Modelo generativo de SQL a partir de consultas en español

Trabajo Terminal No. ____ - ____

*Alumnos: Rubio López Zury Yael, Miranda Chávez Victor Ulises, *Nicolas Hernandez Adair, Perez Sanchez Ives Lancelote*

Directores: Carmona García Enrique Alfonso, Flores Estrada Ituriel Enrique

**e-mail: adairnicolas2@gmail.com*

Resumen - Este trabajo terminal tiene como objetivo crear un modelo de aprendizaje automático que transforme consultas formuladas por usuarios en lenguaje natural (español) a su equivalente en SQL (Structured Query Language) utilizando técnicas de procesamiento del lenguaje natural y aprendizaje máquina. El proyecto busca mejorar significativamente la experiencia del usuario al ofrecer una forma más eficiente y precisa de realizar consultas en bases de datos relacionales.

Palabras clave - Bases de Datos, Inteligencia Artificial, Procesamiento del Lenguaje Natural, SQL

1. Introducción

SQL, también conocido como Structured Query Language, se originó en 1974. Donald Chamberlin y Raymond Boyce crearon un lenguaje para manipular y gestionar los datos almacenados en bases de datos relacionales. En su etapa inicial, el proyecto se conocía como "SEQUEL", que era un acrónimo de Structured English Query Language, lo que demostraba la intención de sus creadores de hacer que el lenguaje fuera similar al idioma inglés para que los usuarios pudieran leer las consultas de manera más natural y comprensible [1]. Posteriormente el nombre se cambió a SQL, el cual se convirtió en el lenguaje de consulta estándar para administrar y manipular bases de datos relacionales.

Aunque SQL puede leerse de forma intuitiva y puede ser muy útil para encontrar información valiosa, acceder al contenido de bases de datos relacionales requiere de preparación técnica. A medida que aumenta la necesidad de información más específica y detallada, las consultas pueden volverse más complejas y difíciles de manejar si no se cuenta con un conocimiento profundo del lenguaje y sobre la estructura de la base de datos en cuestión. Debido a esta limitación, los científicos de la computación han intentado darle solución a esta tarea de conversión de lenguaje natural a SQL, a la cual se denominó “text to SQL”, que busca convertir el texto de una consulta en lenguaje natural a SQL.

Uno de los primeros enfoques propuestos para esta tarea fue el de Woods para el proyecto LUNAR [2]. LUNAR fue capaz de comprender consultas en lenguaje natural relacionadas con una base de datos que contenía análisis químicos de rocas lunares. Sin embargo, tanto el sistema de Woods como muchos de los sistemas desarrollados en los años posteriores tenían una limitación importante: estaban diseñados para trabajar con bases de datos específicas. Lo anterior hacía difícil, e incluso imposible, adaptarlos para trabajar con bases de datos externas.

Desde la aparición de LUNAR y hasta la actualidad, muchos de estos sistemas han sido desarrollados con el inglés en mente como idioma de consulta. Los pocos sistemas que han surgido con el español como idioma de consulta suelen ser diseñados para ser utilizados con una base de datos en especifico [3][4], lo cual limita el alcance del sistema y evita que este pueda ser utilizado con otras bases de datos. Esto nos indica el poco avance sobre la resolución de la tarea ya mencionada pero en el idioma español.

La comunidad de la inteligencia artificial ha enfocado sus esfuerzos en el estudio de la tarea “text to SQL” en el idioma inglés, debido a esto, para el español no se han presentado muchos avances en esta área a pesar de la revolución que ha supuesto el aprendizaje profundo en las tareas de generación de texto. En contraste tenemos a la comunidad de habla inglesa, la cual ha aprovechado los avances en el estado del arte del aprendizaje automático y ha alcanzado nuevos niveles de efectividad mediante distintos enfoques. Por ejemplo, tenemos el propuesto por Cai, que hizo uso de un enfoque basado en grafos al cual denominó SADGA (Structure-Aware Dual Graph Aggregation Network for Text-to-SQL)[5], o el propuesto por Mellah al usar un modelo de lenguaje pre-entrenado como T5 (Text-to-text transfer transformer) [6].

