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

Con el aumento gradual de la información en la actualidad, el proceso de organización y extracción de conocimiento en base a esta se ha convertido en una tarea fundamental. La razón principal, es la necesidad de almacenar dichos datos con el objetivo de ser consultados en un futuro de forma eficiente. Por lo tanto, para llevar a cabo dicho reto, ha sido imprescindible el desarrollo de sistemas capaces de persistir información de forma estructurada y facilitar el acceso a esta.

Las bases de datos orientadas a grafos [1] constituyen herramientas que permiten el almacenamiento y consulta de información de manera escalable y segura. Un ejemplo de estas es *Neo4J* [2], con la cual se puede interactuar a partir del lenguaje de programación *Cypher* [3].

Para utilizar el lenguaje de consulta *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. Por lo tanto, es necesaria una herramienta que permita democratizar dicho proceso, para lo cual se propone un modelo capaz de traducir una consulta en lenguaje natural a un código en *Cypher*. Además, también es objetivo de este trabajo experimentar con los límites del aprendizaje *Zero-Shot* [4] para dicha tarea.

Objetivos

<Objetivos generales>

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

1. Objetivo 1.

[Hablar sobre la estructuracion del documento]

Capítulo 1

Preliminares

Conclusiones y Recomendaciones

Bibliografía

[1] Wikipedia. URL: https://en.wikipedia.org/wiki/Graph_database. Consultado en 18 de octubre de 2023. (Citado en la página 9).

[2] Neo4J. URL: https://neo4j.com/. Consultado en 18 de octubre de 2023. (Citado en la página 9).

[3] Neo4J. URL: https://neo4j.com/docs/getting-started/current/cypher-intro/. Consultado en 18 de octubre de 2023. (Citado en la página 9).

[4] Wikipedia. URL: https://en.wikipedia.org/wiki/Zero-shot_learning. Consultado en 18 de octubre de 2023. (Citado en la página 9).

[5] Dar, H. S., Lali, M. I., Din, M. U., Malik, K. M., and Bukhari, S. A. C. (2019). *Frameworks for querying databases using natural language: a literature review.* arXiv preprint arXiv:1909.01822 (vid. págs. 3, 8-11, 19).

[6] Wu, S., Chen, B., Xin, C., Han, X., Sun, L., Zhang, W., Chen, J., Yang, F., and Cai, X. (2021). *From paraphrasing to semantic parsing: Unsupervised semantic parsing via synchronous semantic decoding.* arXiv preprint arXiv:2106.06228 (vid. págs. 2, 7, 8).

[7] Cai, R., Yuan, J., Xu, B., and Hao, Z. (2021). *SADGA: Structure-Aware Dual Graph Aggregation Network for Text-to-SQL.* Advances in Neural Information Processing Systems, 34, 7664-7676 (vid. págs. 2, 7-9).

[8] Nie, L., Cao, S., Shi, J., Tian, Q., Hou, L., Li, J., and Zhai, J. (2022). *Unifying Semantic Parsing of Graph Query Language with Intermediate Representation.* arXiv preprint arXiv:2205.12078 (vid. págs. 7-9).

[9] Bazaga, A., Gunwant, N., and Micklem, G. (2021). *Translating synthetic natural lan- guage to database queries with a polyglot deep learning framework.* Scientific Reports, 11 (1), 1-11 (vid. págs. 6-8).

[10] GitHub. URL: https://github.com/rolysr/nl2ql. Consultado en 18 de octubre de 2023.

[11] Wikipedia. URL: https://es.wikipedia.org/wiki/GPT-3. Consultado en 18 de octubre de 2023.

[12] IBM. URL: https://www.ibm.com/topics/database-schema. Consultado en 18 de octubre de 2023.

[13] Neo4J. URL: https://neo4j.com/docs/getting-started/current/appendix/example-data/. Consultado en 18 de octubre de 2023.