Master Thesis Defense Ravikiran Bhat

Date: Wednesday, 08.05.2019
Time: 10:30
Room: C219

Title: Application of Model Interpretability Techniques for Short Answer Grading

Abstract:

Automated scoring of short answer type questions via Natural Language Processing (NLP) is a field that has a substantial body of associated research in order to develop methods to reduce workload of manual graders. However, procedures that associates a reasoning behind the automated score to enhance the reliability of the assigned grade is a task that has received little attention in literature. This work focuses on applying and evaluating different strategies of model interpretability to the task of automatic short answer grading as a means of rationalizing the predicted score to the human graders. Pre investigations were carried out on two state-of-the-art model-agnostic interpretation techniques, namely Local interpretable model-agnostic explanations (LIME) and SHapley Additive exPlanations (SHAP). The influence of change in feature extractors and machine learning models on LIME was analyzed, and a quantitative comparison of the interpretation levels from LIME and SHAP was done for different case studies. The results showed that LIME and SHAP tend to give better explanations when answers are restricted to one sentence length, and do not have a large linguistic diversity. Furthermore, LIME's restriction of highlighting only unigram keywords as part of the explanation, and SHAP's exponential time requirement to compute feature attributions, made them unsuitable for our application. As a means of overcoming these challenges, we develop a solution that is model agnostic, is capable of highlights phrases (bigrams and trigrams) in the explanation, and improves the autograding and generation of explanations for new data by involving a human oracle both during initial annotation of data, and later in providing feedback for the predicted scores and explanations. A prototype web based GUI tool was developed to integrate the proposed solution. Tests run with the proposed solution showed a significant improvement in the level of interpretation of the highlighted explanation when compared against the pre investigation results for LIME and SHAP.
