## Assignment 3

(Due: Nov 30, 6am)

**Question 1:** Let  $\mathbf{Q} \in \mathbb{R}^{N \times d}$  denote a set of N query vectors, which attend to M key and value vectors, denoted by matrices  $\mathbf{K} \in \mathbb{R}^{M \times d}$  and  $\mathbf{V} \in \mathbb{R}^{M \times c}$  respectively. For a query vector at position n, the softmax attention function computes the following quantity:

$$\operatorname{Attn}\left(\mathbf{q}_{n}, \mathbf{K}, \mathbf{V}\right) = \sum_{m=1}^{M} \frac{\exp\left(\mathbf{q}_{n}^{\top} \mathbf{k}_{m}\right)}{\sum_{m'=1}^{M} \exp\left(\mathbf{q}_{n}^{\top} \mathbf{k}_{m'}\right)} \mathbf{v}_{m}^{\top} := \mathbf{V}^{\top} \operatorname{softmax}(\mathbf{K} \mathbf{q}_{n}), \tag{1}$$

which is an average of the set of value vectors V weighted by normalized similarity between different queries and keys.

**Question 1a:** Please briefly explain what is the time and space complexity for the attention computation from query  $\mathbf{Q}$  to  $\mathbf{K}$ ,  $\mathbf{V}$ , using the big O notation.

```
Three steps:
```

- 1. Q^T \* K ----> R^(N\*M) time complexity: O(N\*M\*d), space complexity: O(N\*M)
- 2. softmax ----> R^(N\*M) time complexity: O(N\*M), space complexity: O(1)
- 3. V^T \* softmax() ----> R(N\*d) time complexity: O(N\*M\*d), space complexity: O(N\*d)

d and c are constant, then after combination:

Time complexity: O(M\*N) Space complexity: O(M\*N)

**Question 2:** Consider a probabilistic context-free grammar with the following rules (assume that S is the start symbol):

$S \rightarrow NP VP$	1.0
$VP \rightarrow Vt NP$	0.7
$VP \rightarrow VP PP$	0.3
$NP \rightarrow DT NN$	0.8
$NP \rightarrow NP PP$	0.2
$PP \rightarrow IN NP$	1.0
$Vi \rightarrow sleeps$	1.0
$Vt \rightarrow saw \\$	1.0
$NN \rightarrow man$	0.1
$NN \rightarrow woman$	0.1
$NN \rightarrow telescope$	0.3
$NN \rightarrow dog$	0.5
$DT \rightarrow the$	1.0
$IN \rightarrow with$	0.6
$IN \rightarrow in$	0.4

Question 2a: What's the <u>most likely parse tree</u> for the following sentence under this PCFG? Show CYK chart you developed below.

the man saw the woman with the dog

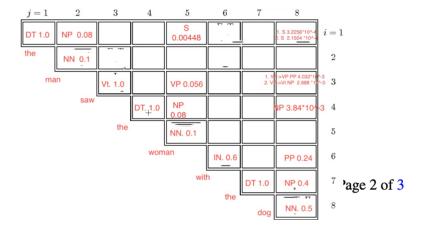
hints: we should put terminal rules into account. there are two possible parse trees.

**Question 2b:** What's the (marginal) probability of the following sentence under this PCFG?

the man saw the woman with the dog

5.376 \* 10^(-4)

hints: the combination of the probabilities of all possible trees



**Question 3:** A trigram language model is also often referred as a second-order Markov language model. It has the following form:

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i \mid X_{i-2} = x_{i-2}, X_{i-1} = x_{i-1})$$

**Question 3a:** Could you briefly explain the advantages and disadvantages of a high-order Markov language model when comparing with the second-order one? Advantages:

Higher-order Markov model considers more information has the ability to capture long-term dependency.

The longer the context on which we train the model, the more coherent the sentences.

## Disadvantages:

increased time and space complexity

sparsity: larger n-gram probability matrices are ridiculously sparse. In other words, sometimes using less context is a good thing, helping to generalize more for contexts that the model hasn't learned much about. We only "back off" to a lower-order n-gram if we have zero evidence for a

**Question 3b:** Could you give some examples in English where English grammar suggests that the second-order Markov assumption is clearly violated.

last night, I ate an apple. "ate" depends on "last night"

the book that describes the development of ... is very interesting. "is" depends on "book"