Con este trabajo terminal se pretende retomar la tarea de convertir consultas en lenguaje natural al lenguaje SQL para el idioma español. Se busca así promover el avance de la inteligencia artificial en la comunidad de habla hispana aprovechando las oportunidades no exploradas por otros investigadores, tales como la independencia de

un modelo para realizar consultas en cualquier base de datos y el uso de nuevas técnicas de aprendizaje profundo. Esto permitirá a los usuarios hacer consultas de manera sencilla y accesible sin invertir mucho tiempo creando las consultas en SQL, y además ayudará a reducir la brecha con la comunidad de habla inglesa en este campo.

2. Objetivo

Objetivo general:

Implementar un modelo generativo de SQL a partir de consultas en español. El modelo podrá ser utilizado como parte de una herramienta que facilite a los usuarios el proceso de consulta de información en bases de datos relacionales. Esto se logrará a través de un enfoque que combina técnicas de procesamiento del lenguaje natural y aprendizaje automático, lo cual permitirá a los usuarios formular preguntas de manera más natural y obtener resultados precisos sin tener que invertir tanto tiempo en redactar su consulta en lenguaje SQL.

Objetivos específicos:

- Realizar una investigación exhaustiva de las estructuras sintácticas y semánticas correctas de SQL y de cómo se relacionan con las expresiones naturales humanas.
- Construir un conjunto de ejemplos etiquetados para entrenar al modelo, incluyendo una gran cantidad de consultas en español, esquemas de bases de datos, y sus correspondientes representaciones en SQL.
- Seleccionar una arquitectura de red neuronal profunda con la capacidad suficiente para capturar patrones complejos de dependencias entre palabras y partes de la frase en el texto en español y sus equivalentes en SQL.
- Crear e implementar un modelo de aprendizaje automático para convertir consultas escritas en español a consultas SQL.
- Asegurar la efectividad y corrección del modelo generado, evaluando y validando a través de pruebas exhaustivas.

3. Justificación

La gestión de bases de datos (BD) es una tarea crítica en cualquier organización moderna para poder manipular y consultar las grandes cantidades de datos. Según un estudio realizado por IDC (International Data Corporation) en 2018, se espera que la cantidad de datos en el mundo alcance los 175 zettabytes para el año 2025 [7]. Sin embargo, la manipulación de estas BD se realiza típicamente utilizando el lenguaje SQL, lo que puede resultar en una actividad difícil y compleja. Incluso la simple tarea de consultar los datos de una persona dentro de la organización requeriría de alguien con conocimientos en SQL para lograr dicha consulta en la BD.

La idea anterior nos conduce a la necesidad crítica de contar con personal especializado en el manejo del lenguaje SQL para poder interactuar con los datos que se encuentran en la BD. Por esta razón, nuestra propuesta es desarrollar un modelo que pueda ser utilizado como parte de las herramientas de trabajo de los profesionales, de manera que contribuya a aumentar la eficiencia y el desempeño de aquellos que trabajen con consultas SQL.

Existen trabajos publicados para abordar esta tarea en el idioma español, los cuales se basan en el análisis de la estructura de la oración [4][8], mientras que otros fuerzan al usuario a escribir su consulta con reglas preestablecidas [3], lo que impide la generación de consultas que no cumplan con estas y reduce la capacidad de expresividad del usuario. Además, algunos modelos están diseñados para trabajar con una sola base de datos [4], lo que imposibilita su uso en otras bases de datos.

Por otro lado, los modelos basados en tecnologías de vanguardia se centran principalmente en el idioma inglés [9]. Dichos modelos excluyen a la comunidad hispanohablante de tener acceso a modelos de inteligencia artificial para resolver esta problemática. Con la llegada de ChatGPT, un gran modelo de lenguaje (LLM por sus siglas en inglés) entrenado con más de 175 mil millones de parámetros [10], se han logrado importantes avances en distintas tareas que involucran el lenguaje natural.Y, aunque herramientas como ChatGPT han avanzado significativamente en distintas labores, incluyendo la de text-to-SQL [11] y su variante para el español, este problema aún no está resuelto, y muchas de las posibles rutas de investigación en este campo para el idioma español aún no han sido exploradas.

