Master Degree in Computer Engineering

Natural Language Processing
Final Exam

February 20th, 2023

1. **[2 points]** With reference to the task of part of speech tagging, define the notion of open class and closed class tags, providing some examples.

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	hug	pug	pun	bun	hugs
freq	10	5	12	4	5

   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 eight iterations (merge operations) in a run of the algorithm, showing the frequency updates at each step.

3. **[5 points]** Let $x_{1:n} = x_1, x_2, \ldots, x_n$ be an input word sequence, and let $\mathcal{Y}(x_{1:n})$ be the set of all possible part of speech tag sequences $y_{1:n} = y_1, y_2, \ldots, y_n$ for $x_{1:n}$. Introduce the family of models called linear chain conditional random fields and explain how these models are used to solve the problem of finding the optimal sequence $\widehat{y}_{1:n}$, defined as

   $$\widehat{y}_{1:n} = \operatorname*{argmax}_{y_{1:n} \in \mathcal{Y}(x_{1:n})} P(y_{1:n} \mid x_{1:n})$$

4. **[2 points]** With reference to the task of syntactic analysis, introduce the notion of constituent, also called phrase. Provide some examples of constituency test, used to identify the constituent structure of a sentence.

(see next page)

5. **[5 points]** With reference to contextualized language models, also called pre-trained language models, answer the following questions.

   (a) Briefly explain the notions of adaptation, feature extraction and fine tuning.

   (b) Introduce and motivate the use of adapter modules in the transformer.

6. **[5 points]** In the context of transition-based dependency parsing, consider the English sentence 'these results suggest that Fli-1 is likely to regulate genes' along with the projective dependency tree consisting of the following unlabeled dependency relations:

head	results	suggest	⟨ROOT⟩	is	is	suggest	is	regulate	is	regulate
dependent	these	results	suggest	that	Fli-1	is	likely	to	regulate	genes

   (a) Draw a graphical representation of the dependency tree above, with arcs directed from the head to the dependent.

   (b) Apply to the above tree the oracle presented in class to construct a sequence of training instances for the arc-standard parser.

7. **[6 points]** Let $x_{1:n} = x_1, x_2, \ldots, x_n$ be a source sentence and let $y_{1:n} = y_1, y_2, \ldots, y_m$ be the target translation.

   (a) Explain how $P(y_{1:n} \mid x_{1:n})$ is modeled in neural machine translation.

   (b) Explain how the encoder-decoder neural architecture is exploited to implement neural machine translation.

8. **[2 points]** Introduce the task of entity linking.