.RDR Threshold Tuning Report

July 13, 2023

About rdr files>
->.Dict: full dictionary, contains tag of all words, if a particular word has more than one tag, the most occurrence is included

->.sDict: short dictionary , contains most occurrence of a words given that a word occurred more than once.

In dictionary , three more default word/tag are there i.e for Numerical, Lower Case letters, Upper Case letters, and unknown, they store tag frequency and one with most occurrence is tagged.

->.RAW: raw corpus from gold standard corpus(each word with slash'\' and tag separated with white space) .

->.Init: Initialized corpus done with using .sDict.

Notes: Initial tagger has some suffix codes that needs to be commented.

Threshold:> default values is (3,2), this was set based on performance on the english data set.

Code snippet from rdr:>

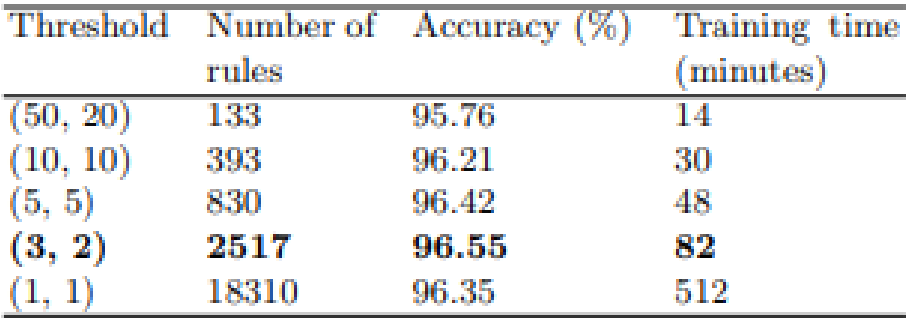
```
improvedThreshold = Threshold[0]
matchedThreshold = Threshold[1]
```

About threshold from published papers:>
.The threshold parameters were tuned on the English validation set. The best value pair (3, 2) was then used in all experiments for all languages.

We apply two threshold parameters: the first threshold is to find exception rules at the layer-2 exception structure, such as rules (4), (10) and (11) in Figure 1, while the second threshold is to find rules for higher exception layers.

) The rule must associate to a highest score value of subtracting B from A in comparison to other ones, where A and B are the numbers of the SO’s Objects which are correctly and incorrectly concluded by the rule respectively. (iii) And the highest value is not smaller than a given threshold. The SCRDR-learner applies two threshold parameters: first threshold is to choose exception rules at the layer-2 exception structure (e.g rules (3), (4) and (5) in figure 1), and second threshold is to select rules for higher exception layers.

Table 2. Pos tagging in accuracy of development data of our approach.



Threshold	Number of rules	Accuracy (%)	Training time (minutes)
(50, 20)	133	95.76	14
(10, 10)	393	96.21	30
(5, 5)	830	96.42	48
(3, 2)	**2517**	**96.55**	**82**
(1, 1)	18310	96.35	512

The above data was done on english data set and got threshold (3,2) as best result. Figure reference:> https://datquocnguyen.github.io/resources/CICLing2011.pdf

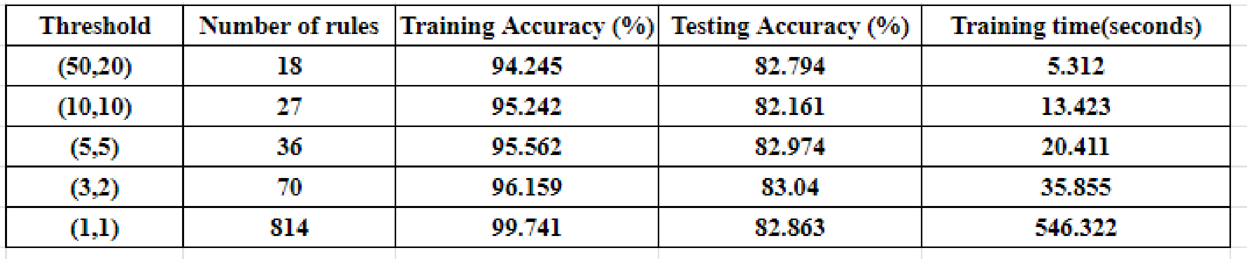
Below Figure is result i got from applying rdr on Tibetan data set on different threshold values.

Notes:

*Same threshold values has been used as like from the references.

*New parameter Testing accuracy has been added.

Number of words in Training File: 16248



Threshold	Number of rules	Training Accuracy (%)	Testing Accuracy (%)	Training time(seconds)
(50,20)	18	94.245	82.794	5.312
(10,10)	27	95.242	82.161	13.423
(5,5)	36	95.562	82.974	20.411
(3,2)	70	96.159	83.04	35.855
(1,1)	814	99.741	82.863	546.322

Training file name: TIB_train_maxmatched_tagged.txt
Testing file name: TIB_test_maxmatched_tagged.txt
Both files in src\word_segmentation_rules_generator\data\

Conclusion:> rdr model is overfitting the data, and needs to be able to generalize more. More study on rdr rules and dictionary need to be done.