de la consulta a una representación en grafos [**?**].

1.1.3. Postprocesamiento

En la fase de Post-Procesamiento, se observan diversas enfoques en las investigaciones. La mayoría de los investigadores realizan un análisis semántico que implica la clasificación según tipos de datos, el uso de ontologías y bases de conocimiento externas para realizar mapeos [**?**] [**?**]. Algunos optan por convertir la consulta en una representación canónica, mientras que otros la transforman en un lenguaje intermedio antes de transpilarla al lenguaje objetivo [**?**]. También hay quienes la codifican directamente en forma de grafo, y algunos la convierten en embeddings [**?**].

1.1.4. Consulta

En la fase de construcción de la consulta, existen tres enfoques principales. El primero es un enfoque manual que implica la búsqueda y formateo de palabras clave como *"where"* y *"select"*, además del mapeo de atributos a tablas [**?**]. Otra vía se centra en el uso de modelos, incluyendo decodificadores y en algunos casos redes neuronales convolucionales [**?**]. También se emplea la construcción de la consulta formal mediante un compilador, especialmente cuando se ha utilizado un lenguaje intermedio entre el lenguaje natural y el formal de consulta[**?**].

1.2. Enfoques basados en técnicas de *prompt engineering* mediante LLMs

Con la creciente atención dada a los modelos de lenguaje a gran escala, se han convertido en un componente esencial en el procesamiento del lenguaje natural. A medida que aumenta el tamaño de los modelos preentrenados, también está cambiando gradualmente su uso. A diferencia de modelos como BERT [**?**] y T5 [**?**], que requieren un proceso de entrenamiento con una pequeña cantidad de datos, modelos como GPT-3 [**?**] requieren un diseño de un texto de entrada para generar resultados deseados. El reciente modelo de *ChatGPT* [**?**], que emplea Aprendizaje por Reforzamiento para la Retroalimentación Humana (RLHF) [**?**], simplifica el diseño de textos de entrada de calidad, lo que permite una mejor utilización de la capacidad de ZSL de modelos preentrenados a gran escala de manera conversacional. Debido a la sólida capacidad de dichos en la generación de código [**?**] y al hecho de que los modelos de generación de código suelen requerir una gran cantidad de datos anotados para producir buenos resultados [**?**], un modelo de generación de código de ZSL se considera fundamental.

Para la tarea específica de generar código de un lenguaje de consulta formal como *SQL* y *Cypher* se han utilizado distintos enfoques basados en dos de las principales técnicas de *prompt engineering*:

- **Zero-Shot Learning**: Esta técnica se enfoca en la capacidad de un modelo para comprender y generar código en un lenguaje de consulta sin requerir ejemplos específicos de entrenamiento en ese lenguaje en particular. En otras palabras, el modelo puede realizar esta tarea "desde cero", sin conocimiento previo del lenguaje. Algunos estudio interesantes se han realizado principalmente en la tarea de traducir lenguaje natural a lenguaje *SQL* [**?**] [**?**], los cuales permiten inferir la

calidad mínima que estos modelos pueden alcanzar en la realización de dicha tarea [**?**].

- **Few-Shot Learning**: En contraste, el enfoque de Few-Shot Learning (FSL) se basa en la idea de que el modelo tiene acceso a un pequeño número de ejemplos (pocos ejemplos) en el lenguaje de consulta deseado para mejorar su capacidad de generar código en ese lenguaje. Esto puede ser especialmente útil cuando se necesita una adaptación rápida a un nuevo lenguaje o contexto. Tal y como muestran algunos resultados experimentales, este enfoque puede tener resultados superiores a varios modelos basados en *fine-tuning* [**?**].

  [INSERTAR FIGURAS DE ZSL y FSL]

  A pesar del uso comprobado de dichos enfoques, todavía existe una escasez de estudios orientados a la traducción de lenguaje natural a código de consulta a BDOG como por ejemplo *Cypher* [**?**], por lo tanto es visible la necesidad de elaborar experimentos enfocados en dicha tarea dicha tarea.

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).