Team Projects MNLP

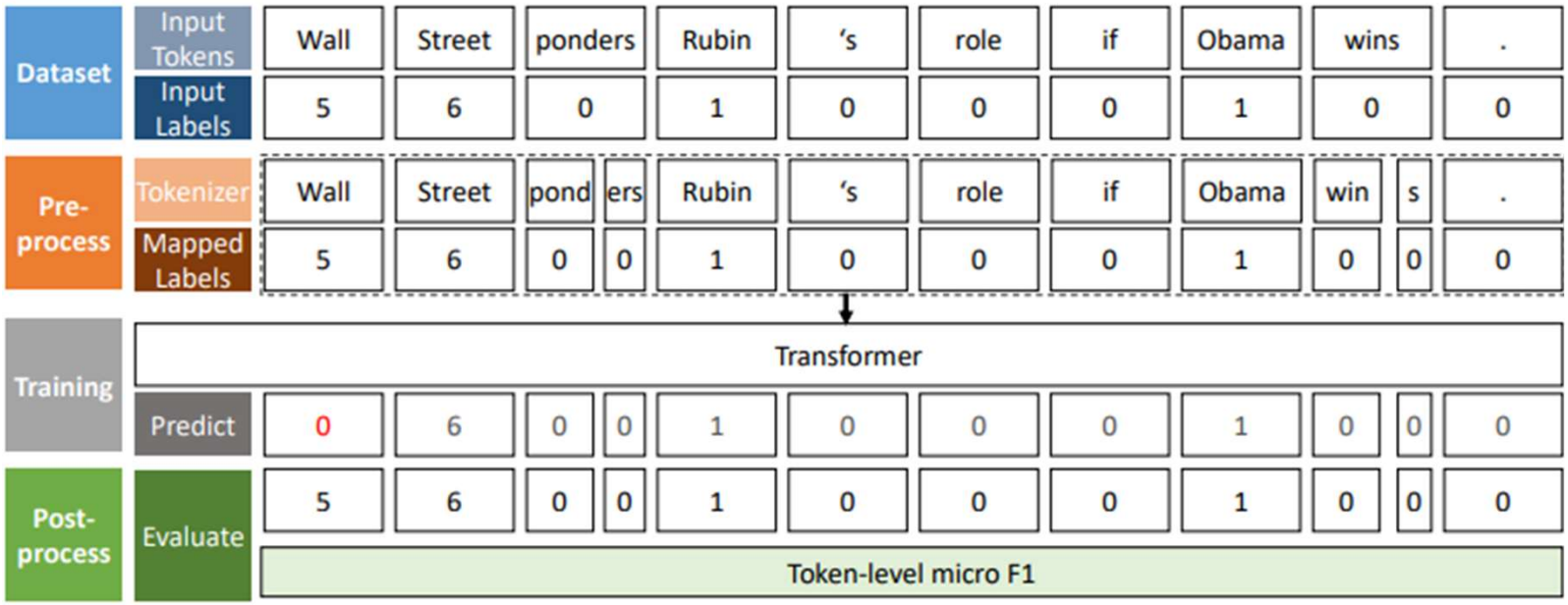
Chair XII for Natural Language Processing

Prof. Dr. Goran Glavaš
Benedikt Ebing

Project Overview

- All groups tackle the same task
  - Bilingual specialization followed by Named Entity Recognition
- Project presentation (~10-15 minutes) followed by short QA (~5 minutes) on 23rd July
- Coaching sessions on demand (up to 2) ➔ schedule via e-mail
- Grading on 4-point scale from 0 to 3 points that count towards the exam bonus
  - Keep the reading assignments in mind
- “Research” like project ➔ Strong empirical results are not necessary

Token Classification for Named Entity Recognition



Task Details

- Base Model: small pre-trained encoder (i.e., XLM-R base)
- Goal:
  1. Implement bilingual (EN+Target Language) specialization (for a target language of choice) ➔ Continual pretraining (MLMing) on a bilingual corpus
  2. Implement zero-shot cross-lingual evaluation on named entity recognition
- Datasets for Token Classification:
  - Source Language: CoNLL 2003 English & WikiANN English
  - Target Language: One of MasakhaNER
- Infrastructure:
  - Google Colab/Kaggle

1. Bilingual Specialization

- Find a text corpus for English and the target language of choice
  - Choose a target language from MasakhaNER
  - Target language should not be included in the pretraining of XLM-R
- Do continual pretraining (MLMing) on the bilingual corpus (EN+Target Language)

2. Token Classifcation

- Fine-Tune on Named Entity Recognition in the following settings (minimum):
  - 3 random seeds
  - Base Model (w/o bilingual specialization)
    - On ConLL
    - On WikiANN
    - One few-Shot setting with 100 instances from the validation set of MasakhaNER
  - Bilingual Specialized Model
    - On ConLL
    - On WikiANN
    - One few-Shot setting with 100 instances from the validation set of MasakhaNER
- Evaluate the model on the target language data from MasakhaNER using the SeqEval implementation of F1

Technical Roadmap – Lightning Module

- Write the Lightning Module
  - Use „xlm-roberta-base“ as encoder
  - Write your own model head for
    - Language Modeling
    - Token Classification
  - Implement all relevant methods of the lightning module

Technical Roadmap - Lightning Data Module

- Write the Lightning Data Module
  - Datasets to use for bilingual specialization:
    - Up to you!
  - Datasets to use for task fine-tuning (train, validation):
    - https://huggingface.co/datasets/eriktks/conll2003
    - https://huggingface.co/datasets/unimelb-nlp/wikiann
  - Datasets to use for final evaluation (test):
    - https://huggingface.co/datasets/masakhane/masakhaner

Technical Roadmap – Training

- Write the final training script
  - Decide on the number of epochs to train for
    - Bilingual Specialization
    - Fine-Tuning for Token Classification (ConLL/WikiANN)
    - Few-Shot
  - Decide on other important hyperparameters (Learning Rate, Learning Rate Scheduler, Sequence Length for Bilingual Specialization, …)
  - Test the model performance on the last checkpoint