THE UNIVERSITY OF ARIZONA

Great Learning
POWER AHEAD

Applied NLP - Project - [Marks 10]

Note: Each Question carries 0.25 marks including all sections.

I. True or False

1. We can use the spacy library to pre-process documents in Spanish. __True__

2. In keras, a TimeDistributed layer is a wrapper that allows a Recurrent layer to return an output for every token in the sequence. __True__

3. When using bag-of-words, if we sort the vocabulary alphanumerically the resulting vectors will be the same as if we sort it randomly. __True__

4. To represent out-of-vocabulary words with one-hot encoding we can use a vector where all the values are zeros. __True___

5. We could represent all the 10,000 words of a vocabulary with word embeddings of 10 dimensions. __True__

6. You trained a logistic regression model to predict if a text contains misinformation or not. For an input article, the model returns a predicted value of 0.43, so you could classify the article as containing misinformation. __False___

7. If you were working on a binary text classification problem and all the examples in the training set were positive (equal to 1), the cross-entropy would be equal to

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i))$$

. ___False___

8. Training a bi-directional Recurrent Neural Network is more efficient than a unidirectional Recurrent Neural Network because it requires less parameters to capture long dependencies. _____False___

9. In a LSTM, if the output of the forget gate is a vector with all zeros, the unit should forget all the information from the cell state. __True___

10. Training a Neural Network for a sequence-to-sequence problem, we usually need to pad/truncate the input sequences to have the same length, but we don't need to pad/truncate the output sequences. _____False__

11. Contextual word embeddings can handle word polysemy. __True___

12. A Pre-trained Language Model that has been already fine-tuned on a specific task cannot be longer fine-tuned on a different one. __False___

13. A sparse self-attention can be used to process longer sequences because its computational requirements grow quadratically. ___**False**___

14. Just like GPT, training chatGPT does not involve human supervision. _**False**__

15. A sigmoid function only accepts input values between 0 and 1. _**True**__

II. Fill in the Blank

1. The __**adjudicator**__ is in charge of resolving discrepancies among the annotations produced by the annotators.

2. _**Stop words**_are words that can be filtered out from textual data because they are so frequent that they provide little information.

3. You have the sentence *"Time flies when you're having fun."* tokenized by words and annotated with Part-of-Speech. To represent this annotation following the BIO schema, there should be _**6**_ tokens with the label O.

4. A n-gram language model that only attends to the previous word in the sequence, is called a ___**bigram**__language model.

5. The dot product of 2 word-embeddings is 10. If their magnitudes were 5 and 4, their cosine similarity would be _**0.5**__.

6. You are working on a text classification problem with 3 classes, and you have implemented a model with a softmax in the output layer. For a specific input, the model returns the following probabilities: 0.31, 0.14 and _**0.55**___.

7. A character-based tokenization of the sentence *"Can't wait, it's almost vacation time."* would result in __**38**___tokens.

HINT: Do not include the double quotes.

8. The maximum value of BLEU's Brevity Penalty is ___**1.0**____.

9. ___**Attention mechanism**__allows a deep-learning model to selectively focus on certain parts of the input sequence based on the relevance of each token to the others.

10. Given the following confusion matrix for a sentiment analysis model: Prediction

		positive	Negative	neutral
Truth	positive	37	5	0
	negative	3	17	8
	neutral	16	21	32

The macro-average f1 score is **0.609** .

HINT: A multi-class confusion matrix for N classes can be converted into N one-vs-all binary confusion matrices.

III. Multiple Choice

1. What AutoClass of the transformers library could be used to instantiate a pre-trained model for a sequence labeling task?

a. AutoModelForTokenClassification

b. AutoModelForMaskedLM

c. AutoModelForSequenceClassification

d. AutoModelForCausalLM

Answer: a. AutoModelForTokenClassification

2. Which of the following pre-processing steps should always be taken?

a. None, it depends on the task.

b. Word tokenization

c. Sentence segmentation

d. Lemmatization

Answer: b. Word tokenization

3. A word that is very frequent in a document but very infrequent in the rest of the documents in a corpus will have:

a. High tf and high idf.

b. High tf and low idf.

c. Low tf and high idf.

d. Low tf and low idf.

Answer: b. High tf and low idf.

4. Given the co-occurrence probabilities ratio $p(w_1|w_2)/p(w_1|w_3) = 10$, GloVe will learn embeddings for w_1, w_2 and w_3 such that:

a. w_1 and w_2 are closer together than w_1 and w3.

b. w_1 and w_3 are closer together than w_1 and w2.

c. w_2 and w_3 are close together but far apart from w1.

d. w_1, w_2 and w_3 are all close together.

