Course	20241_csci_544_30063: Applied Natural Language Processing
Test	Quiz 1
Started	1/16/24 5:02 PM
Submitted	1/16/24 5:09 PM
Due Date	1/16/24 5:15 PM
Status	Completed
Attempt Score	100 out of 100 points
Time Elapsed	6 minutes out of 10 minutes
Instructions	Rules and Regulations: 1. You can only submit your answers once. 2. Once you submit your answer you can no longer see the question or edit your answer 3. If you have any technical problems during the quiz, please take a screenshot and send us. 4. This is an individual quiz and you have to do it without any outside help
Results Displayed	All Answers, Submitted Answers, Correct Answers, Feedback, Incorrectly Answered Questions

Question 1

10 out of 10 points

Which of the following statements about feature extraction are **WRONG**:

Selected Answers:

- Capitalization is a significant feature for sentiment classfication task
- Negation words are important for named entity recognition task

Answers:

- Capitalization is a significant feature for sentiment classfication task
- Function and stop words are useless for sentiment classfication task
- Negation words are important for named entity recognition task
- Preposition is important for named entity recognition task

Question 2

10 out of 10 points

Which of the following statements regarding word vocabulary construction for feature extraction is **WRONG**?

Selected Answer: Mapping each word to their overall counts

Answers:
- Mapping each word to a word id
- Mapping each word to their overall counts
- Calculating the counts of each word and selecting the top K words with the most counts
- The word id for each each word is unique

Question 3

10 out of 10 points

Which of the following method is **NOT** discriminative?

Selected Answer: Naive Bayes

Answers:
- Neural Networks (NNs)
- Support Vector Machine (SVM)
- Logistic Regression
- Naive Bayes

Question 4

10 out of 10 points

To compute the Inverse Document Frequency (IDF) of a word, we divide the number of documents in the corpus by the total counts of documents without that word and perform a logarithm operation, i.e., idf_i=log #documents/#documents without i

Selected Answer: False

Answers:
- True
- False

Question 5

10 out of 10 points

Which of the following methods are normally used to represent each data instance into a vector?

Selected Answers:
- Term Frequency - Inverse Document Frequency (TF-IDF)
- Bag of Words (BoW)

Answers:
- Term Frequency
- Term Frequency - Inverse Document Frequency (TF-IDF)

Inverse Document Frequency

Bag of Words (BoW)

Question 6

10 out of 10 points

Which of the following model is **NOT** appropriate for text classfication tasks?

Selected Answer: HMM

Answers:
Logistic regression
HMM
Linear Classifier
SVM

Question 7

10 out of 10 points

Which of the following operation is **NOT** included in tokenization:

Selected Answer: Removing rare characters

Answers:
Removing rare characters
Removing extra spaces
Removing external URL links
Removing non-alphabetical characters

Question 8

10 out of 10 points

With the Bag of Words (BoW), given the Vocab = [good, bad, nice, expensive, love], we can convert the following sentence "*It is a good idea to give a nice present to someone you love*" into a vector.
NOTE: please split the element by comma and write the vector in the following format: [<element$_0$>,<element$_1$>,<element$_2$>,<element$_3$>...], e.g., [0,0,0,0,0]

Selected Answer: [1,0,1,0,1]

Correct Answer:

Evaluation Method	Correct Answer	Case Sensitivity
Exact Match	[1,0,1,0,1]	

Question 9

10 out of 10 points

Which of the following limitations are **TRUE** for Bag of Words (BoW)?

Selected Answers:
Resulting vectors are highly sparse
Contexture information is discarded

Answers:
- Word dependencies are overlooked
- Resulting vectors are highly sparse
- Rare words dominate the vector representation
- Contexture information is discarded
- Word dependencies are overlooked

Question 10

10 out of 10 points

Which of the following task-model pair is **inappropriate**:

Selected Answer: Naive bayes for text generation

Answers:
- LSTM for sequence labeling
- Tree-structured models for syntactic parsing
- Transformers for machine translation
- Naive bayes for text generation

Wednesday, March 6, 2024 4:17:26 PM PST

← OK

Course	20241_csci_544_30063: Applied Natural Language Processing
Test	Quiz 2
Started	1/23/24 5:01 PM
Submitted	1/23/24 5:10 PM
Due Date	1/23/24 5:15 PM
Status	Completed
Attempt Score	94.16666 out of 100 points
Time Elapsed	9 minutes out of 10 minutes
Results Displayed	All Answers, Submitted Answers, Correct Answers, Incorrectly Answered Questions

Question 1

10 out of 10 points

Word2Vec is a Factorization-based word embedding approach.

