Summary: Adaptive Testing and Debugging of NLP Models

BY JIGNESH SURYAKANT SATAM

Authors:
Marco Tulio Ribeiro, Microsoft Research, USA
Scott M. Lundberg, Microsoft Research, USA

Affiliations: Microsoft Research, University of Washington

Problem Addressed:
The paper addresses the problem of testing and debugging natural language processing (NLP) models. Testing and debugging are crucial steps in the development of NLP models to ensure their accuracy and reliability. However, traditional testing and debugging methods can be time-consuming and often require extensive manual effort. Moreover, the existing methods might only work for a very restrictive class of bugs in some cases. The paper proposes a new framework for adaptive testing and debugging of NLP models, which can improve the efficiency and accuracy of the testing and debugging process. The authors propose a new methodology called AdaTest which comprises the use of large-scale language models and human feedback to automate the writing of unit tests in the target model. Another key issue the authors claim to address is that their method fixes bugs without introducing any new bugs or issues.

Prior Work:
Prior work has focused on developing various methods for testing and debugging NLP models, including error frameworks for testing, error analysis, and crowdsourcing. However, these methods have their limitations, such as being inefficient or not capturing all possible errors. Some previous work has also explored using active learning methods to improve testing and debugging, but these methods are often computationally expensive and require large amounts of labeled data. Based on the author's research, existing approaches are not capable of identifying and fixing undesirable behaviors.

Unique Contributions of This Paper:
The paper proposes a new framework for adaptive testing and debugging of NLP models, which combines active learning and interpretability techniques. The framework is designed to automatically generate tests based on the model's weaknesses and use the generated tests to

improve the model's accuracy. The generated tests are also used to provide insights into the model's behavior, making it easier to identify and debug errors.

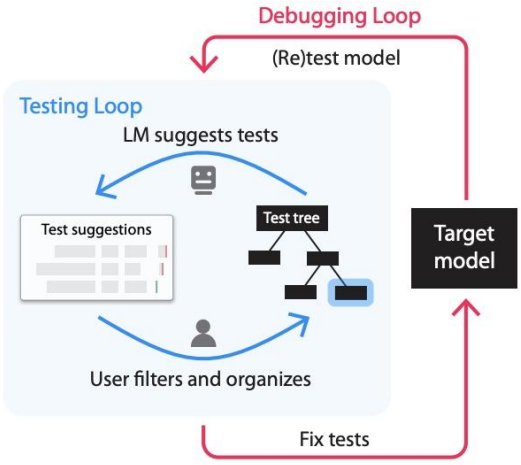

Fig 1. AdaTest Framework [1]

The framework is adaptive in that it can adjust the testing and debugging process based on the behavior of the model. The framework consists of two main components: (1) A Testing Loop that generates tests for the target NLP model and (2) a debugging loop that iteratively refines the target model based on failures of tests.

The testing loop encompasses a combination of test suggestion generation, assigning scores to suggestions by the model, and the user accepting and organizing them. To perform the evaluation of the model behavior, tests are organized within a test tree. In this tree, each node is a topic. Such trees are constructed by users which enhances the capability of the user and language model to focus topic-by-topic. Within this step, the user performs an organization step wherein they decide whether language model suggestions should be accepted or discarded. Next comes the debugging step where the user fixes the bugs from the testing step. Furthermore, the testing and debugging steps are repeatedly performed during which the AdaTest framework adapts without any intervention. Thus the debugging step produces a satisfactory model by transcending the boundaries of the given specifications. This is supported by different examples in the paper, one of which also demonstrates via examples how an existing test tree adapts to a new model even in cases where such trees are constructed based on different models.

Evaluation Method:

The proposed framework was evaluated on several NLP tasks, including text classification, sentiment analysis, and question answering. The results showed that the framework was able to identify and correct errors in the models, leading to improved accuracy. The paper also provides a detailed analysis of the framework's performance under different conditions and compares it

with existing testing and debugging methods. The authors evaluate their framework on two capabilities: bug finding and bug fixing. The bug-finding capacity is evaluated by using tools like Dynabench. Authors claim users consistently were able to find more bugs than other tools. In terms of fixing bugs, authors used tools like Checklist wherein they claim their framework's debugging step proved to be superior to the competitors. The results show that the framework is effective at identifying errors in the models and improving the quality of the models. The authors also compare their approach to traditional testing techniques and show that their approach is more effective at identifying errors. To support these evaluations, the paper further includes real-life case studies and examples which despite having faced issues show superior results with the AdaTest framework.

Citations:
As of April 2023 Google scholar indicates the lead author, Marco Tulio Ribeiro, has 16899 citations, while the second author, Scott M. Lundberg, has 18996 citations.

Conclusion:
The paper proposes a new framework for adaptive testing and debugging of NLP models, which combines active learning and interpretability techniques. To measure the usability of their framework, the authors performed tests with users of varied ages and backgrounds with no data science experience. Their experiments indicated that users with no computer science background using the Ada test framework found 10x more failures in the models. This means the proposed framework does not require users with prior training or knowledge to effectively use the framework. The components of users and language model make this framework superior as the user provides feedback that a language model lacks and on the other hand the language model provides computations that are not feasible by users to perform manually. The framework has the potential to save significant amounts of time and effort in the development of NLP models and can help researchers and practitioners to develop more accurate and reliable NLP models. The authors have demonstrated AdaTest's effectiveness on classification models like sentiment analysis and task detection, generation models, and per-token models (NER). Further showing the relevance of their work, their test framework has shown evidence of success on old, mature applications as well as new applications. The author's work immensely contributes towards effective, easy-to-use, and cheaper alternatives to existing techniques in the market.

References:
[1] https://aclanthology.org/2022.acl-long.230.pdf