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

Exam 1 Review

Name (Print):_____ Date:_____

1 Linguistic Studies

1. **True/False:** Lexical ambiguity refers to a word having multiple meanings.
2. **True/False:** Morphology studies the structure of words and their components like roots and affixes.
3. **True/False:** Syntax focuses on arranging words to form meaningful sentences.
4. **Multiple Choice:** Which linguistic field studies meaning in context?
   (a) Syntax
   (b) Pragmatics
   (c) Morphology
   (d) Phonology
5. **Fill in the Blanks:** _____ studies the rules governing sentence structure in a language.
6. **True/False:** Language ambiguity is one of the key challenges in natural language understanding for computers.
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, _____ , simulated a psychotherapist by reflecting user inputs.

2 Tokenization

9. **True/False:** Subword tokenization breaks words into smaller units like prefixes, suffixes, or roots.
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:** _____ tokenization is commonly used in large language models to handle rare or unknown words.

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

13. **Multiple Choice:** Which of the following is a unique example of stemming? that is it is different 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.

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

16. **Fill in the Blanks:**_____ 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.

4 NLP Tasks and Regular Expressions

18. **True/False:** New York is an example of Named Entity Recognition (NER) task.

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

5 Optimization Techniques

20. **True/False:** Model distillation reduces the size of a neural network while maintaining its performance.

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

22. **Fill in the Blanks:** _____ 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 is used to human preferences.

24. **True/False:** Locality Sensitive Hashing (LSH) is used for approximate nearest neighbor search in high-dimensional data.

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

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:** _____ is an example of dynamic context-aware embeddings