Nuestra propuesta no busca competir con herramientas como ChatGPT, sino aportar al campo de la generación de consultas SQL a partir de consultas en español, esperando que este trabajo pueda servir como una base sólida para futuras investigaciones en el área de la generación de consultas SQL. Así que, el presente trabajo buscará

implementar la combinación de técnicas de procesamiento de lenguaje natural (PLN) y aprendizaje automático (ML) para transformar consultas de lenguaje natural (en español) a consultas en SQL.

Tomando en cuenta los problemas encontrados en los modelos existentes para el idioma español, se busca generar un modelo versátil. Esto significa que el modelo debe ser capaz de procesar de manera efectiva una amplia variedad de preguntas y consultas complejas en diferentes contextos y dominios. Para lograr esto, se debe trabajar en la creación de un modelo lo suficientemente flexible y adaptable para manejar diferentes tipos de consultas y contextos.

Los usuarios potenciales de esta herramienta son empresas, organizaciones e individuos que requieren el acceso a información contenida en bases de datos. Los beneficios a los que podrán acceder incluyen el aumento de la productividad a la hora de trabajar con bases de datos SQL, gracias a la facilidad y rapidez con la que podrán realizar consultas a una base de datos relacional.

Buscamos proveer una documentación clara y completa del modelo propuesto y su implementación. Esto incluirá la creación del documento actual y los próximos a realizar durante el proceso del trabajo terminal, igualmente una documentación técnica que permita a otros desarrolladores e investigadores entender y replicar el modelo propuesto.

En resumen, el desarrollo de este trabajo presenta un gran desafío debido a la necesidad de diseñar e implementar un modelo de PLN y ML preciso y confiable. Además la elaboración de este proyecto formará parte de los avances para la investigación de esta tarea en el idioma español, lo que requiere de una rigurosa evaluación y prueba del prototipo resultante. En cuanto a la viabilidad del proyecto, se cuenta con el tiempo y recursos necesarios para llevar a cabo las diferentes etapas del desarrollo y evaluación del modelo.

4. Productos o resultados esperados

Se entregará un modelo capaz de transformar consultas formuladas por usuarios en lenguaje natural (español) a su equivalente en SQL. El usuario deberá introducir una consulta al modelo sobre un dato contenido en la base de datos con la que se esté trabajando, de la cual deberá introducir su esquema, y éste devolverá la consulta en SQL a la pregunta previamente formulada.

Debido a la complejidad de la tarea, el alcance del modelo estará conformado pero no limitado a consultas de tipo “SELECT-FROM-WHERE”. Se espera extender la funcionalidad de las consultas con estructuras más complejas, como las que se generan al incluir funciones de agregado, por ejemplo, “COUNT”, “MAX”, “MIN”, entre otras. El realizar consultas más complejas será determinado una vez que se genere y analice el dataset resultante.

Para el entrenamiento de nuestro modelo necesitaremos un conjunto de datos que contenga preguntas en español sobre información contenida en tablas de datos y su equivalente como consultas en SQL. Existen varios conjuntos de datos que buscan aportar esta tarea, pero la gran mayoría de ellos están en inglés [9]. Un ejemplo de un conjunto de datos con contenido en español es ‘MultiSpider’, el cual representa el conjunto de datos de texto a SQL multilingüe más grande que existe hasta la fecha y cubre siete idiomas [12]; por lo cual un producto del presente proyecto será un conjunto de datos con las características mencionadas, el cual tomará como punto de partida los datos proporcionados por MultiSpider, y extenderá al mismo por medio de la traducción de conjuntos de datos ya existentes.

Adicionalmente, se crearán un conjunto de herramientas auxiliares que nos ayudarán a generar el dataset con el que se entrenará el modelo y, de esta manera, crear el dataset se convertirá en un proceso más eficiente, ayudándonos a cumplir con los tiempos y objetivos previstos.

