Questioning your maintenance data and get accurate answers

Ziyu Zhao

PhD Student

Theme 1
2023-03-17

Virtual - Researchers Catch-up host online from Curtin University

Can we retrieve information from a structured database by asking questions like ChatGPT?

What technical building blocks are necessary to enable a natural language interface to a structured database, be it relational or graph-based?
When you access and try understanding structured maintenance data, for example, a graph database, you may question this database using your everyday language. Using everyday language or what we call 'natural language' to question a database will not give you an accurate answer as it is not executable in a database.
Ziyu's presentation will take you through how she has developed a baseline using a technique called 'Text-to-Cypher: A Baseline Natural Language Interface for Property Graph Databases'.

What is a Cypher query, you ask?

Well, Cypher is Neo4j's query language developed in 2011. It is one of the NoSQL query languages. Neo4j tops the DB engine ranking of Graph DBMS (https://db-engines.com/en/ranking). Different storage and query paradigms have different characteristics, especially in terms of representation and query expressivity, and scalability. While SQL, the first database query language, comprises a sophisticated data structure and expressive query language, NoSQL trades schema and query expressivity for scalability. Her proposed baseline has enormous potential in the maintenance domain as it has been trained on her proposed parallel corpus, which spans 155 domains.

Ziyu's 'Text-to-Cypher' work features two T5-style baseline models. T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks. The availability of such corpus and baselines can help develop and evaluate new transformer-based methods in understanding the text-to-NoSQL query language generation problem to support maintainers without coding experience and domain knowledge to access structured database data in natural language.