Special Topic in Computers 2 (ELL 881)

Natural Language Processing (Minor 2)

Total Marks: 30 Date: 23.03.23

---

1. State whether the following statements are **True** or **False** (provide **brief explanation**):
2x5 = 10

1. The word **bank** is an example of homonymy. **True** (usage of 'bank' as financial institution vs. slope)
2. A very deep feedforward network suffers from the vanishing gradient problem. **True** (successive matrix multiplication will result in diminishing gradient if eigenvalues < 0)
3. Pointwise Mutual Information is biased towards frequent words. **False,** it is biased towards infrequent words
4. Beam search needs memory in quadratic order of the input length. **False,** linear complexity
5. LSTM solves the vanishing gradient problem via residual connections. **False,** uses gating mechanism to solve vanishing gradient

2. Given below is a cooccurrence matrix of six words. Use it to answer the following questions:

	Cat	**Dog**	**Bird**	**Fish**	**Mouse**	**Rabbit**
Cat	0	200	50	20	300	70
Dog	200	0	30	40	150	90
Bird	50	30	0	70	20	10
Fish	20	40	70	0	10	50
Mouse	300	150	20	10	0	80
Rabbit	70	90	10	50	80	0

a. What is the PMI between "cat" and "mouse"?

PMI(cat, mouse) = $\log_2$ N*count(cat, mouse)/(count(cat)*count(mouse))
= $\log_2$ C(RU)*count(cat, mouse)/Sum(row(cat))*Sum(row(mouse)) = −0.005637511

b. Calculate (a) with add-2 smoothing.

$\log_2$ C(RU+2*21)*(count(cat, mouse)+2)/(Sum(row(cat))+12)*(Sum(row(mouse))+12)
= −0.003408211

2+2 = 4

3. Mention two disadvantages of count-based methods to generate word embeddings that are solved using direct-prediction-based methods. How would you deal with the problem of out-of-vocabulary words while using word-embedding methods like Word2Vec/GloVe?
Disadvantages of count-based methods:
a. Biased towards larger counts, b. Primarily used for word similarity

Out-of-vocab solution: Subword tokenization

2+2 = 4

4. Assume an LSTM with a hidden state dimension of 128. It is processing a sequence of 10 vectors, each of dimension 128 (assume that there is no input embedding projection).

a. What is the total number of parameters in this LSTM?

8x(128x128) + 4*128 = 131584

b. How many matrix multiplications are needed to perform this processing?

8 for each token, 80 total

2+4 = 6

5. State the advantages and disadvantages of a vanilla LSTM vs. a vanilla RNN with attention (at least 2 for each) in the context of machine translation. What are the issues (at least 2) with a CNN to design a language model? 4+2 = 6

Advantage of vanilla LSTM over vanilla RNN with attention:

1) Better capture long-range dependencies in encoder as well as decoder
2) Better at resolving exploding as well as vanishing gradient problem

Disadvantage of vanilla LSTM over vanilla RNN with attention:

1) Encoder-decoder bottleneck
2) Alignment between source and target tokens not explainable