5. Metodología

Optamos por aplicar la metodología CRISP-ML(Q) debido a sus beneficios, ya que es muy recomendable para proyectos de ML [13]. La primera ventaja que observamos es que nos da un marco específico para el desarrollo del proyecto, lo que nos ayuda a evitar errores típicos y asegura una adecuada y efectiva implementación del modelo. Además, esta metodología nos permite trabajar de forma iterativa, lo que nos ayudará a mejorar continuamente nuestro modelo a medida que acumulamos nueva información o retroalimentación. Finalmente, esta metodología destaca el valor de las pruebas y validaciones adecuadas, asegurándonos de que nuestro modelo sea preciso y confiable antes de ponerlo en producción.

A continuación se presenta una breve descripción de las actividades que se realizarán en cada fase de la metodología:

- Fase 1: Comprender el problema. El objetivo de esta fase es identificar el problema que debe resolverse, que es el desarrollo de un modelo que convierta el lenguaje natural en SQL y permita a los usuarios ejecutar consultas de bases de datos sofisticadas. Se determinarán los objetivos y especificaciones de la herramienta, así como las dificultades o limitaciones tecnológicas que puedan surgir durante el desarrollo.

- Fase 2: Exploración de datos. Los datos necesarios para crear el modelo se recopilarán durante esta fase y se evaluarán. Se examinarán en profundidad las bases de datos accesibles, su arquitectura interna y las consultas más frecuentes que se realizan sobre ellas. Además, se recopilarán ejemplos de consultas en lenguaje natural para el entrenamiento del modelo de aprendizaje automático.

- Fase 3: Preparación de datos. Los datos necesarios para entrenar el modelo de aprendizaje automático se prepararán en esta etapa. Para que las consultas de lenguaje natural de ejemplo se puedan usar como entrada del modelo, se editarán y ordenarán.

- Fase 4: Modelamiento. El modelo de aprendizaje automático necesario para traducir consultas en lenguaje natural a SQL se desarrollará durante esta fase. El modelo se entrenará utilizando técnicas de procesamiento de lenguaje natural y aprendizaje automático utilizando las muestras de datos preparadas del paso anterior. Se evaluará el rendimiento del modelo y se realizarán los cambios necesarios.

- Fase 5: Evaluación. Esta etapa incluirá la evaluación del modelo terminado. La corrección y la fiabilidad del modelo para convertir consultas en lenguaje natural a SQL se probarán exhaustivamente. También se evaluará la usabilidad del modelo resultante.

- Fase 6: Implementación. Se realizaron las últimas pruebas antes de ser implementada y presentada ante el comité de evaluación.

6. Cronograma

Miranda Chávez Víctor Ulises

Tarea	Agos	Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abril	Mayo	Jun
Evaluación y selección de datasets											
Recopilar el dataset											
Análisis profundo de datasets seleccionados											
Preprocesamiento del dataset para su posterior tratamiento de traducción											
Traducción automática											
Etiquetar traducciones obtenidas											
Corrección de traducciones incorrectas											
Evaluación TT1											
Selección de algoritmos de aprendizaje automático											
Diseño de arquitectura de modelo											
Creación del modelo											
Entrenamiento del modelo											
Validar el rendimiento del modelo											
Reentrenar modelo en base a los descubrimientos hechos en la fase anterior											
Evaluar el modelo en condiciones de producción											

Documentar el proceso de elaboración del modelo de aprendizaje automático y los experimentos											
Evaluación TT2											

Nicolás Hernández Adair

Tarea	Agos	Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abril	Mayo	Jun
Evaluación y selección de datasets											
Recopilar el dataset											
Análisis profundo de datasets seleccionados											
Preprocesamiento del dataset para su posterior tratamiento de traducción											
Traducción automática											
Etiquetar traducciones obtenidas											
Corrección de traducciones incorrectas											
Diseño de tubería de datos											
Selección de algoritmos de aprendizaje automático											
Creación de la tubería de datos											
Entrenamiento del modelo											
Validar el rendimiento del modelo											
Reentrenar modelo											

en base a los descubrimientos hechos en la fase anterior											
Evaluar el modelo en condiciones de producción											
Documentar el proceso de elaboración del modelo de aprendizaje automático y los experimentos											
Evaluación TT2											

