CSSE 315 – Natural Language Processing
Rose-Hulman Institute of Technology

Exam 1 Review Solutions

Name (Print):_____ Date:_____

1 Linguistic Studies

1. **True/False:** Lexical ambiguity refers to a word having multiple meanings. **True**
   Note: Lexical ambiguity refers to the phenomenon where a single word has multiple meanings, making it inherently tied to the concept of meaning (example - bank: financial institution and river bank)

2. **True/False:** Morphology studies the structure of words and their components like roots and affixes. **True**

3. **True/False:** Syntax focuses on arranging words to form meaningful sentences. **True**
   Note: It focuses on the rules and principles that govern sentence structure and the arrangement of words

4. **Multiple Choice:** Which linguistic field studies meaning in context?

   (a) Syntax
   (b) **Pragmatics**
   (c) Morphology
   (d) Phonology

   Note: Pragmatics is a subfield of linguistics focused on how language is used in specific contexts and how meaning is influenced by those contexts. It goes beyond the literal meanings of words (semantics) to include speaker intention.

5. **Fill in the Blanks:** _____ studies the rules governing sentence structure. **Syntax**

6. **True/False:** Language ambiguity is one of the key challenges in natural language understanding for computers. **True**

7. **Multiple Choice:** Why is human language difficult for computers to process? Select all that apply.

   (a) **Variability in grammar and syntax**
   (b) **Cultural and contextual nuances**
   (c) **Ambiguity in meaning**
   (d) Only one meaning

8. **Fill in the Blanks:** The first chatbot, **ELIZA**, simulated a psychotherapist by reflecting user inputs.

2 Tokenization

9. **True/False:** Subword tokenization breaks words into smaller units like prefixes, suffixes, or roots. **True**

10. **Multiple Choice:** Which of the following is a tokenization method? Select all that apply.

    (a) **Character-level tokenization**
    (b) **Word-level tokenization**
    (c) **Subword-level tokenization**
    (d) **All of the above**

11. **Fill in the Blanks: Subword** tokenization is commonly used in large language models to handle rare or unknown words.
    Note: Subword tokenization methods like Byte Pair Encoding (BPE) and WordPiece are widely used in models such as BERT, GPT. Rare or unknown words are broken into smaller units, such as prefixes, suffixes, or root segments - so the model can still understand and process parts of the word, even if the full word is not in its vocabulary

12. **True/False:** Lemmatization reduces words to their dictionary form, while stemming may produce linguistically invalid root forms. **True**

13. **Multiple Choice:** Which of the following is a unique example of stemming that differs from lemmatization?

    (a) Running → Run
    (b) Happier → Happy
    (c) **Studies → Studi**
    (d) Better → Good

3 Large Language Models and Transformer Architecture

14. **True/False:** Transformer models process input tokens sequentially, one at a time. **False**
    Note: Transformers process all input tokens simultaneously using mechanisms like self-attention. This enables them to analyze relationships between tokens in parallel rather than sequentially.

15. **Multiple Choice:** Which component of the Transformer architecture is responsible for capturing relationships between tokens?

    (a) Positional encoding
    (b) **Attention mechanism**
    (c) Feedforward neural network
    (d) Dropout layer

    Note: The attention mechanism enables the model to focus on relevant parts of the input sequence when processing a particular token.

16. **Fill in the Blanks: Attention** is a mechanism in transformers that allows models to focus on relevant parts of the input.

17. **Multiple Choice:** Which of the following distinguishes GPT from BERT?

    (a) **GPT is a unidirectional model, while BERT is bidirectional.**

    (b) BERT generates text, while GPT only classifies text.

    (c) GPT uses transformers, while BERT does not.

    (d) BERT is trained on smaller datasets compared to GPT.

    GPT (Generative Pre-trained Transformer) processes text in a left-to-right manner. BERT considers both the left and right context simultaneously to understand the meaning of a word or phrase within a sentence.

4 NLP Tasks and Regular Expressions

18. **True/False:** "New York" is an example of Named Entity Recognition (NER) task. **True**
    Note: "New York" is identified as a proper noun and categorized as a location.

19. **Multiple Choice:** Which of the following is NOT an NLP task?

    (a) Sentiment Analysis

    (b) Named Entity Recognition

    (c) **Sorting Algorithms**

    (d) Machine Translation

5 Optimization Techniques

20. **True/False:** Model distillation reduces the size of a neural network while maintaining its performance. **True**
    Note: In model distillation, a large, pre-trained model (called the teacher model) transfers its knowledge to a smaller model (called the student model).

21. **Multiple Choice:** Quantization in NLP models typically refers to:

    (a) Reducing model accuracy

    (b) **Using smaller numerical representations for weights**

    (c) Increasing the size of embeddings

    (d) Compressing input data

    Note: Quantization replaces high-precision floating-point numbers (e.g., 32-bit or 16-bit floats) with lower-precision numbers (e.g., 8-bit integers).

22. **Fill in the Blanks: Quantization** is an optimization technique used to reduce computational requirements while preserving accuracy.

23. **True/False:** Reinforcement learning with human feedback (RLHF) uses Reward Model to predict human preferences. **True**
    Note: A separate model, the Reward Model, is trained to predict human preferences by learning from the rankings provided by the evaluators. The Reward Model outputs a scalar value representing how well a model's output aligns with human preferences.

24. **True/False:** Locality Sensitive Hashing (LSH) is used for approximate nearest neighbor search in high-dimensional data. **True**
    Note: Locality Sensitive Hashing (LSH) and SCANN (Scalable Nearest Neighbors) both address the challenge of finding approximate nearest neighbors in high-dimensional data but differ significantly in methodology, performance, and use cases. SCANN combines techniques like partitioning and quantization to accelerate similarity search. LSH uses hash probabilities and groups similar data points into the same bucket, ensuring that nearby points in the original space have a high probability of landing in the same bucket.

25. **Multiple Choice:** SCANN is optimized for:

    (a) Sorting text alphabetically

    (b) **Accelerating vector similarity search**

    (c) Grammar checking in NLP

    (d) Tokenizing text

6 Embeddings and Vector Search

26. **True/False:** Word2Vec is a dynamic embedding. **False**
    Note: Word2Vec creates fixed vector representations for words. Each word has one static embedding regardless of the context in which it appears.

27. **Multiple Choice:** Which type of database is most commonly used for vector search?

    (a) Relational database

    (b) Graph database

    (c) **Vector database**

    (d) Document database

28. **Fill in the Blanks: BERT embeddings** are an example of dynamic context-aware embeddings.
    Note: A word's embedding changes based on its usage in a sentence. For example, "bank" in "river bank" and "bank account" would have different embeddings because the context is considered.