Programming Language Representation with Semantic-level Structure

Anonymous Author(s)

ABSTRACT

Evaluating natural language model (NLP) on testset does with held-out accuracy is limited to show its quality assurance because the held-out datasets are often not comprehensive. While the behavioral testing over multiple general linguistic capabilities are employed, it relies on manually created test cases, and is still limited to measure its comprehensive performance for each linguistic capability. In this work, we introduce Fuzz-CHECKLIST, an NLP model testing methodology. Given a linguistic capability, The Fuzz-CHECKLISTfinds relevant testcases to test the linguistic capability from existing datasets as seed inputs, generates sufficient number of new test cases by fuzzing the seed inputs based on their context-free grammar (CFG). We illustrate the usefulness of the Fuzz-CHECKLISTby showing input diversity and identifying critical failures in state-of-the-art models for NLP task. In our experiment, we show that the Fuzz-CHECKLISTgenerates more test cases with higher diversity, and finds more bugs.

ACM Reference Format:
Anonymous Author(s). 2022. Programming Language Representation with Semantic-level Structure. In *Proceedings of ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2022)*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

NLP testing is crucial part for developing a reliable NLP model. In NLP, testing is mainly used to check the ML model's performance on hold-out set with respect to the accuracy of the model. The hold-out data refers to a portion of dataset that is held out of the datasets used for training machine learning models. Generally, the hold-out set is extracted via Train-Validation-Test split.

2 BACKGROUND

3 RELATED WORK

4 APPROACH

5 RESEARCH QUESTIONS FOR EVALUATION

6 EXPERIMENT

7 RESULT

REFERENCES