Pérez Sánchez Ives Lancelote

Tarea	Agos	Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abril	Mayo	Jun
Evaluación y selección de datasets											
Recopilar el dataset											
Análisis profundo de datasets seleccionados											
Preprocesamiento del dataset para su posterior tratamiento de traducción											
Traducción automática											
Etiquetar traducciones obtenidas											
Corrección de traducciones incorrectas											
Evaluación TT1											
Diseño de tubería de datos											
Selección de algoritmos de aprendizaje automático											

Creación de la tubería de datos											
Entrenamiento del modelo											
Validar el rendimiento del modelo											
Reentrenar modelo en base a los descubrimientos hechos en la fase anterior											
Evaluar el modelo en condiciones de producción											
Documentar el proceso de elaboración del modelo de aprendizaje automático y los experimentos											
Evaluación TT2											

Rubio Lopez Zury Yael

Tarea	Agos	Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abril	Mayo	Jun
Evaluación y selección de datasets											
Recopilar el dataset											
Análisis profundo de datasets seleccionados											
Preprocesamiento del dataset para su posterior tratamiento de traducción											
Traducción automática											
Etiquetar traducciones obtenidas											
Corrección de											

traducciones incorrectas											
Selección de algoritmos de aprendizaje automático											
Creación del modelo											
Entrenamiento del modelo											
Validar el rendimiento del modelo											
Reentrenar modelo en base a los descubrimientos hechos en la fase anterior											
Evaluar el modelo en condiciones de producción											
Documentar el proceso de elaboración del modelo de aprendizaje automático y los experimentos											
Evaluación TT2											

7. Referencias

[1] D. D. Chamberlin, “Early History of SQL", *IEEE Annals Hist. Comput.*, vol. 34, núm. 4, pp. 78–82, 2012, doi: 10.1109/MAHC.2012.61.

[2] W. A. Woods, “Progress in natural language understanding: an application to lunar geology”, en *Proceedings of the June 4-8, 1973, national computer conference and exposition on - AFIPS ’73*, New York, New York: ACM Press, 1973, p. 441. doi: 10.1145/1499586.1499695.

[3] D. Alconada, “AlcoNQL : Herramienta de consulta SQL por medio de lenguaje natural”, Universitat Oberta de Catalunya, 2013. [En línea]. Disponible en: https://openaccess.uoc.edu/bitstream/10609/18849/6/diealcoTFC0113memoria.pdf

[4] M. Bonilla, “Traductor de consultas del lenguaje natural a SQL”, Universidad Central “Marta Abreu” de Las Villas, 2011. [En línea]. Disponible en: https://dspace.uclv.edu.cu/bitstream/handle/123456789/9225/%5b06-28%5d%20Trabajo%20de%20Diploma_Marlen%20FINAL%20OK%20.pdf?sequence=1&isAllowed=y

[5] R. Cai, J. Yuan, B. Xu, y Z. Hao, “SADGA: Structure-Aware Dual Graph Aggregation Network for Text-to-SQL”. arXiv, el 17 de enero de 2022. Consultado: el 24 de abril de 2023. [En línea]. Disponible en: http://arxiv.org/abs/2111.00653

[6] Y. Mellah, A. Rhouati, E. H. Ettifouri, T. Bouchentouf, y M. G. Belkasmi, “SQL Generation from Natural Language: A Sequence-to-Sequence Model Powered by the Transformers Architecture and Association Rules”, *Journal of Computer Science*, vol. 17, núm. 5, pp. 480–489, may 2021, doi: 10.3844/jcssp.2021.480.489.

