Kasdi Merbah University – OUARGLA
Faculty of New Information and Communication Technologies (FNTIC)
Department of Computer Science and Information Technologies

Exam (Semester 1)
-MASTER 2 - AI

Course: Natural Language Processing
Duration: 1 hour and 30 minutes
(No authorized documents)

---

1) What are word embeddings?

   1. **Vectors representing words, such that semantically similar words are represented by similar vectors.**
   2. Vectors used to compress the meaning of a text in sequence-to-sequence problems.
   3. A mechanism that made neural networks more efficient, leading to the birth of the Transformer neural network.

2) What do nodes and edges in a semantic networks typically represent?

   1. Nodes represent relationships, and edges represent concepts.
   2. Nodes represent words, and edges represent semantic relationships between the words.
   3. **Nodes represent concepts, and edges represent relationships between the concepts.**

3) What is tokenization in NLP?

   1. **The process of splitting text into tokens, which are usually individual words but they can be also single characters or subwords.**
   2. The process of inferring tokens from text, such as noun phrase, verb phrase, noun, adjective, etc.
   3. The process of solving ambiguities in sentences, linking words to their meanings in lexical databases.

4) What is text vectorization?

   1. The process of converting text into an unordered set of words.
   2. **The process of converting text into a numerical vector that most machine learning models can understand.**
   3. The process of converting text into vectors of words that are easier to use by machine learning models.

5) What is a bag of words representation?

   1. A representation where text is represented as the set of its words, disregarding word order but considering grammar.
   2. **A representation where text is represented as the set of its words, disregarding grammar and even word order but keeping multiplicity.**
   3. A representation where text is represented as the ordered list of its words, disregarding grammar but keeping multiplicity.

6) What do we mean by inflection when talking about natural languages?

   1. **It's the modification of a word to express different grammatical categories such as tense, case, and gender.**
   2. It's the modification of a word over time to express different meanings.
   3. It's the tone of voice of a specific word relative to its context.

7) Which of the following reduces words always to a base-form which is an existing word?

1. Stemming.
2. **Lemmatization.**
3. Tokenization.
4. Normalization

8) Which of the following can leverage context to find the correct base-form of a word?

1. Stemming.
2. Normalization
3. **Lemmatization.**
4. Tokenization.

9) Which of the following is typically faster?

1. **Stemming.**
2. Lemmatization.

10) Stopwords are the words that, in a typical corpus, have the…

1. **Most occurrences.**
2. Least occurrences.

11) As a consequence of TF-IDF:

1. **Rare words have high scores, common words have low scores**
2. Common words have high scores, rare words have low scores

12) Choose the correct option.

1. Word embeddings models are often finetuned models that can be retrained for specific use cases.
2. Word embeddings models must be trained from scratch for each specific use case.
3. **Word embeddings models are often pre-trained models that can be finetuned for specific use cases.**

13) A morpheme is…

1. A meaningful sound
2. A meaningful word
3. **The smallest unit of meaning in a language**
4. A sentence structure

14) An inflectional morpheme is used for…

1. Changes the grammatical category of a word
2. Adds extra meaning to a word
3. **Marks grammatical relationships, such as tense or plural**
4. Combines two words to form a new one

15) Which of the following is an example of an inflectional morpheme?

1. -er (as in "teacher")
2. **-s (as in "cats")**
3. -ly (as in "quickly")
4. un- (as in "unhappy")

16) Which morphological process involves combining two or more words to create a new one?

1. Derivation
2. **Compounding**
3. Inflection
4. Prefixing

17) Which morphological process involves changing the grammatical category of a word?

1. Inflection
2. **Derivation**
3. Compounding
4. Reduplication

18) Which technique is used to represent words as dense vectors in a continuous vector space?

1. One-Hot Encoding
2. Bag-of-Words
3. **Word Embedding**
4. TF-IDF

19) What is the primary purpose of a thesaurus in NLP?

1. Spell checking
2. **Synonym identification**
3. Named entity recognition
4. Speech synthesis

20) What is WordNet primarily used for in NLP?

1. Speech synthesis
2. Named entity recognition
3. **Semantic analysis and lexical relations**
4. Text summarization

21) In WordNet, what does a synset represent?

1. A synonym
2. A set of homophones
3. **A set of words with similar meanings**
4. A part-of-speech tag

22) What is PARSEVAL in the context of NLP?

Answer:

PARSEVAL refers to a set of evaluation metrics used to assess the performance of syntactic parsers. It involves comparing the output of a parser with a gold standard (manually annotated syntactic structures) to measure the accuracy and effectiveness of the parser.

23) Consider the following PCFG (Probabilistic Context-Free Grammar) with non-terminal symbols S, NP, VP, and terminal symbols N (noun), V (verb), and Det (determiner). The associated probabilities are as follows:

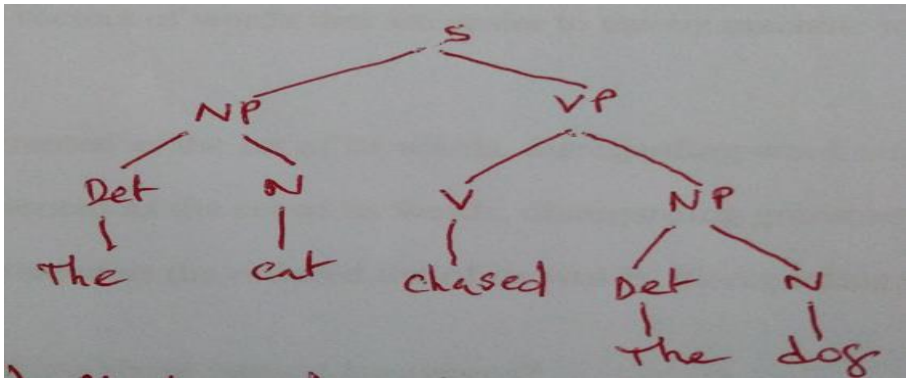
S → NP VP [0.6]
NP → N [0.4]
NP → Det N [0.6]
VP → V [0.5]
VP → V NP [0.5]
Det → "the" [0.8]
N → "cat" [0.7]
N → "dog" [0.3]
V → "chased" [0.8]

Given the sentence "The cat chased the dog" :

1. Construct a parse tree for the sentence.
2. Compute the probability of generating this sentence according to the PCFG.

Answer:

1) The parse tree



2) The probability is calculated by multiplying the probabilities of the production rules used in the derivation.
To compute the probability, multiply the probabilities of the production rules used in the derivation:
P(S -> NP VP) * P(NP -> Det N) * P(Det -> "The") * P(N-> "cat")* P(VP -> V NP) * P(V -> "chased") * P(NP -> Det N) * P(N -> "the") * P(N-> "dog")
= 0.6 * 0.6 * 0.8 * 0.7 * 0.5 * 0.8 * 0.6 * 0.8 * 0.3
= 0,01161