Quiz: Information Extraction with NLP

TOTAL POINTS 10

1. Which of the following is not true about BERT's inner word representations? **0 / 1 point**

- ○ Each unique word can have exactly one vector representation
- ○ The representation of a word depends on the words around it
- ○ Words which are similar in meaning are typically close as vector
- ◉ None of the above

! **Incorrect**

2. True or False: the start and end vectors are fixed throughout training **0 / 1 point**

- ◉ True
- ○ False

! **Incorrect**

3. Which of the following is a difference between BERT and LSTM models? **1 / 1 point**

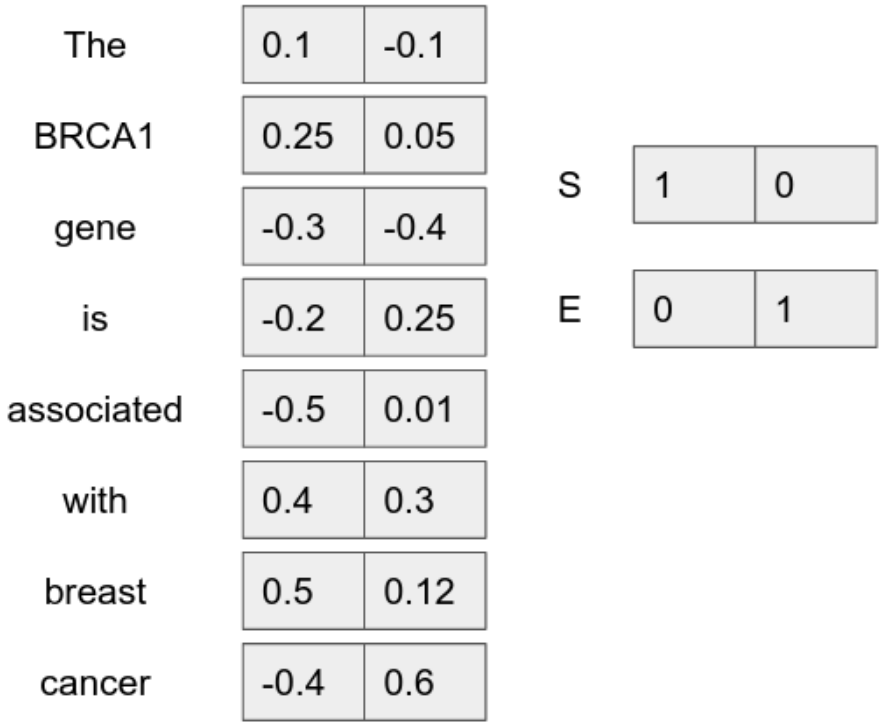
- ○ BERT can be trained on multiple languages, while LSTMs cannot

- ○ BERT is trained using backpropagation while LSTMs are not
- ◉ BERT takes entire sequences as input, while LSTM models process words one by one
- ○ BERT uses regular word vectors, while LSTMs use contextualized word vectors

✓ **Correct**

Explanation: A major difference between BERT and LSTMs is that BERT process an entire sequence of input, while LSTMs only in words one by one. This enables greater parallelization and results in better training among a variety of tasks.

4. Given the following word vectors and start and end vectors, determine the start and end of the sequence of interest. **1 / 1 point**



- ○ start: The, end: cancer

- start: gene, end: associated
- start: with, end: gene
- start: breast, end: cancer

Correct

Explanation: The start scores are [the: 0.1, BRCA1: 0.25, gene: -0.3, is: -0.2, associated: -0.5, with: 0.4, breast: 0.5, cancer: -0.4] while the end scores are [the: -0.1, BRCA1: 0.05, gene: -0.5, is: 0.25, associated: 0.01, with: 0.03, breast: 0.12, cancer: 0.6]. The the pair of words which maximize the sum of start and end scores and where the start token is before the end token is start: breast and end: cancer, which corresponds to answer D.

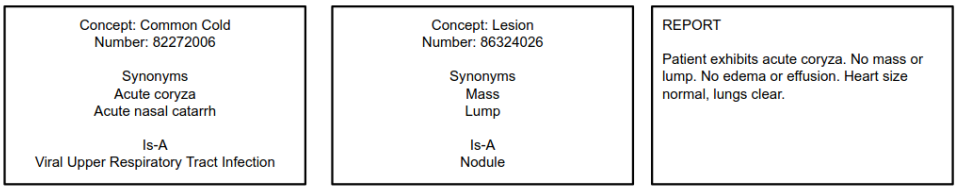
5. You find that a radiology report mentions “edema”. Which of the following can you immediately conclude? **1 / 1 point**

- The x-ray contains edema
- The x-ray contains pneumonia
- The x-ray does not contain edema
- None of the above

Correct

Explanation: The correct answer is none of the above, since we are missing negation information about the mention of edema. A is false since the mention could have been preceded by “no”, so we can’t conclude that it is present. For similar reasons, we can’t say that it’s not necessarily not present. Since nothing is mentioned about pneumonia, we can’t say whether its present or not. Therefore the correct answer is none of the above.

6. Use the following entry in SNOMED CT to help determine the positive labels for this x-ray report. **1 / 1 point**



- common cold: 0, lesion: 0
- common cold: 0, lesion: 1
- common cold: 1, lesion: 1
- common cold: 1, lesion: 0

Correct

Explanation: First, report mentions acute coryza, which we can see from the SNOMED CT cards is a synonym for the common cold. Since it is a positive mention, we can safely say that the patient has a common cold. However, while mass is synonymous mass and lump, which are mentioned, they are negated. Therefore the label should be common cold: 1, lesion: 0.

7. Let's see why F1 is used instead of the regular mean of precision and recall. Let's say the mean of precision and recall is at least 0.75. Which of the following could be the true value of the precision? **0 / 1 point**

- 0.75
- 0.5
- Both

- (●) Neither

! **Incorrect**

8. Now let's say F1 score is at least 0.75. Now which of the following values of precision are possible? **0 / 1 point**

- ( ) 0.75
- (●) 0.5
- ( ) Both
- ( ) Neither

! **Incorrect**

9. Compute the F1 score for pneumonia and mass separately based on the following retrieved labels and ground truth: **0 / 1 point**

	Label		Ground Truth	
Example	Pneumonia	Mass	Pneumonia	Mass
1	1	1	0	1
2	1	0	1	1
3	0	1	0	1
4	0	0	1	0

- ( ) (0.5, 0.83)

- (0.5, 0.8)
- (0.75, 0.8)
- (●) None of the above

! **Incorrect**

10. Now compute the F1 score for all labels jointly: **1 / 1 point**

	Label		Ground Truth	
Example	Pneumonia	Mass	Pneumonia	Mass
1	1	1	0	1
2	1	0	1	1
3	0	1	0	1
4	0	0	1	0

- 1.35
- 0.61
- (●) 0.66

None of the above

Correct

Explanation: The overall recall is ⅗, while the overall precision is ¾. Therefore the F1 score is 2*⅗*¾ / (⅗ + ¾) = 18/20 / 27/20 = 18/27 ~ 0.66. Therefore the correct answer is C. Note that it is not B, which is the harmonic mean of the individual class F1 scores, since 2*0.5*0.8 / (0.5 + 0.8) ~ 0.62, and it is not A, which is the arithmetic mean of the overall recall and precision.