Answer: a. w1 and w2 are closer together than w1 and w3.

5. The range of the output values of a relu function is:

a. [0, ∞)

b. (-∞, ∞)

c. [0, 1]

d. [0, 10]

Answer: a. [0, ∞)

6. A model for Named Entity Recognition is able to identify all the entities in a test set, however the model is only able to predict one token per entity. For example, for the named entity *"Frida Kahlo"*, the model only identifies *"Frida"*. Using a relaxed evaluation, the precision of the model would be:

a. 1

b. 0

c. 0.5

d. Depends on the total number of tokens per entity.

Answer: c. 0.5

7. Both GTP and BERT are based on the transformer architecture, but they only use part of it:

a. GPT uses the decoder and BERT uses the encoder.

b. GPT uses the encoder and BERT uses the decoder.

c. Both use the encoder.

d. Both use the decoder.

Answer: a. GPT uses the decoder and BERT uses the encoder.

8. Which of the following statements about the Embedding layer is not true?

a. It is a lookup table that maps words to their indices.

b. Its weights can be trained as parameters of a neural network.

c. It can be initialized with random vectors.

d. It can be a square matrix.

Answer: d. It can be a square matrix.

9. Which of the following NLP approaches is most suitable for sentence segmentation?

a. Sequence Labeling

b. Text Classification

c. Sequence-to-Sequence

d. Language Modeling

Answer: c. Sequence-to-Sequence

10. Which of the following corruptions of the tokenized input *"SGD is an optimizer. It learns from errors."* would not be used for pre-training BART?

a. "SGD is an error. It learns from optimizer."

b. "It learns from errors. SGD is an optimizer. "

c. "from errors. SGD is an optimizer. It learns"

d. "SGD. an optimizer. It learns. errors."

Answer: d. "SGD. an optimizer. It learns. errors."

IV. Short Answer

1. Why do we need to train Tfidfvectorizer of scikit-learn on the training data?

Training the TfidfVectorizer on the training data is essential to ensure that the vectorization process accurately reflects the characteristics of the specific training dataset. It allows the vectorizer to learn the vocabulary from the training data, calculate the inverse document frequency (IDF) based on the entire training corpus, and establish a consistent transformation for both the training and test datasets. By fitting the vectorizer to the training data, data leakage from the test set is prevented, and realistic representations of documents are obtained, making it a fundamental step in building machine learning models for text data.

2. What information should be included in the annotation guidelines?

The annotation guidelines should include essential information such as the task's description, the annotation schema or labelling rules, exemplar annotations, instructions for addressing ambiguity, and quality control measures. These elements ensure clarity and consistency in the annotation process.

3. It is possible to make Beam Search behave as Greedy Search. How?

Yes, it is possible to make Beam Search behave as Greedy Search. One way to do this is to set the beam size to 1. This means that at each time step, only the single best word will be kept. This is equivalent to the Greedy Search algorithm, which always selects the word with the highest probability at each time step.

4. The [CLS] token provides an aggregate representation of the sequence that is used to fine-tune BERT for text classification tasks. What is the [CLS] token used for during BERT pre-training?

During BERT pre-training, the [CLS] token has a dual role: it is used for sentence classification in the "Next Sentence Prediction" task, where BERT learns to predict whether two input sentences are consecutive or not, and it also provides an aggregated representation of the entire input sequence. This aggregated representation is valuable for various downstream tasks and can be fine-tuned for text classification, named entity recognition, and other NLP tasks, making the

[CLS] token a central element in BERT's pre-training and its versatility in handling various language understanding tasks.

5. How can we help to distinguish training examples from different tasks when pre-training a multitask seq-to-seq model?

To help distinguish training examples from different tasks when pre-training a multitask sequence-to-sequence model, several approaches can be employed. One effective strategy is to incorporate task-specific tokens or prefixes in the input sequences, indicating the associated task for each example. Additionally, using separate training datasets for distinct tasks, introducing auxiliary objectives that predict task identification, implementing task dropout, and designing task-specific architectural components, such as multi-head models, all contribute to task differentiation. Including domain or metadata tags and organizing mini-batches by task further aids in emphasizing the task-specific learning, ensuring that the model can effectively handle a range of sequence-to-sequence tasks in a multitask setting.

Happy Learning ☺