Report for HW5

CS 519 Natural Language Processing

Prof. Liang Huang

Group member: Gune.Aditya, Hao.Liu, Ke.Huang

Part 1

1. We do the <unk> stuff because we want to get rid of noise. Word only appeared once will have a high chance to get similar small probabilities which would make it hard when we generate the word. The train.dict can help us in further decoding. Like in Q2.d
2. This is a right-expanding tree. The new child will only be add at right. And this VP’ make it easier for getting the probs. The alternatives can be NP-PP-ADVP instead of VP’, in this case the NP-PP-ADVP may only appear once, which is not what we want.

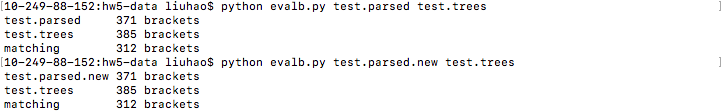
No this is not in CNF because it has a unary rule like NP -> DT. But in another case the unary rule is a preterminal of a lexical rule. So it depends on how we treat it.

1. We have 226 Binary rules, 47 unary rules and 326 lexical rules

Part 2

1. For there is a circle in one of the transaction in these two pcfgs. So it’s hard to count the exactly number. The strings can be really weird.
2. toy.parsed

c&d. we got no NONE for the result since the <unk> has been handled as an internal symbol.

e.

Since the <unk> has been handled internally. The two parsing accuracies are exactly the same and diff -b indicates that the trees are also the same. The reason why it is different from the original tree is because the training data may provide high probability transaction for the given test.txt which is not right or not what we want.