Project plan:
An Open Source Danish Knowledge Graph Language Model

Søren Winkel Holm, Asger Laurits Schultz
s183911, s183912

March 10, 2021

Natural Language Processing (NLP) is one of the subfields of Artificial Intelligence with the clearest practical applicability. It is, however, also one of the most data hungry fields, resulting in a limited success in transferring methods from English to low-resource language domains such as Danish. A possible mitigation of this limitation of statistical learning is the introduction of explicit knowledge in the modelling, which will be explored in this project.

The contextualized word representation model LUKE (Language Understanding using Knowledge Embeddings) combines explicit knowledge of named entities mined from Wikipedia with the Deep Learning in the form the Transformer architecture. The primary goal of this bachelor project is to produce and analyze daLUKE, a Danish entity-aware model following the LUKE architecture. The NLP subtask of Named Entity Recognition (NER) will be the main approach to examine the performance of the model.

As a part of the project, we aim to

- reproduce existing Danish NER results on a number of public NER datasets and reproduce the NER results of the English LUKE
- pre-train a Danish LUKE-based model on the Danish Wikipedia
- fine-tune daLUKE on NER and compare its' performance and predictions to existing NER models - particularly daBERT

- optimize the performance of daLUKE, possibly by changes to data and architecture, and perform a open-source release of the pre-trained model

Learning outcomes

1. Understand and apply a Machine Learning approach that combines explicit knowledge with statistical learning
2. Apply the training of Deep Natural Language Processing models to a low resource language
3. Implement an open-source Deep Learning model, allowing easy reproduction and application
4. Compare and analyze performance of Natural Language Processing models on an entity-related task and discuss their practical use

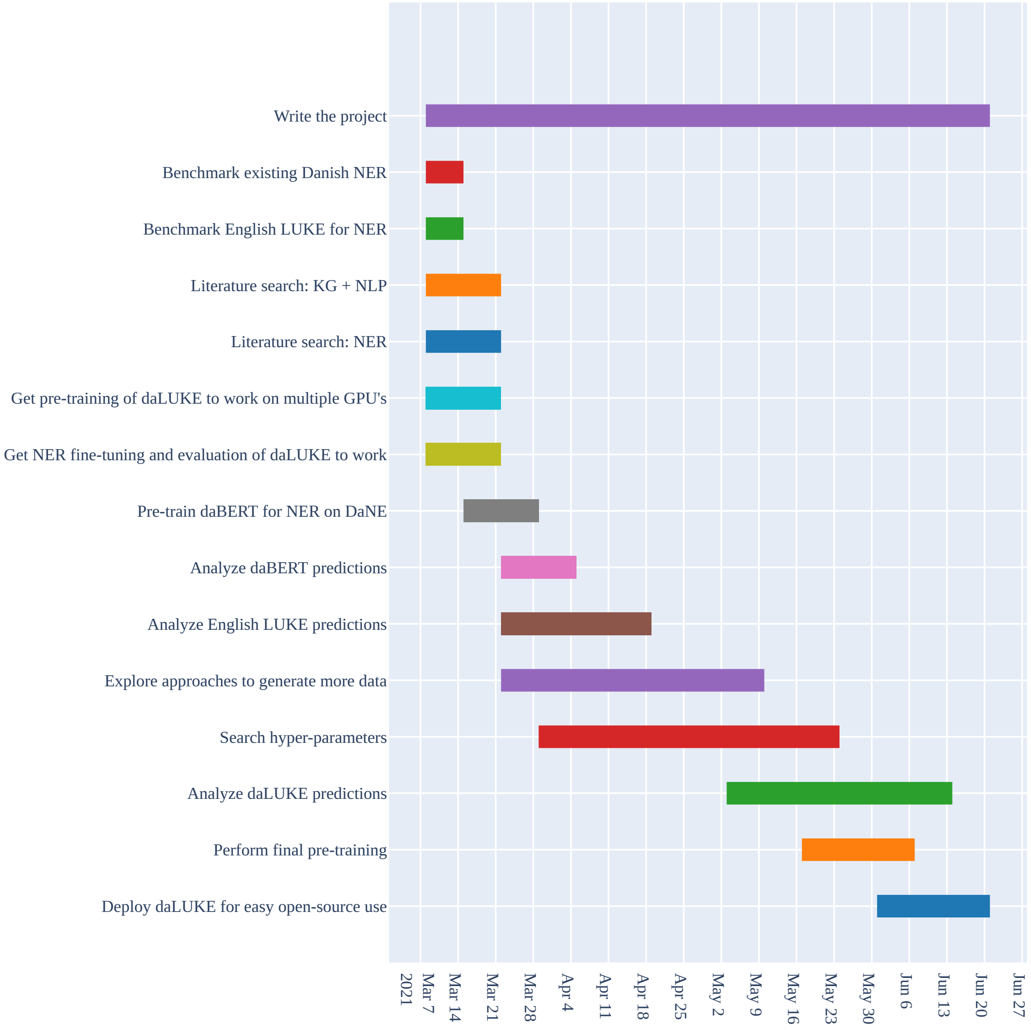

Figure 1: Gantt chart showing the current, quite tentative, project timeline