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

Rolando Sánchez Ramos ROLSANCHEZ@YANDEX.COM
Grupo C411

Tutor(es):
Dr. Alejandro Piad Morffis, *MatCom, Universidad de La Habana*

Resumen

Este trabajo tiene como objetivo experimentar las capacidades del aprendizaje *Zero-Shot* en el proceso de traducción de una consulta en lenguaje natural a un código del lenguaje *Cypher*, con el fin de facilitar el proceso de extracción de conocimiento en bases de datos en forma de grafo, como es el caso de *Neo4J*. Se realiza una introducción breve a la problemática y se mencionan las principales técnicas en el estado del arte para darle solución a la misma. Luego, se expone el flujo de trabajo del experimento, el cual posee como componente fundamental el modelo *GPT-3*. Finalmente, se realiza un experimento con 108 consultas sintéticas para validar la calidad del modelo mencionado para la tarea propuesta, obteniéndose como resultado un 45.37 % de acierto en las respuestas de las consultas generadas al ser ejecutadas sobre una base de datos *Neo4J*. De esta forma, se obtiene una interesante base para continuar explorando las capacidades del enfoque *Zero-Shot* para la traducción de texto a código.

Abstract

This work aims to test the capabilities of *Zero-Shot* learning in the translation process of a natural language query into a *Cypher* language code, in order to facilitate the knowledge extraction process in databases in graph form, as is the case of *Neo4J*. A brief introduction to the problem is made and the main techniques in the state of the art are mentioned to solve it. Then, the workflow of the experiment is exposed, which has the *GPT-3* model as a fundamental component. Finally, an experiment is carried out with 108 syntactic queries to validate the quality of the mentioned model for the proposed task, obtaining as a result a 45.37 % hit in the responses of the queries generated when executed on a *Neo4J* database. In this way, an interesting base is obtained to continue exploring the capabilities of the *Zero-Shot* approach for text-to-code translation.

Palabras Clave: Inteligencia Artificial, Gran Modelo de Lenguaje, Aprendizaje Zero-Shot.
Tema: Inteligencia Artificial, Procesamiento de Lenguaje Natural.

1. Resumen Extendido

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 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.

1.1 Estado del Arte

El problema de traducir un texto en lenguaje natural a una consulta formal para interactuar con una base de conocimientos ha sido una tarea ampliamente estudiada los campos de la Inteligencia Artificial y el Procesamiento de Lenguaje Natural. Ejemplos son los enfoques basados en reglas [5], redes neuronales convolucionales y recurrentes [6] [7], técnicas de compilación

con análisis sintáctico y semántico [8] y de manera reciente, el uso de *transformers* [9].

1.2 Propuesta e Implementación [10]

Se propone el uso del modelo *GPT-3 (text-davinci-003)* [11], un Gran Modelo de Lenguaje entrenado para tareas como la traducción de texto y la generación de código. También, se diseñó un componente denominado *Graph Contractor*, capaz de interactuar con una base de datos *Neo4J*. El flujo de funcionamiento del experimento sería: El modelo *GPT-3* recibe una consulta en lenguaje natural y un esquema [12] de la base de datos a utilizar (base de datos sobre películas famosas) [13], luego este ofrece la consulta en *Cypher* correspondiente a la entrada dada y es ejecutada sobre la base de datos. La idea fundamental del experimento consiste en utilizar el modelo mencionado sin proveerle ejemplos de cómo son las consultas de *Cypher* ni haber sido entrenado para este tipo específico de traducción, lo cual se conoce como aprendizaje *Zero-Shot*.

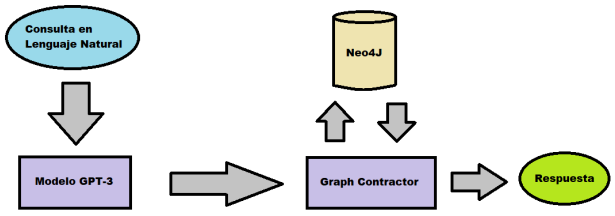

Figura 1: Arquitectura del experimento.

1.3 Resultados Experimentales

Para evaluar la calidad de la propuesta de solución se generaron 108 casos de prueba en forma de pares de consultas en lenguaje natural con su código de *Cypher* correspondiente. Todas las consultas de *Cypher* generadas por el modelo fueron correctamente compiladas, y de estas, 49 (45.37 %) dieron exactamente la misma respuesta que las consultas de prueba al ser ejecutadas sobre la base de datos.

Referencias

[1] Wikipedia. URL: https://en.wikipedia.org/wiki/Graph_database. Consultado en 22 de abril de 2023.

[2] Neo4J. URL: https://neo4j.com/. Consultado en 22 de abril de 2023.

[3] Neo4J. URL: https://neo4j.com/docs/getting-started/current/cypher-intro/. Consultado en 22 de abril de 2023.

[4] Wikipedia. URL: https://en.wikipedia.org/wiki/Zero-shot_learning. Consultado en 22 de abril de 2023.

[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 22 de abril de 2023.

[11] Wikipedia. URL: https://es.wikipedia.org/wiki/GPT-3. Consultado en 22 de abril de 2023.

[12] IBM. URL: https://www.ibm.com/topics/database-schema. Consultado en 22 de abril de 2023.

[13] Neo4J. URL: https://neo4j.com/docs/getting-started/current/appendix/example-data/. Consultado en 22 de abril de 2023.