A hybrid method of feature extraction and naive bayes classification for splitting identifiers

ABSTRACT

Nowadays, integrating natural language processing techniques on software systems has caught many researchers' attentions. Such integration can be represented by analyzing the morphology of the source code in order to gain meaningful information. Feature location is the process of identifying specific portions of the source code. One of the most important information lies on such source code is the identifiers (e.g. Student). Unlike the traditional text processing,the identifiers in the source code is formed as multi-word such as 'Employee-Name'. Such multi-words are not divided using white space, instead it can be formed using special characters (e.g. Employee_ID), CamelCase (e.g. EmployeeName) or using abbreviations (e.g. EmpNm). This makes the process of extracting such identifiers more challenging. Several approaches have been performed to resolve the problem of splitting multi-word identifiers. However, there is still room for improvement in terms of accuracy. Such improvement can be represented by utilizing more robust features that have the ability to analyses the morphology of identifiers. Therefore, this study aims to propose a hybrid method of feature extraction and Naïve Bayes classifier in order to separate multi-word identifiers within source code. The dataset that has been used in this study is a benchmark-annotated data that contains large number of Java codes. Multiple experiments have been conducted in order to evaluate the proposed features independently and with combinations. Results shown that the combination of all features have obtained the best accuracy by achieving 64.7% of f-measure. Such finding implies the usefulness of the proposed features in terms of discriminating multi-word identifiers.

Keyword: Feature location; Split identifiers; Feature extraction; Naive bayes; Source code