Master Degree in Computer Engineering

Natural Language Processing
Final Exam

July 21st, 2023

1. **[2 points]** In the context of the linguistic subfield of morphology, answer the following questions.
   (a) Introduce the notions of root and affix, providing some examples in English language.
   (b) What is the distinction between inflectional morphology and derivational morphology? Again, provide some examples in English language.

2. **[6 points]** Some text T has been tokenized based on white spaces. The resulting dictionary and word frequencies are reported in the following table

word	post	postal	port	report
freq	21	5	13	17

   Apply the byte pair encoding algorithm to derive subword tokens for T, using the character '_' to mark the end of each word. Report and comment each of the first nine iterations (merge operations) in your run of the algorithm, including the frequency updates.

3. **[5 points]** With reference to static word embeddings, answer the following questions.
   (a) Introduce the basic idea underlying the skip-gram algorithm, and define the two probabilities $P(+ \mid w, u)$ and $P(- \mid w, u)$.
   (b) Define and discuss the loss function used by the skip-gram algorithm.

4. **[5 points]** Introduce the neural model called sentence-BERT and discuss in which tasks the model is exploited. Explain how training and inference are carried out for sentence-BERT.

(see next page)

5. **[7 points]** Assume some Hidden Markov model, and let $w_{1:n}$ be an input sequence.

   (a) Define the probability of $w_{1:n}$ under the model, written $P(w_{1:n})$.

   (b) Specify the forward algorithm for the computation of $P(w_{1:n})$ in polynomial time.

6. **[2 points]** In the context of phrase structure parsing, introduce the problem of PP-attachment and provide an example in English language.

7. **[6 points]** In the context of text-based question answering, answer the following questions.

   (a) Introduce the distinction between factoid and non-factoid questions by means of some examples.

   (b) Introduce the machine reading task and the notions of query, passage and span. Define the span probability $P(p_i, \ldots, p_j \mid q, p)$ and explain how this probability is approximated by means of start and end probabilities.

   (c) Present and discuss the neural approach to machine reading using contextual embeddings produced by BERT, which we have introduced in class.