Master Degree in Computer Engineering

Natural Language Processing
Final Exam

July 18th, 2024

1. **[1 points]** With respect to the investigation of natural language, introduce the two different positions of rationalism and empiricism, as presented in lecture 1.

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

word	magnus	melior	melius	maior	peius
freq	21	13	19	17	7

   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 ten iterations (merge operations) in a run of the algorithm, including the frequency updates.

3. **[6 points]** In the context of POS tagging for German, consider the hidden Markov model consisting of the transition and emission matrices reported below. These matrices are specified in terms of **costs** (negative log probabilities). As an example, the transition from DET to NOUN has score 5, and the transition from NOUN to DET has score 19.

	NOUN	VERB	DET	⟨EOS⟩
⟨BOS⟩	7	17	5	21
NOUN	10	5	19	5
VERB	10	14	6	5
DET	5	19	18	21

	Hans	Haus	das	kaufte
NOUN	5	5	19	17
VERB	17	18	19	4
DET	15	16	5	19

   Consider the German sentence 'Hans kaufte das Haus' (translation: Hans bought the house). In the following table each entry represents the **lowest cost** of reaching the associated word/POS pair, starting from ⟨BOS⟩. Use the Viterbi algorithm to fill in the table and to calculate the lowest cost POS tag sequence. Report the intermediate computation for each entry in the table.

	⟨BOS⟩	Hans	kaufte	das	Haus	⟨EOS⟩
⟨BOS⟩	0	–	–	–	–	–
NOUN	–					–
VERB	–					–
DET	–					–
⟨EOS⟩	–	–	–	–	–	

(see next page)

4. **[3 points]** Introduce the notion of cross-lingual **static** word embedding and answer the following questions.
   (a) Which classes of objective functions are minimized by cross-lingual word embedding methods?
   (b) Which types of data sources are most suitable in training of cross-lingual word embeddings?
5. **[5 points]** In the context of so-called **contextualized** word embeddings, introduce the BERT neural network, and answer the following questions.
   (a) Describe the bidirectional encoding implemented by BERT.
   (b) Describe and motivate the masked language modeling task.
   (c) Describe and motivate the next sentence prediction task.
6. **[3 points]** Provide the mathematical definition of projective dependency tree and non-projective dependency tree. Report some simple examples of both structures for English language.
7. **[7 points]** In the context of neural machine translation, answer the following questions.
   (a) Introduce the model equations used by the inference algorithm for the encoder-decoder architecture based on recurrent neural network and **static** context vector.
   (b) Introduce the model equations used by the inference algorithm for the encoder-decoder architecture based on recurrent neural network and **dynamic** context vector.
8. **[2 points]** In the context of natural language processing and machine learning, introduce the general problem of structured prediction and define it in a mathematical way as an optimization problem. Discuss an example of a natural language processing task that can be formulated as a structured prediction problem.