Selected Answer: False

Answers: True

False

Question 2

10 out of 10 points

Which of the following is the ReLU function?

Selected Answer: $\sigma(x) = max(0,x)$

Answers: $\sigma(x) = min(0,x)$

$\sigma(x) = max(0,x)$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Question 3

10 out of 10 points

Which of the following statements accurately describes a step in the backpropagation algorithm for training a neural network?

Selected Answer: Calculating the gradient of the loss with respect to the weights and biases.

Answers:
- Forward pass, where inputs are passed through the network to compute the output.
- Adjusting weights and biases based on the difference between predicted and actual outputs.
- Calculating the gradient of the loss with respect to the weights and biases.
- Initializing network parameters randomly before training.

Question 4

10 out of 10 points

Which of the following is not a hyper-parameter when training a NN?

Selected Answer: Weights of each layer

Answers:
- Batch size
- Learning rate
- Weight decay
- Weights of each layer

Question 5

10 out of 10 points

The hyperparameters in feedforward neural networks are learnable parameters.

Selected Answer: False

Answers:
- True
- False

Question 6

10 out of 10 points

In the context of Word2Vec, what is the purpose of sub-sampling?

Selected Answer: Sub-sampling involves skipping frequent words during training to improve efficiency and focus on informative words.

Answers:
- Sub-sampling is used to randomly remove a portion of the training data to reduce model overfitting.
- Sub-sampling involves skipping frequent words during training to improve efficiency and focus on informative words.
- Sub-sampling is a technique to augment the training dataset with additional samples for better model generalization.
- Sub-sampling refers to adjusting the learning rate during training to control the convergence speed of the model.

Question 7

10 out of 10 points

Feedforward neural networks is good at modeling sentences with various lengths.

Selected Answer: False

Answers:
- True
- False

Question 8

4.16666 out of 10 points

What is/are **true** about word embeddings?

Selected Answers:
- It is a dense vector per word
- It can be used to compute the similarity of distance between two different words

Answers:
- It contains contextual information from the sentence
- It is a dense vector per word
- It can be used to compute the similarity of distance between two different words
- It contains contextual information from the sentence
- We can perform linear operations on top of word embeddings

Question 9

10 out of 10 points

Which of the followings is **False** about non-convex optimization problems?

Selected Answer: It's guaranteed to converge to a global optimum

Answers:
DNNs are non-convex
There are multiple local optimal points
It's guaranteed to converge to a global optimum
Initialization will impact the final convergence point

Question 10

10 out of 10 points

Which of the followings is/are the advantage(s) of CNN?

Selected Answers:
Parallel computation
Simple architecture

Answers:
Parallel computation
Simple architecture
Small context window
Good at modeling long dependencies

Wednesday, March 6, 2024 4:23:18 PM PST

← OK

Course	20241_csci_544_30063: Applied Natural Language Processing
Test	Quiz 3
Started	1/30/24 5:01 PM
Submitted	1/30/24 5:10 PM
Due Date	1/30/24 5:20 PM
Status	Completed
Attempt Score	90 out of 100 points
Time Elapsed	8 minutes out of 10 minutes
Results Displayed	All Answers, Submitted Answers, Correct Answers, Feedback, Incorrectly Answered Questions

Question 1

10 out of 10 points

What is the primary advantage of using Conditional Random Fields (CRF) over HMMs in sequence labeling tasks?

Selected Answer: CRFs allow for the incorporation of rich, overlapping features, capturing complex dependencies in the data.

Answers:

CRFs are generative models, while HMMs are discriminative models.

CRFs allow for the incorporation of rich, overlapping features, capturing complex dependencies in the data.

CRFs can handle sequential data with variable-length observations, while HMMs require fixed-length observations.

CRFs are more computationally efficient than HMMs for most sequence labeling tasks.

Question 2

10 out of 10 points

Which is NOT a fundamental property of homogeneous Hidden Markov Models (HMMs)?

Selected Answer: The assumption that the number of hidden states is infinite.

Answers:
- The assumption that transition and emission probabilities do not depend on the position in the Markov chain.
- The assumption that the future state of the system depends only on the current state.
- The assumption that the number of hidden states is infinite.
- The assumption that given the current hidden state, the observations are conditionally independent of each other.

Question 3

10 out of 10 points

Which of the following statements about decoding strategy in HMMs is/are WRONG?

