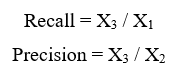
# **3.0 MODEL EVALUATION**

The model performance is measured through the calculation of precision and recall for both real-word and non-word errors detection. These metrics are chosen as they are widely used in many spell-checker evaluations studies (Van Zaanen and Van Huyssteen, 2003; Starlander and Popescu-Belis, 2002). Recall serves to indicate the model’s sensitivity in detecting potential errors while precision indicates the accuracy of the model’s detection. In other words, higher recall may indicate that the model is highly sensitive while high precision indicates that the detection made by the model is accurate. The calculation of recall and precision require few parameters, as shown in below formulae. X1 is the total number of errors in the testing corpus while X2 is the total number of detections. Finally, X3 is total number of correct detections.



# **4.0 RESULT AND DISCUSSION**

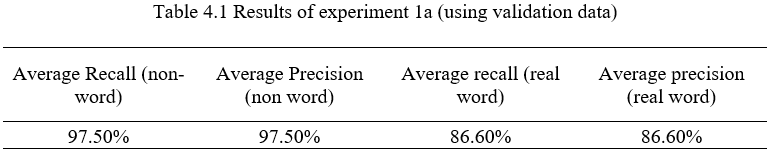
The table below summarizes the experiments conducted for the purpose of the evaluation.



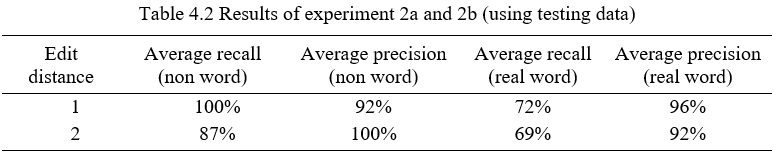
Figure 4.1 Result for different variation of testing

In order to ensure that the model is adequately trained, the validation corpus is first used to check the model performance where it was tested in two different experiments. In the first experiment, the validation corpus was submitted to the spell-checker without any modification to establish a baseline for the number of words that are unknown to the system. This step was taken as it is reasoned that the 60% of the dataset may not capture sufficient amount of words. By establishing the number of unrecognized words in the validation corpus, the calculation of precision and recall in the next experiment can be done more accurately and will not be confused with the synthetically introduced errors.

Following the establishment of the baseline, 26 errors (13 real word, 13 non-word) are introduced into the validation corpus for second testing. the detection for non-word are higher than detection of real words among all the testers. Furthermore, the results also suggest that the model was able to detect approximately 98% of the non-word errors and 98% of those detection were correct. Such performance was fairly good, which may also mean that the sampling technique used was effective in ensuring wide range of words from different chapters are captured in the dictionary. However, the performance for real word was much lower, with only 87% for both precision and recall.



Based on the model performance, the system is deemed complete and ready for the subsequent testing. Therefore, the validation set is then processed to tokens and bi-grams and subsequently added into the dictionary. Similar to the steps above, the remaining 20% of the data is used to test the re-trained model that contains the initial 60% of training data and 20% of validation data. The table below summarizes the results of the testing using the remaining 20% of unseen data. Using the corpus, errors with 1 and 2 edit distance were introduced and performance of the model on different edit distance is evaluated.



Based on the results above, it can be seen that the overall performance of the detection of non-words is still better than the performance of real word detection. This is expected as the detection of non-word errors rely purely on the words’ presence in the stored dictionary. On the other hand, the detection of real word error is less straightforward. As shown in the results, the model was only able to detect 60 to 70% of the real word errors. This may suggest that the model’s sensitivity to errors and threshold may need further adjustment. Nonetheless, majority of the detection that were made by the model are accurate, as the precision is approximately 90% for both types of edit distance.