1

This report has handled the problem of blend word detection using GED, LED and Jaro-Winkler similarity. Firstly, it proposed a data pre-processing method to remove non-standard words. Next, it introduced methodologies and evaluation metrics used concisely. In addition, some experiments has been simulated to show the performance of different string matching algorithms evaluated by precision and recall clearly. Finally, the author studied some relevant literature and analyzed the feasibility of methodologies used critically – LED,GED and Jaro-Winkler similarity are powerful for approximate strings matching problem, but they are not suitable for blend word detection.

2

There are three strengths in this report.

Firstly, the report is overall very well written and clearly organized. It is divided eight parts and help reader to understand its work easily.

Secondly, this report has used three approximate strings matching algorithms we has learnt in subject to resolve the problem of blend word detection suitably.

Thirdly, the simulation has been designed and implement very good by comparing the performance of GED, LED and Jaro-Winker similarity because of clear tables and results.

3

However, there are also two drawbacks in this report and should be improved.

Initially, the explanations of implement details are not 100% clear. For instance, as a reader, I am confused about why the values of Jaro-Winker similarity have been selected in part 5.3.

Next, although author has not mentioned directly, I suppose that the computing complexity of this implementation are high. It will be better if the author can take some actions to reduce the complexity and speed up the whole implementation rather than remove non-standard words in candidates.

1

This report has handled the problem of blend word detection by finding whether the left and right substrings exist in dictionary. There are two parts of this report proposed detection system – one is string splitter and another is substring searching. In string splitter subsystem, candidates has been split to left and right substring. In string searching subsystem, author has used N-gram algorithm and global edit distance to judge whether substrings are in the dictionary and then detect whether the candidates is blend words or not.

2

There are two strengths in this report.

Firstly, the idea of word split and substring matching is very good and the simulation of implementation is well done using N-gram and global edit distance. For example, author give the parameter of N-gram and global edit distance to detect blend words from candidates.

Secondly, in other to reduce the time of simulation, author preprocessed the dataset by selecting one candidate from 16 candidates, which will help author to speed up by 16 times.

3

However, there are also three drawbacks in this report and should be improved.

Initially, the organization of this report is not very formal. For instance, section 5 just have one point but the author has written subsection 5.1. in addition, this report does not included relevant literature.

Next, the implement details are not 100% clear and lack of explanations and proper reference of many terminologies. For example, the definition of N-gram and global edit distance is not introduced which improves the difficult for reader to understand report, especially those who have not familiar with string matching algorithm and blend word detection.

Finally, the evaluation result is not trustworthy because it shows the precision of Jaccard similarity is 40% but the recall is only 3%. I can not understand the result.

This report has handled the problem of blend word detection using prefix tree, suffix tree, Global Edit Distance and Jaro-Winkler similarity. Firstly, the author mined many patterns of blend words using diagrams and charts clearly. Secondly, it proposed a data pre-processing method to remove non-standard words such as the word with less than three letters or more than 14 letters. Next, it introduced methodologies and evaluation metrics used concisely and proposed a blend word detection algorithm based on GED and Jaro-Winker similarity. In addition, some experiments has been simulated to show the performance of the algorithms proposed by author evaluated by precision and recall clearly. Finally, the author studied some relevant literature and analyzed the feasibility of methodologies used critically–GED and Jaro-Winkler similarity are powerful for approximate strings matching problem, but they are not suitable for blend word detection.

2

There are three strengths in this report.

Firstly, the report is overall very well written and clearly organized. It is divided eight parts and help reader to understand its work easily.

Secondly, a number of patterns of blend words are mined and proposed three feasible hypotheses.

Thirdly, this report has used two approximate strings matching algorithms we has learnt in subject and used an effective data structure for search prefix and suffix to resolve the problem of blend word detection suitably. The algorithm proposed by the author is clear and efficient to detect whether a word is blend word or not.

3

However, there are also two drawbacks in this report and should be improved.

Initially, the explanations of implement details are not 100% clear. As the reason of space limitation, the author have not explained why the values of parameters have been selected in algorithm.

In addition, the author used too much space to introduce the patterns of blend word and less space to do more than one detection methods. I hope the author can pay more attention on the methodology and experiments.