[7] D. Reinsel, J. Gantz y J. Rydning. (Noviembre, 2018). The Digitization of the World from Edge to Core, IDC Analyze The Future, [En línea], Disponible en: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[8] A. Bautista, “Traducción de consultas del lenguaje natural español a SQL que involucran agrupamiento”, 2014.

[9] B. Qin et al., “A Survey on Text-to-SQL Parsing: Concepts, Methods, and Future Directions”. arXiv, el 29 de agosto de 2022. Consultado: el 12 de abril de 2023. [En línea]. Disponible en: http://arxiv.org/abs/2208.13629

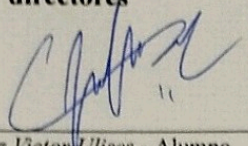
[10]L. Ouyang et al., “Training language models to follow instructions with human feedback”. arXiv, el 4 de marzo de 2022. doi: 10.48550/arXiv.2203.02155.

[11]A. Liu, X. Hu, L. Wen, y P. S. Yu, “A comprehensive evaluation of ChatGPT’s zero-shot Text-to-SQL capability”. arXiv, el 11 de marzo de 2023. Consultado: el 24 de abril de 2023. [En línea]. Disponible en: http://arxiv.org/abs/2303.13547

[12] L. Dou et al., “MultiSpider: Towards Benchmarking Multilingual Text-to-SQL Semantic Parsing”. arXiv, el 27 de diciembre de 2022. doi: 10.48550/arXiv.2212.13492.

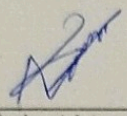
[13] S. Studer *et al.*, “Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology”. arXiv, el 24 de febrero de 2021. doi: 10.48550/arXiv.2003.05155.

8. Alumnos y directores

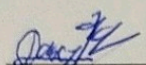
Firma: 

Miranda Chávez Victor Ulises.- Alumno
de la carrera de Ing. en Inteligencia Artificial en ESCOM, Especialidad en Inteligencia Artificial, Boleta: 2021630456, Tel. 5573338323, ulimirandachavez@hotmail.com.

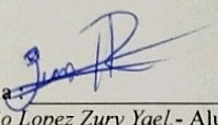
CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Artículo 11 Fracc. V y Artículos 108, 113 y 117 de la Ley Federal de Transparencia y Acceso a la Información Pública.
PARTES CONFIDENCIALES: Número de boleta y teléfono.

Firma : 

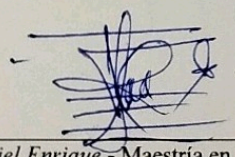
Nicolas Hernández Adair.- Alumno
de la carrera de Ing. en Inteligencia Artificial en ESCOM, Especialidad en Inteligencia Artificial, Boleta: 2021630570, Tel. 5519371732, adairnicolas2@gmail.com

Firma : 

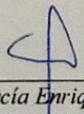
Pérez Sánchez Ives Lancelote.- Alumno
de la carrera de Ing. en Inteligencia Artificial en ESCOM, Especialidad en Inteligencia Artificial, Boleta: 2021630530 , Tel .5539973428 , lancelote.ps@gmail.com.

Firma : 

Rubio Lopez Zury Yael.- Alumno
de la carrera de Ing. en Inteligencia Artificial en ESCOM, Especialidad en Inteligencia Artificial, Boleta: 2021630638 , Tel. 5546798950, Zrubiol1700@alumno.ipn.mx

Firma : 

Flores Estrada Ituriel Enrique.- Maestría en Tecnologías de la Información, Inteligencia de Negocios y Análisis de Datos por la Universidad de Carnegie Mellon. Profesor de ESCOM. Áreas de Interés: Inteligencia Artificial, Procesamiento de lenguaje natural, Redes Neuronales, Minería de datos. Email: iflorese@ipn.mx

Firma : 

Carmona García Enrique Alfonso.- Maestría en Sistemas Computacionales Móviles por el IPN. Profesor de ESCOM. Áreas de interés: Inteligencia artificial, Procesamiento del Lenguaje Natural, Aprendizaje Automático. Email. eacarmona860920@gmail.com