1. WorNet and vectors both capture relationships between words.

   (a) (2 marks) What are the ways they represent relationships?

WordNet collect similar words together.

Vector (Glove): will update word vectors based on the other words which appear with it at the same time. such as fruit and apple will have closer vector.

WordNet: A database of labeled relationships between words.

Vectors: Location in the space, e.g., proximity can indicate similarity.

   (b) (1 mark) What are the benefits of WordNet's approach?

simple, more straight forward, base on human understanding

   (c) (1 mark) What are the benefits of using vectors?

it may find hiden pattern among words, since it mainly based on the occurence of paired words.

Twooptions: (a) low human effort as they are determined automatically, (b)can represent soft relationships or varying degrees of a relationship

2. (1 mark) What does it mean if two words have low distributional similarity?

For dot: it may lead by small dimension of vectors or two vecotrs are on the different direction

For cos: it means 2 vectors toward different direction

The words are used in different contexts

3. (1 mark) How could a Bag of Words model be adapted to account for word senses?

we can compare word vector's similarity to determine their sense is similar or not.

这里只是bag of words**不是**cbow

Identify the word sense of each word and modify the entries in the bag to include the specific sense. For example, instead of inserting "bat," insert "bat (the animal)."

4. (1 mark) Why do we usually care more about True Positives than True Negatives?

True negative can not clearly show your model is working. Consider a dataset include many false cases, even your model cannot capture true label at all, it will still get pretty good score. However, that's not what we want our model to do.

5. (1 mark) When we talked about interpretability of hidden states in RNNs, we said things like "This position turns on inside quotes." What does 'turns on' mean?

It means each state can have some information about previous token.

The value of that position in the hidden state (or cell state) is non-zero or significantly different from zero.

6. What do each of the following labels mean in NER?

(a) (1 mark) B-PERSON

Begin of a person Entity

(b) (1 mark) I-LOCATION

Inside a location entity

(c) (1 mark) O

outside entity

7. (1 mark) What do random sampling, top-k sampling, and top-p sampling all have in common?

they all choose the top examples (even with different methods),
they all use random sampling.

8. (1 mark) In beam search, when does an item get removed from the beam?

I think you mean a path. So the path will be removed when we find new top m highest-score sequence.

9. (1 mark) In the sentence below, which nouns would be put in the same clusters by a perfect coreference resolution system?

”I came to the room with the chocolate. I saw the chocolate. I ate the chocolate.”

chocolate

[I, I, I], [the room], and [the chocolate, the chocolate, the chocolate]

10. (1 mark) Does the Viterbi algorithm only work with RNNs? Why / Why not?

No, as long as we have previous input and current state, it should work. Because it satisfy the requirements which include a transition and emmision model.

11. (1 mark) Consider the decoder in an encoder-decoder.

(a) (1 mark) Where does the first input come from?

from decoder input

It is a fixed value that we provide <sos>

(b) (1 mark) Where do all the other inputs come from?

from encoder

Each input token is the output from the previous step of the decoder

12. (1 mark) We saw a range of different equations for different types of attention. What are those equations all calculating?

the relationship between each token pairs and generate a context vector for each token

13. (1 mark) What is the value of having multiple heads in attention?

it can capture more information by allowing each head focus on different area, such as one focus on long distance, one focus on short distance, etc.

14. (1 mark) What does a residual connection do with an input vector?

it just keep a back up of original vector.

It passes the vector through a computation (e.g., a feedforward layer) and then adds the original vector back afterwards

15. One way to train a language model is with token edit detection. Another model (the editor) edits some of the input tokens and the model you are training (the predictor) has to predict ”edited” or ”same” for every token in the input.

(a) (1 mark) What would go wrong if edit most of the tokens?

If you edit too much tokens, then the sentence lose its original meaning, so the LM cannot detect the edit token base on rest unchange token because their quantity is not enough

The model might struggle to learn because there is no context to inform the decision about what has been edited

(b) (1 mark) What would go wrong if we only edited one of the tokens?

In most cases a token can be replace without leading any error, such as replace apple with banada in most case work

The model might learn very slowly because there is a very limited training signal

16. (1 mark) How does an n-gram language model use data to inform its probabilities?

current token only rely on previous n -1 tokens

N-grams in the data are counted and the counts are used to determine probabilities by dividing one count by another

17. (1 mark) Given an example of how you can use the probabilities from a language model to classify a news article into one of these four topics: politics, entertainment, sport, tech.

Prompt

I will collect frequent words for each topic, and use LM base on the frequency of those words to classify the final topic

Provide the article as input, followed by the prompt ‘the topic of this article is:’, then look at the probability distribution for the next token. Compare the four topic options and select the one with the highest probability

18. (1 mark) When we use RAG, we calculate the similarity between things. What are we comparing and why?

Internal token and external resources. Because we want to find precise external link.

The query is compared with each piece of text in the database to identify the most relevant texts.

19. (1 mark) When doing annotation, one option is to edit the outputs of a model.

(a) (1 mark) Where does that model come from?

from previous work, trained on other data

(b) (1 mark) What problem(s) could that cause?

the model might not accurate or have certain patterns

20. Consider the table of annotation counts for two people below.

		Annotator 1	
		Label A	Label B
Annotator 2	Label A	5	5
	Label B	5	5

(a) (1 mark) What will Cohen's Kappa be in this case?

Pe= P(1=A)P(2=A)+ P(1=B)(P2=B) = 0.5*0.5+0.5*0.5=0.5

Pe= 10/20 * 10/20 + 10/20 * 10/20

Po = 10/20=0.5
Cohen = 0

(b) (1 mark) What would happen if all of the numbers were doubled?

same result

21. (2 marks) RLHF involves a model with a purpose we had not discussed earlier in the unit. What is the model and what is its purpose?

reward model, to reward prefered answer.