Assignment: Research Engineer in Natural Language Processing

RISE Research Institutes of Sweden

Please read the instructions carefully and feel free to ask clarification questions if anything is unclear. You should complete the assignment within ten days of receiving the email. Once you have completed the assignment, you should notify us by email. In your email, you should include a link to the Github repository where you have uploaded your solution.

Resources:

- MultiNERD Named Entity Recognition dataset: https://huggingface.co/datasets/Babelscape/multinerd?row=17
- Accompanying paper: https://aclanthology.org/2022.findings-naacl.60.pdf
- Please feel free to use any suitable compute resource to which you have access. If you do not have access to suitable resources, you may wish to consider using Google Colab: https://colab.google/

Instructions:

Using the MultiNERD Named Entity Recognition (NER) dataset, complete the following steps to train and evaluate a Named Entity Recognition model for English:

1. Familiarize yourself with the dataset and the task
2. Find a suitable LM model on HuggingFace Model Hub (https://huggingface.co/models). This can be a Large Language Model (LLM) or any type of Transformer-based Language Model
3. Filter out the non-English examples of the dataset
4. Fine-tune your chosen model on the English subset of the training set. This will be system **A**
5. You will now train a model that will predict only five entity types and the O tag (I.e. not part of an entity). Therefore, you should perform the necessary pre-processing steps on the dataset. All examples should thus remain, but entity types not belonging to one of the following five should be set to zero: PERSON(PER), ORGANIZATION(ORG), LOCATION(LOC), DISEASES(DIS), ANIMAL(ANIM)
6. Fine-tune your model on the filtered dataset that you constructed in step 5. This will be system **B**
7. Pick a suitable metric or suite of metrics and evaluate both systems **A** and **B** using the test set
8. Create a new Github repository and upload your code solution. You should include a README file with instructions on how to run your code, and a requirements.txt file to create the environment needed to run your code
9. Write a short paragraph (maximum 200 words) highlighting the main findings of your experiments as well as any limitations of your approach