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:

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

which is an average of the set of value vectors $\mathbf{V}$ 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)

PCFG

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 most likely parse tree 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

	$j=1$	2	3	4	5	6	7	8	
	DT 1.0	NP 0.08			S 0.00448			1. S 3.2256*10^-4 2. S 2.1504 *10^-4	$i=1$
the		NN 0.1							2
man			Vt. 1.0		VP 0.056			1. VP->VP PP 4.032*10^-3 2. VP->Vt NP 2.688 *10^-3	3
saw				DT. 1.0	NP 0.08			NP 3.84*10^-3	4
the					NN. 0.1				5
woman						IN. 0.6		PP 0.24	6
with							DT 1.0	NP 0.4	7
the								NN. 0.5	8
dog									

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

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

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 higher-order n-gram

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"