Cells in yellow color are under 0.5 avg accy

5 methods are used to corrupt the structure

|  |  |
| --- | --- |
| Method | Description |
| Inhibit | Change loss value (cross entropy) to mean=0.5 ,var =0.05 for specific POS tag position ( NOUN , VERB, ADJ, ADJ+NOUN) |
| Random Permute Token | Random permutation of input token sequence |
| Random Merge Token | Merge token randomly with min2, max 3 |
| Merge <POS> Token | Only Merge surrounding token of specificy POS tag |
| Permute+Merge | Combine randome permute token and random merge |

Remark :

The top 10 merge count are different between seq 5,10 and 25,50

|  |  |  |
| --- | --- | --- |
| Random Merge Seq 5, 10 | Random Merge Seq 25, 50 | Merge Only ADJ |
| ('NOUN\_NOUN', 1299403),  ('NOUN\_ADJ', 794882),  ('VERB\_NOUN', 782272),  ('NOUN\_VERB', 684696),  ('ADJ\_NOUN', 501560),  ('NOUN\_NOUN\_NOUN', 433874),  ('NOUN\_NOUN\_ADJ', 388947),  ('NOUN\_VERB\_NOUN', 296444),  ('ADJ\_DET', 191582),  ('VERB\_NOUN\_NOUN', 148550) | ('ADJ\_ADJ\_ADJ', 4748367),  ('ADJ\_ADJ', 3725680),  ('NOUN\_NOUN', 3126343),  ('NOUN\_ADJ', 1781641),  ('VERB\_NOUN', 1771679),  ('NOUN\_VERB', 1507486),  ('NOUN\_NOUN\_NOUN', 1481755),  ('NOUN\_NOUN\_ADJ', 1118719),  ('NOUN\_VERB\_NOUN', 979389),  ('ADJ\_NOUN', 872810) | ('ADJ\_ADJ\_ADJ', 2131074),  ('ADJ\_ADJ', 1516601),  ('ADJ\_NOUN', 260508),  ('ADJ\_ADJ\_NOUN', 156022),  ('NOUN\_NOUN', 136952),  ('NOUN\_ADJ\_NOUN', 105874),  ('NOUN\_ADJ', 93044),  ('NOUN\_VERB', 87431),  ('DET\_ADJ', 79607),  ('NOUN\_NOUN\_NOUN', 56365) |

2. For random token permutation, a measurment between the original POS pattern and after permutation is add to track the difference

For seq 5, 10 , training and validation is rougthly 0.5-0.6

For seq 25 , 50 training and validation are different , training : 0.6 , validation : 0.7-0.8