Engineering Village™

1. **On the Importance of Building High-quality Training Datasets for Neural Code Search**
Sun, Z.; Li, L.; Liu, Y.; Du, X.; Li, L. **Source:** *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, 1609-20, 2022; **ISBN-13:** 978-1-4503-9221-1; **DOI:** 10.1145/3510003.3510160; **Conference:** 2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE), 25-27 May 2022, Pittsburgh, PA, USA; **Sponsor:** IEEE Computer Society;ACM SIGSOFT;IEEE Technical Council in Software Engineering; **Publisher:** IEEE, Piscataway, NJ, USA

Author affiliation:
Monash University, Melbourne, VIC, Australia
Tongji University, China

Abstract: The performance of neural code search is significantly influenced by the quality of the training data from which the neural models are derived. A large corpus of high-quality query and code pairs is demanded to establish a precise mapping from the natural language to the programming language. Due to the limited availability, most widely-used code search datasets are established with compromise, such as using code comments as a replacement of queries. Our empirical study on a famous code search dataset reveals that over one-third of its queries contain noises that make them deviate from natural user queries. Models trained through noisy data are faced with severe performance degradation when applied in real-world scenarios. To improve the dataset quality and make the queries of its samples semantically identical to real user queries is critical for the practical usability of neural code search. In this paper, we propose a data cleaning framework consisting of two subsequent filters: a rule-based syntactic filter and a model-based semantic filter. This is the first framework that applies semantic query cleaning to code search datasets. Experimentally, we evaluated the effectiveness of our framework on two widely-used code search models and three manually-annotated code retrieval benchmarks. Training the popular DeepCS model with the filtered dataset from our framework improves its performance by 19.2% MRR and 21.3% Answer@1, on average with the three validation benchmarks. (0 refs.) **Inspec controlled terms:** data handling - learning (artificial intelligence) - natural language processing - query processing

Uncontrolled terms: building high-quality training datasets - neural code search - training data - neural models - high-quality query - code pairs - code search datasets - code comments - famous code search dataset - natural user queries - dataset quality - model-based semantic filter - semantic query cleaning - code search models - manually-annotated code retrieval benchmarks - filtered dataset

Classification Code: C7250R Information retrieval techniques - C6130 Data handling techniques - C6180N Natural language processing - C1140Z Other topics in statistics

IPC Code: G06F7/00 - G06F17/20 - G06N20/00 - G06F16/00

Treatment: Practical (PRA)

Database: Inspec