

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

Rolando Sánchez Ramos

Grupo C411

ROLSANCHEZ@YANDEX.COM

Tutor(es):

Dr. Alejandro Piad Morffis, *MatCom, Universidad de La Habana*

Resumen

En este trabajo, presentamos enfoques de aprendizaje Zero-Shot para la extracción de conocimiento a partir de lenguaje natural, con el objetivo de mejorar la comprensión y el análisis de información en texto. Revisamos el estado del arte y proponemos una arquitectura novedosa basada en el modelo GPT-3 para abordar este desafío. A través de experimentos, demostramos que nuestra propuesta alcanza una efectividad del 45.37% en la traducción de consultas sintéticas generadas, lo que indica un avance significativo en la capacidad de comprensión y procesamiento del lenguaje humano en tareas de extracción de conocimiento. Nuestros resultados también revelan oportunidades para mejorar la precisión y robustez de los enfoques Zero-Shot en futuras investigaciones.

Abstract

In this paper, we present Zero-Shot learning approaches for knowledge extraction from natural language, with the goal of enhancing the understanding and analysis of textual information. We review the state-of-the-art and propose a novel architecture based on the GPT-3 model to tackle this challenge. Through experiments, we demonstrate that our approach achieves *a*45.37% effectiveness for the synthetic queries generated, indicating significant progress in the understanding and processing of human language in knowledge extraction tasks. Our results also reveal opportunities to improve the accuracy and robustness of Zero-Shot approaches in future research.

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 [?] constituyen herramientas que permiten el almacenamiento y consulta de información de manera escalable y segura. Un ejemplo de estas es *Neo4J* [?], con la cual se puede interactuar a partir del lenguaje de programación *Cypher* [?].

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* [?] 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. Ejemplos son los enfoques basados en reglas [?], redes neuronales convolucionales y recurrentes [?], técnicas de compilación con análisis sintáctico y semántico [?] y de manera reciente, el uso de *transformers* [?].

1.2 Propuesta e Implementación

Se propone el uso del modelo *GPT-3 (text-davinci-003)* [?], entrenado para tareas como la traducción de texto y la generación de código. También, se diseñó un componente denominado *GraphContractor*, 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 [?] de la base de datos a utilizar (base datos sobre películas famosas) [?], 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*.

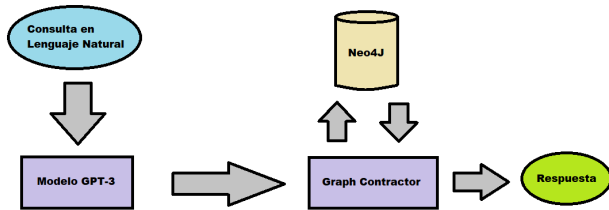


Figure 1: Arquitectura del experimento.

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.

References

- [1] Wikipedia. URL: https://en.wikipedia.org/wiki/Graph_database. Consultado en April 21, 2023.
- [2] Neo4J. URL: <https://neo4j.com/>. Consultado en April 21, 2023.
- [3] Neo4J. URL: <https://neo4j.com/docs/getting-started/current/cypher-intro/>. Consultado en April 21, 2023.
- [4] Wikipedia. URL: https://en.wikipedia.org/wiki/Zero-shot_learning. Consultado en April 21, 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 language to database queries with a polyglot deep learning framework*. Scientific Reports, 11 (1), 1-11 (vid. págs. 6-8).
- [10] Wikipedia. URL: <https://es.wikipedia.org/wiki/GPT-3>. Consultado en April 21, 2023.
- [11] IBM. URL: <https://www.ibm.com/topics/database-schema>. Consultado en April 21, 2023.
- [12] Neo4J. URL: <https://neo4j.com/docs/getting-started/current/appendix/example-data/>. Consultado en April 21, 2023.