Programming Language Representation with Semantic-level Structure

Anonymous Author(s)

ABSTRACT

NLP (NLP) technique becomes one of the core techniques for developing text analytics applications. For developing the NLP applications, the applications are required to achieve high reliability before it goes to market. The trustworthiness of the prevalent NLP applications is obtained by measuring the accuracy of the applications on held-out dataset. However, evaluating NLP on testset does with held-out accuracy is limited to show its quality 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-CHECKLIST finds 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-CHECKLIST by showing input diversity and identifying critical failures in state-of-the-art models for NLP task. In our experiment, we show that the Fuzz-CHECKLIST generates 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

Software testing is the cruicial process when developing software. It evaluates an attribute or capability of the software and determines that it meets the requirements by examining the behavior of the software under test. Software testing in the early stage of the development finds bugs, and fixing them saves amount of costs ensuring that high quality of the software to users.

NLP testing is crucial part for developing a reliable NLP applications. For developing NLP applications, 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

ISSTA 2022, 18-22 July, 2022, Daejeon, South Korea
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn