Natural Language

MAP 2 (40 minutes)

Student number: _____

Student name: _____

This MAP has two parts:

- The first part contains multiple-choice questions. There is only one correct answer for each question. Multiple-choice questions might have different scores.
- Write down your answer in the table bellow (**mandatory!**). The scoring only considers this table. If you make a mistake, cross out the wrong answer and write the new answer. If you leave two or more answers, the question will be considered wrong.
- The second part contains open question(s). Limit the size of responses to the available space. You can answer in **Portuguese or English**. Write legibly or your answer might not be evaluated.

During this evaluation there is no clarification of doubts. If you detect an error, mark it. If you are right, you will be given the full score for that question.

Good luck!

Table for Multiple-Choice Answers

Scoring: correct answer = 1, wrong answer = $-1/\#options$, no answer = 0.

Question	Chosen answer	Question	Chosen answer
1		6	
2		7	
3		8	
4		9	
5		10	

Professor corner (ignore this, please):

# Correct	# Wrong	# Not Answered	Open Questions

Algorithms

Viterbi

$i \leftarrow 1$
while $i < N$ **do**
  SS(i, 1) = P(w_1 | L_i)*P(L_i | $< s >$)
  BP(i, 1) = 0
  $i++$
end while
$t \leftarrow 2$
while $t < n$ **do**
  $i \leftarrow 1$
  while $i < N$ **do**
    SS(i, t) = $max_{j=1,\ldots,N}$ SS(j, t-1)*P(L_i | L_j)*P(w_t | L_i)
    BP(i, t) = j that resulted in the maximum score
    $i++$
  end while
  $t++$
end while
C(n) = i that maximizes SS(i, n)
$i \leftarrow n - 1$
while $i > 1$ **do**
  $i--$
  C(n) = BP(C(i+1), i+1)
end while

- N = number of tags
- n = number of words in the sequence

- Data structures:
  - SS (sequence score) – records the score of the best sequence found up to a given position with category L.
  - BP (Back Pointer) – records the previous state to a given state
  - C – records the best sequence of tags.

CKY

Algorithm 3 CKY

$j \leftarrow 1$
while $j < n$ **do**
  [1, j] = {A : A $\rightarrow$ $w_j \in$ R}
  $j++$
end while
$i \leftarrow 2$
while $i < n$ **do**
  $j \leftarrow 1$
  while $j < n - i + 1$ **do**
    [i, j] = $\cup_{m=1}^{i-1}${A : A $\rightarrow$ B C $\in$ R, B $\in$ [m, j], C $\in$ [i - m, j + m]}
    $j++$
  end while
  $i++$
end while
if $S_0 \in$ [n, 1] **then**
  W $\in$ L(G)
end if

Multiple-Choice Questions (10 values)

Some examples of possible multiple-choice questions (not covering the whole set of materials):

1. Which of the following is a characteristic of context-free embedding models?

   a) They generate a single word embedding for each word in the vocabulary
   b) They adjust word embeddings based on the word's context within each sentence
   c) They adjust word embeddings based on the word's context within a paragraph
   d) They adjust word embeddings based on the word's left and right dependencies
   e) None of the above

   Answer: a)

2. Which of the following is an example of derivational morphology?

   a) Adding "s" to "cat" to form "cats"
   b) Changing "run" to "running"
   c) Adding "un-" to "happy" to form "unhappy"
   d) Using "am" instead of "is"
   e) None of the previous

   Answer: c)

3. In Dependency Grammars, which of the following is **NOT** true?

   a) Arcs represent grammatical relations
   b) Cycles are allowed
   c) Each vertex has one incoming arc (except the root)
   d) There is one root with no incoming arcs
   e) spaCy package offers facilities to generate dependency graphs

   Answer: b)

4. What is the main purpose of semantic parsing in computational semantics?

   a) To map user input into syntactic structures
   b) To convert natural language into a meaning representation
   c) To perform word sense disambiguation
   d) To generate sentences from a meaning representation
   e) To classify the intent behind a user's input

   Answer: b)

5. What are the pros of using Large Language Models?

   a) Low computational cost
   b) High interpretability
   c) The ability to apply them to a wide range of tasks
   d) Full accuracy
   e) Minimal energy consumption

   Answer: c)

Open Questions (10 values)

There will be two types of open-questions: a) Two questions to explain NLP concepts (2.5 points each); b) One question to demonstrate your knowledge about an NLP task/problem (5.0 points).

Show us that you understand the concept(s)/problem. No formulas are needed to get the full score. In the following you can find examples of such questions. Notice that they not cover the whole set of materials.

1. Question (2.5 points): What is the BIO notation in NLP? Illustrate your answer with an example in which you show how the BIO notation can be used in an NLP task of your choice.
   Possible answer: BIO (Beginning, Inside, Outside) is a tagging scheme used for labeling sequences, such as named entities in NLP. 'B' marks the beginning of an entity, 'I' marks tokens inside the entity, and 'O' marks tokens outside any entity. For example, in the sentence "John lives in Paris": John (B-PER), lives (O), in (O), Paris (B-LOC). This can be used in a Named Entity Recognition (NER) task to identify person and location names.

2. Question (2.5 points): What is WordNet? How can it be used to detect semantic similarity between two sentences?
   Possible answer: WordNet is a lexical database of English words organized into sets of synonyms called synsets. It also records various semantic relationships between words—such as synonymy, antonymy, hypernymy, and hyponymy. By analyzing how closely related the words in two sentences are within WordNet's hierarchy, we can estimate how similar the sentences are in meaning. This involves mapping each word to its corresponding synset and calculating similarity scores based on the distance between synsets in the hierarchy. The individual word similarities are then combined.

3. Question (2.5 points): Present one advantage and one limitation of using First Order Logic (FOL) to represent meaning in NLP?
   Possible answer: Examples of advantages of FOL: they are human-understandable (unlike vector representations) and provide a support for logical inference (allow systems to derive conclusions and verify relationships between statements). Examples of limitations: require a structured, well-defined small domain and translating sentences into formal FOL expressions is challenging (as you have seen).

4. Question (5 points): You are given a collection of 10 000 documents about the Olympic Games from 10 journalists, clearly identified. A new collection of texts about the same Olympic Games is also given to you and you are asked to associate these texts to their authors (from the same group of journalists). Considering what you have learned in this course, describe: a) your approach; b) how you will evaluate your approach; c) its pros and cons; d) problems you might have with the data. Realistic answers mentioning the concepts, methods, and resources you have studied in NLP will be highly scored. Figures and examples are also welcome.

Possible Answer 1: To associate the new unlabeled texts with the known journalists, I would use a supervised learning approach. First, I'd preprocess the texts through tokenization, stopword removal, and lemmatization, followed by extracting features like TF-IDF and others (e.g., sentence length, punctuation, n-grams). I would then train a classifier such as Naive Bayes or an SVM, using the labeled documents to learn each journalist's writing style. For evaluation, I'd use cross-validation and metrics like accuracy to assess the model's performance. A confusion matrix would help identify where authors are confused with others. Pros: Both SVM and Naive Bayes are computationally inexpensive and can be trained quickly on smaller datasets. Both are easy to implement. Cons: These models may not capture complex patterns and nuances in writing style compared to deep learning models. Problems with the data could include imbalance among journalists, noisy texts, and overlapping styles between journalists.

Possible Answer 2: To tackle authorship attribution, I would utilize deep learning with a pretrained language model such as BERT. I would fine-tune the pretrained model on the labeled documents, adding a classification head for author identification. For evaluation, I would employ metrics like accuracy to assess the model's performance, while a confusion matrix would identify specific authors that are often misattributed. Pros: The model can capture deep contextual and stylistic features without extensive feature engineering. Cons: Problems with the data could include imbalance among journalists, noisy texts, and overlapping styles between journalists.