Selected Answers:
- The greedy decoding algorithm runs in $O(mk^2)$ time, where m is the sequence length, k refers to the number of observed tags.
- Greedy decoding considers the entire history of states.

Answers:
- Viterbi decoding guarantees the globally optimal state sequence, while greedy decoding does not.
- The greedy decoding algorithm runs in $O(mk^2)$ time, where m is the sequence length, k refers to the number of observed tags.
- Greedy decoding is more computationally efficient than Viterbi decoding.
- Greedy decoding considers the entire history of states.

Question 4

10 out of 10 points

Regarding HMMs and Maximum Entropy Markov Models (MEMMs), HMMs are generative models, while MEMMs are discriminative models.

Selected Answer: True

Answers:
- True
- False

Question 5

10 out of 10 points

Log-linear models are a type of linear regression model used primarily for continuous data rather than categorical data.

Selected Answer: False

Answers: True
False

Question 6

0 out of 10 points

Which of the following statements about PyTorch is TRUE?

Selected Answer: PyTorch operates at a single level of abstraction, providing only a high-level API for deep learning tasks.

Answers:
PyTorch operates at a single level of abstraction, providing only a high-level API for deep learning tasks.

PyTorch does not support automatic differentiation (AutoGrad) for computing gradients.

PyTorch seamlessly integrates with NumPy, allowing for easy conversion between NumPy arrays and PyTorch tensors.

Mini-batching is a technique in PyTorch used for training neural networks sequentially, one sample at a time.

Question 7

10 out of 10 points

In the context of Part-of-Speech (POS) Tagging, which is/are classified as open class?

Selected Answers: Nouns
Adjectives

Answers:
Prepositions
Nouns
Adjectives
Determiners

Question 8

10 out of 10 points

Generative models focus on modeling the conditional probability of the output given the input, enabling them to generate new data samples and gain a deeper understanding of the underlying data distribution.

Selected Answer: False

Answers: True
False

Question 9

10 out of 10 points

Which is NOT structured prediction?

Selected Answer: Text classification

Answers:
- Information extraction
- Named entity recognition
- Machine translation
- Text classification

Question 10

10 out of 10 points

Which functions should be implemented to construct a custom dataset using torch.utils.data.Dataset?

Selected Answers:
- __init__
- __getitem__
- __len__

Answers:
- __init__
- __getitem__
- __len__
- forward

Wednesday, March 6, 2024 4:53:29 PM PST

← OK

Course	20241_csci_544_30063: Applied Natural Language Processing
Test	Quiz 4
Started	2/6/24 5:00 PM
Submitted	2/6/24 5:10 PM
Due Date	2/6/24 5:20 PM
Status	Completed
Attempt Score	90 out of 100 points
Time Elapsed	10 minutes out of 10 minutes
Results Displayed	All Answers, Submitted Answers, Correct Answers, Feedback, Incorrectly Answered Questions

Question 1

10 out of 10 points

What is the primary goal of syntactic parsing?

Selected Answer: To identify the grammatical structure of a sentence.

Answers:
- To identify the grammatical structure of a sentence.
- To identify the sentiment of a sentence.
- To identify the topic of a sentence.
- To identify the meaning of a sentence.

Question 2

10 out of 10 points

Which of the following RNN architectures is specifically designed to handle the vanishing gradient problem by using gates to control the flow of information?

Selected Answer: Long Short-Term Memory (LSTM)

Answers:
- Long Short-Term Memory (LSTM)
- Multi-Layer Perceptron (MLP)

Autoencoder

Convolutional Neural Network (CNN)

Question 3

10 out of 10 points

In the context of RNNs, what is the "vanishing gradient problem"?

Selected Answer: The issue where gradients become too small, effectively preventing the network from learning long-range dependencies

Answers:
The issue where gradients become too small, effectively preventing the network from learning long-range dependencies

The problem of gradients disappearing when the learning rate is too high

The problem of gradients vanishing because of too many layers in the network

The issue where gradients become too large and cause the model to become unstable

Question 4

10 out of 10 points

In a basic RNN, what information is used as input for the next time step?

Selected Answer: Both the current input and the previous hidden state

Answers:
Both the current input and the previous hidden state

Only the current input

None of the above

Only the previous hidden state

Question 5

10 out of 10 points

Which of the following are NOT problems of Simple Recurrent Neural Networks (RNN)?

Selected Answer: Small context window

Answers:
Small context window

Not modeling future information

Vanishing gradients

Inefficient sequence computation

Question 6

10 out of 10 points

In dependency parsing, a sentence is represented as a directed acyclic graph.

