Politecnico di Torino
Dipartimento di Automatica e Informatica
Deep NLP (01VIXSM)
Written Exam simulation

January 23rd, 2022

Name: _____
Surname: _____
Student ID: _____

Exam rules:

- The present exam consists of 6 pages (including this cover page) and 7 questions overall. Any inconsistencies/printing errors in the written exam content must be reported to the teacher *at the beginning* of the exam.
- Exam duration: *60 minutes*.
- Withdraw is allowed only *at the end* of the exam.
- The exam is *closed-book*. Electronic devices, mobile phones, smart watches, and extra papers (even blank papers) are *not allowed*.
- Closed-ended questions: cross the right answer (just one) at pag. 2. Wrong or missing answers to closed-ended questions will receive *no penalty*.
- Open questions: write your answers below the text of the question. If you need more space please use the last page (i.e., pag. 6) and/or the back side of the paper.

Evaluation grid

Question:	1	2	3	4	5	6	7	Total
Points:	1	1	1	1	4	6	6	20
Score:								

1. (1 point) What of the following statement holds **True** for the Transformer architecture?
   - The encoder consists of a stack of six layers. Each layer has two sub-layers. The first is a multi-head self-attention module, and the second is a simple, position-wise recurrent neural network.
   - The encoder consists of a stack of six layers. Each layer has two sub-layers. The first is a masked multi-head self-attention module, and the second is a simple, position-wise feed-forward neural network.
   - The encoder consists of a stack of two layers. The first one is based on a multi-head self-attention module, and the second is a simple, position-wise feed-forward neural network.
   - The encoder consists of a stack of two layers. The first one is based on a multi-head self-attention module, and the second is a simple, position-wise recurrent neural network.
   - None of the above.
2. (1 point) The FastText embedding model
   - considers n-grams in place of entire words.
   - considers n-grams beyond entire words.
   - Considers n-grams and syntactical dependencies instead of entire words.
   - Considers n-grams and syntactical dependencies beyond entire words.
   - None of the above.
3. (1 point) In a RASA chatbot architecture a story
   - consists of a sequence of entities, intents, or actions.
   - consists of a sequence of intents and actions.
   - consists in a conversation between a user and an AI assistant expressed in plain text.
   - consists of a set of decision rules applied on top the provided intents.
   - None of the above.
4. (1 point) MultiLingual BERT:
   - do not allow end-users to encode input text snippets longer than 512 tokens.
   - separately processes text snippets written in different languages.
   - does not support finetuning.
   - requires a step of language identification.
   - None of the above.

5. (4 points) Describe the **distributional hypothesis**

    1. State the initial hypothesis.
    2. Provide examples of applications in English.
    3. Enumerate at least two NLP models/techniques that rely on the aforesaid hypothesis.

    5.1: The distribution hypothesis states that words that occur in similar contexts tend to have similar meanings (in terms of semantic similarity).

    5.2: *I go to the bank to ask about new financial instruments.*

    My power bank is out of order: I cannot charge my mobile phone.

    In the foresaid examples, word *Bank* was used in different contexts. I can capture the underlying context by looking into the surrounding words (e.g., *financial* in the former case, *power/charge* in the latter).

    5.3: Word2Vec is an established word embedding model that leverages the distributional hypothesis to infer high-dimensional vector representations of words. It is based on a fixed-size sliding window, which is used to dynamically define the context of the center word. According to the distributional hypothesis, words within the window are likely to be semantically related to the center word.

6. (6 points) Describe the process of **Statistical Machine Translation (SMT)**

    1. State of general problem.
    2. Describe the pipeline.
    3. Compare SMT with the main alternative Machine Translation solutions.

    6.1: Statistical Machine Translation (SMT) entails translating a sentence in one language, called source language, into another language, called target language.

    6.2: The SMT process is typically divided into three parts: a language model, a translation model, and a classifier. The language model assigns probabilities to the target sentences, the translation model assigns probabilities to the source-target sentence pairs based on the likelihood that one is a translation of the other, and a classifier attempts to find the target sentence for which the product of the language and translation model probabilities is maximal.

    6.3: The other solutions are *rule-based MT* and *Deep Learning-based MT*. Rule-based MT relies on dictionaries and grammars covering the main semantic, morphological, and syntactic properties of the text. Unlike SMT, rule-based MT models are less efficient, not flexible, and heavily depend on humanly-provided rules. Deep Learning-based MT adopts end-to-end deep learning models in which neural networks are used to find the correspondence between the source and target languages. Unlike SMT, they often rely on encoder-decoder architectures that do not require advanced feature engineering.

7. (6 points) Describe the LSARank summarization algorithm (*Using Latent Semantic Analysis in Text Summarization and Summary Evaluation* by Josef Steinberger and Karel Jezek).

   1. Define the summarization problem addressed by the paper.
   2. Explain why LSA can be useful for summarization purposes.
   3. Describe the procedure used by LSARank to retrieve the most relevant sentence in the original documents to put into the summary.

   7.1 LSARank addresses the multidocument extractive summarization problem, which entails extracting a sentence-based summary from a collection of documents. It can be potentially applied to multilingual collections (where all documents are written in the same language).

   7.2 LSA can be used to extract the main topics covered by the analyzed documents. The output summary is intended to report all the key topics descriptors with minimal redundancy.

   7.3 The evaluation process compares a summary with the original document from an angle of n most salient topics. The main steps are enumerated below.

   - Perform Singular Value Decomposition to get the reduced version of $U \cdot \Sigma \cdot V^T$. Remove from the original decomposition the dimensions whose singular values fall under the half of the highest singular value.
   - For each term vector in $U \cdot \Sigma$ compute its length, i.e., $s_k = \sqrt{\sum u_{k,i}^2 \cdot \sigma_i^2}$.
   - Put into the summary the sentence with maximal values in s.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.