Exam Web Data Processing Systems 2019/2020

Instructions

The questions sum up to 100 points. You get 5 extra points. The sum of the obtained points will be converted on a decimal scale (but it won't be higher than 10, e.g., 105→10). Answers must be given in English. No external material can be consulted.

Questions

1. Describe one (non trivial) technique to tokenize the text and mention some problematic inputs with this technique (if any). How can you overcome these problems? **[10 pts]**

2. Describe the task of Entity Recognition (ER) **[3 pts]**. When evaluation a ER method, what is an important metric that we should consider in a knowledge extraction pipeline? **[4 pts]** How can we use neural networks to perform ER? **[5 pts]**

3. Many techniques for Entity Linking (EL) perform a syntactic matching between the entity mention and the entity labels in the Knowledge Base. What kind of techniques can we use for this purpose and which are their advantages/disadvantages? **[6 pts]** Describe why syntactic matching is often not enough and what we can do to improve the performance of EL **[7 pts]**.

4. Which are the three main challenges discussed in the class that we need to face when building a Open IE (OIE) engine? **[7 pts]** Do you believe that Entity Linking can improve the precision and recall of OIE? Motivate your answer **[6 pts]**.

5. Rules are used in many different tasks. Formalize and describe two popular types of rules and report two potential scenarios where they can be used **[6 pts]**. Using rules is only one of the possible ways that we can use to infer new knowledge from a Knowledge Graph. Can you (briefly) describe two other methods? **[6 pts]**

6. Describe the notions of standard confidence and PCA-based confidence (reporting the corresponding formulae) used by AMIE and report the main difference between them **[6 pts]**. If we assume that everything that is not in the database is false (closed world assumption), is it true that the output of the PCA-based confidence equals to the one of the standard confidence? Motivate your answer **[6 pts]**.

7. Given in input a database I and a ruleset Π, describe (in a formal way) the materialization process **[6 pts]**. Moreover, describe the restricted and the skolem chase procedures and point out at least one difference between them **[7 pts]**.

8. Describe the problem of fact-checking. Which are the types of fake content, their sources, and which are the assets that we can use to verify them **[5 pts]**? Describe the system DeClarE for performing fact-checking that we discussed in the class **[5 pts]**. What are the differences between this system and ExFaKT? **[5 pts]**