Selected Answer: True

Answers: True

False

Question 7

10 out of 10 points

Dependency parsing represents sentence structure as a tree with words as nodes and directed edges showing their grammatical dependencies.

Selected Answer: True

Answers: True

False

Question 8

10 out of 10 points

Bidirectional RNNs allow processing information in both directions (forward and backward) simultaneously, potentially capturing more context.

Selected Answer: True

Answers: True

False

Question 9

10 out of 10 points

Which of the following is NOT a type of syntactic parser?

Selected Answer: Semantic parser.

Answers: Rule-based parser.

Probabilistic parser.

Neural network-based parser.

Semantic parser.

Question 10

0 out of 10 points

What is a "projective" dependency parse tree?

Selected Answer: [None Given]

Answers:

It focuses on noun phrases and verb phrases only.

Lines do not intersect except for parent-child connections.

The tree captures long-distance dependencies accurately.

Each word has exactly one parent node.

Wednesday, March 6, 2024 8:42:41 PM PST

← OK

Course	20241_csci_544_30063: Applied Natural Language Processing
Test	Quiz 5
Started	2/13/24 5:07 PM
Submitted	2/13/24 5:16 PM
Due Date	2/13/24 5:20 PM
Status	Completed
Attempt Score	87.5 out of 100 points
Time Elapsed	9 minutes out of 10 minutes
Results Displayed	All Answers, Submitted Answers, Correct Answers, Feedback, Incorrectly Answered Questions

Question 1

0 out of 10 points

Which of the following statements is consistent with the empirical findings of semi-supervised machine translation discussed in class? Select all that apply.

Selected Answers:

Noisy sentences in target language are often added to boost the robustness of the encoder.

Noisy sentences in source language are often added to boost the robustness of the decoder.

Answers:

Noisy sentences in target language are often added to boost the robustness of the decoder.

Noisy sentences in target language are often added to boost the robustness of the encoder.

Noisy sentences in source language are often added to boost the robustness of the decoder.

Noisy sentences in source language are often added to boost the robustness of the encoder.

Question 2

10 out of 10 points

When calculating the attention scores, which of the following attention types involves learnable weights? Select all that apply.

Selected Answers: Multiplicative
Additive

Answers: Dot-product
Multiplicative
Additive

Question 3

10 out of 10 points

In the vanilla encoder-decoder architecture with simple RNNs, (select all that apply)

Selected Answers: a single encoding vector needs to capture all the information about the source sentence.
longer sequences can lead to vanishing gradients.

Answers: at each time step, all of the latent source representations are used.
a single encoding vector needs to capture all the information about the source sentence.
longer sequences can lead to vanishing gradients.

Question 4

10 out of 10 points

The IBM translation models discussed in class are word-alignment models in ____.

Selected Answers: Statistical Machine Translation

Answers: Statistical Machine Translation
Neural Machine Translation

Question 5

10 out of 10 points

Non-autoregressive decoding generates each token in the output sequence independently of the others, while autoregressive decoding generates each token based on the previously generated tokens.

Selected Answer: True

Answers: True
False

Question 6

10 out of 10 points

Which of the following statements on Multilingual NMT is true? Select all that apply.

Selected Answers:
- The limited amount of annotated multilingual data is one of the lasting problems in NMT.
- Applying zero-shot transfer in Many-to-Many NMT helps with the issue of having no parallel data available between a pair of source and target languages.

Answers:
- The limited amount of annotated multilingual data is one of the lasting problems in NMT.
- It is recommended that we keep individual vocabulary sets for each language.
- Applying zero-shot transfer in Many-to-Many NMT helps with the issue of having no parallel data available between a pair of source and target languages.

Question 7

10 out of 10 points

In semi-supervised neural machine translation, if monolingual data from the target side is available, ____ can be used.

Selected Answers: back-translation

Answers:
- back-translation
- self-learning

Question 8

10 out of 10 points

Given the following words and their corresponding counts: {“high”: 2, “hello”: 3, “height”: 2}, what will be added to the vocabulary after the first iteration of byte pair encoding (BPE)?

Selected Answer: he

Correct Answer:

Evaluation Method	Correct Answer	Case Sensitivity
Exact Match	he	*Case Sensitive*
Exact Match	"he"	*Case Sensitive*
Exact Match	"he": 5	*Case Sensitive*
Exact Match	"he":5	*Case Sensitive*

Question 9

10 out of 10 points

Which of the following evaluation metrics is NOT contextualized/embedding-based? Select all that apply.

