# Transforming Text-to-SQL Datasets into Closed-Domain NER Benchmark

Zikun Fu[1], Chen Yang[2], Heidar Davoudi[1], Ken Pu[1]
[1]Ontario Tech University
[2]Northeastern University
Zikun.Fu@ontariotechu.net, Yang.Chen9@northeastern.edu, Heidar.Davoudi@ontariotechu.ca, Ken.Pu@ontariotechu.ca

## 1 Introduction

Closed-domain named entity recognition (CD-NER) involves extracting entities from a fixed set of values in a database, which can be very large. Unlike general-purpose NER, CD-NER leverages domain-specific properties to improve extraction. However, the lack of high-quality benchmark datasets limits progress. CD-NER involves extracting entities from text that belong to elements of a structured database, such as table and column names, or partial tuple values. This domain-specific set can be extensive, containing billions of entities, which makes extraction challenging. Unlike open-domain NER, CD-NER requires handling specialized vocabulary, leveraging domain-specific context, and managing a large fixed pool of entities. The main challenge is accurately identifying entities within this closed set while dealing with the complexities of database size and specificity.

In text-to-SQL translation, benchmark datasets like BIRD[1] and Spider[2] have advanced research and set baselines. This paper addresses the lack of CD-NER resources by converting text-to-SQL benchmarks into CD-NER benchmarks. Our method creates a CD-NER benchmark using structured features from text-to-SQL datasets, providing a reliable evaluation resource for closed-domain entity extraction.

## 2 Methodology

To address the absence of high-quality CD-NER benchmarks, we transformed text-to-SQLs dataset into a CD-NER benchmark. BIRD[1] and Spider[2] offers natural language (NL) questions paired with SQL queries, along with database schemas and contents from over 175 diverse domains.

**Entity Extraction and Linking:** Our approach focused on extracting entities such as table names, column names, and values from SQL queries using SQLParse. After extraction, these entities were validated against the corresponding database schemas to ensure accuracy. To align the entities with the NL questions, we applied n-gram tokenization to the NL and matched the resulting tokens to entities based on similarity, which we measured using Levenshtein distance. Tokens that exceeded the similarity threshold were then linked to their respective database entities.

## 3 Future Work

Future work includes exploring alternative similarity metrics, incorporating human-annotated linking to improve accuracy, and extending entity types beyond tables, columns, and values. We also plan to fine-tune language models to establish baselines and compare them with state-of-the-art NER models on classification tasks.

## References

[1] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C.C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[2] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium, 2018. Association for Computational Linguistics. URL: https://aclanthology.org/D18-1425, doi:10.18653/v1/D18-1425.