Selected Answers: METEOR

Answers: METEOR
BERTScores
BLEURT

Question 10

7.5 out of 10 points

Which of the following <input, output> pairs is usually formatted as a seq2seq generation task? Select all that apply.

Selected Answers:
- <a sequence of image frames, action classes>
- <speech in the original language, text in the target language>
- <a collection of movie reviews, a summarizing description of the reviews>

Answers:
- <a sequence of image frames, action classes>
- <speech in the original language, text in the target language>
- <a collection of movie reviews, a summarizing description of the reviews>

Wednesday, March 6, 2024 9:30:14 PM PST

← OK

Course	20241_csci_544_30063: Applied Natural Language Processing
Test	Quiz 6
Started	2/20/24 5:00 PM
Submitted	2/20/24 5:10 PM
Due Date	2/20/24 5:20 PM
Status	Completed
Attempt Score	80 out of 100 points
Time Elapsed	9 minutes out of 10 minutes
Results Displayed	All Answers, Submitted Answers, Correct Answers, Feedback, Incorrectly Answered Questions

Question 1

10 out of 10 points

How does the multi-head attention mechanism contribute to the transformer model's ability to process sequential data?

Selected Answer: By enabling the model to attend to different parts of the sequence simultaneously, capturing a broader range of dependencies.

Answers:

By allowing the model to focus on a single representation of the input sequence, simplifying the learning process.

By enabling the model to attend to different parts of the sequence simultaneously, capturing a broader range of dependencies.

By strictly increasing the model's focus on local dependencies, thereby improving performance on tasks requiring detailed textual analysis.

By reducing the overall depth of the network, thus lowering the computational cost associated with training transformers.

Question 2

0 out of 10 points

Token_1: {q:1 ,k:2 ,v:3}

Token_2: {q:0 ,k:0 ,v:3}

Token_3: {q:3 ,k:1 ,v:3}

Token_4: {q:6 ,k:6 ,v:3}

When computing the attention score for the third input element, ___ will have the largest impact.

Selected Answer: Token_3

Answers:
Token_1
Token_2
Token_3
Token_4

Question 3

10 out of 10 points

Transformers can only be used for natural language processing tasks, such as translation and sentiment analysis, and cannot be adapted for image recognition or generative tasks.

Selected Answer: False

Answers:
True
False

Question 4

10 out of 10 points

Token_1: {q:1 ,k:2 ,v:3}

Token_2: {q:0 ,k:0 ,v:3}

Token_3: {q:3 ,k:1 ,v:3}

Token_4: {q:6 ,k:6 ,v:3}

When computing the attention score for the second input element, token_1 has more impact than token_3.

Selected Answer: False

Answers:
True
False

Question 5

0 out of 10 points

In the context of multi-head attention within transformers, why is the dimensionality of key, query, and value vectors divided by the number of heads?

Selected Answer: To ensure that the computational complexity per head is reduced, making the model more efficient.

Answers:
- To increase the computational complexity and improve the model's accuracy.
- To maintain the model's overall parameter count unchanged despite increasing the number of heads.
- To ensure that the computational complexity per head is reduced, making the model more efficient.
- To decrease the amount of memory used by each attention head, facilitating the processing of longer sequences.

Question 6

10 out of 10 points

In the Transformer architecture, the dimensions of the query, key, and value vectors must always be identical to ensure the proper calculation of attention scores in the self-attention mechanism.

Selected Answer: False

Answers:
- True
- False

Question 7

10 out of 10 points

The original Transformer model, as introduced in the paper "Attention is All You Need," utilizes convolutional layers to extract spatial features before applying self-attention mechanisms.

Selected Answer: False

Answers:
- True
- False

Question 8

10 out of 10 points

Token_1: {q:1 ,k:2 ,v:3}

Token_2: {q:0 ,k:0 ,v:3}

Token_3: {q:3 ,k:1 ,v:3}

Token_4: {q:6 ,k:6 ,v:3}

Ignoring d_k, what is the attention score for the second input element.

Selected Answer: 1/4

Answers: 1/4

0

1

1/2

Question 9

10 out of 10 points

The self-attention mechanism in transformers allows them to process all parts of the input data simultaneously, making them inherently parallelizable and more efficient than RNNs for sequence processing tasks.

Selected Answer: True

Answers: True

False

Question 10

10 out of 10 points

Despite their efficiency in parallel processing, transformers inherently require more memory than RNNs for processing long sequences due to their self-attention mechanism.

Selected Answer: True

Answers: True

False

Wednesday, March 6, 2024 10:05:18 PM PST

← OK