FIN-CLARIAH 2024/25

W.P 2.2. TRAINING ENVIRONMENTS

Authors:
Erik Axelson
FIN-CLARIN / University of Helsinki
Collaborators:
Sam Hardwick
Katri Tegel
CSC
Veronika Laippala
Turku NLP / University of Turku

THE GOALS

We aim to provide interactive online training environments for humanities scholars. The training environments will be implemented as **Jupyter Notebook self-study courses**. Their topics will be about creating language modules and adapting and developing transformers and large language models for the automated annotation of spoken and written data.

We will focus on support for **Finnish language or minority languages such as Sami, other Fenno-Ugric languages or Finland Swedish**, as there are fewer resources available for them and it is possible to implement something that is not already available. One possibility is also to use Finnish learner language parallel corpora.

EXISTING TOOLS

We currently have modules for morphological analysis, named entity recognition as well as sentiment analysis of various languages, such as Finnish, Swedish and English. We also offer self-study courses in morphology in CSC's Jupyter Notebooks environment.

PRELIMINARY WORK

Practicalities that need to be taken care of before starting the work:

- What already exists on CSC environment (Puhti, Lumi)
- How these resources can be accessed from Jupyter Notebooks
- Personal data restrictions on Puhti and how these can be handled
- How language models and data can be processed in Secure Desktop environment (SD)
- Using Lumi environment for personal work

Also, collect a list of best practices from others who have already done this kind of work.

DELIVERABLE 2.2.1 TRANSFORMER TRAINING FOR SPECIALISED DATA

This deliverable will be a self-study course implemented as Jupyter Notebooks. Its purpose is to **learn to build up a language model from scratch** in the CSC computing environment using one or more existing resources of Language Bank of Finland, but not limited to them.

The course will **start with simple morphologies** and then continue with more advanced topics. The models will either be distributed via CSC's Puhti environment or downloaded from an external resource and possibly modified.

Examples of morphologies include Finnish, Sami languages and sign languages. More advanced topics could be **spelling correction, sentiment annotation and predictive text input**.

The purpose of the course is to **introduce the scholar to the subject step by step**. The course outline could consist of the following items:

1. Introduction to Transformer Models
2. Setup and Installation of Necessary Libraries
3. Data Preprocessing and Analysis
4. Model Building
5. Model Training
6. Evaluation
7. Fine-tuning and Improvements

Collaborators in the deliverable include CSC and Turku NLP.

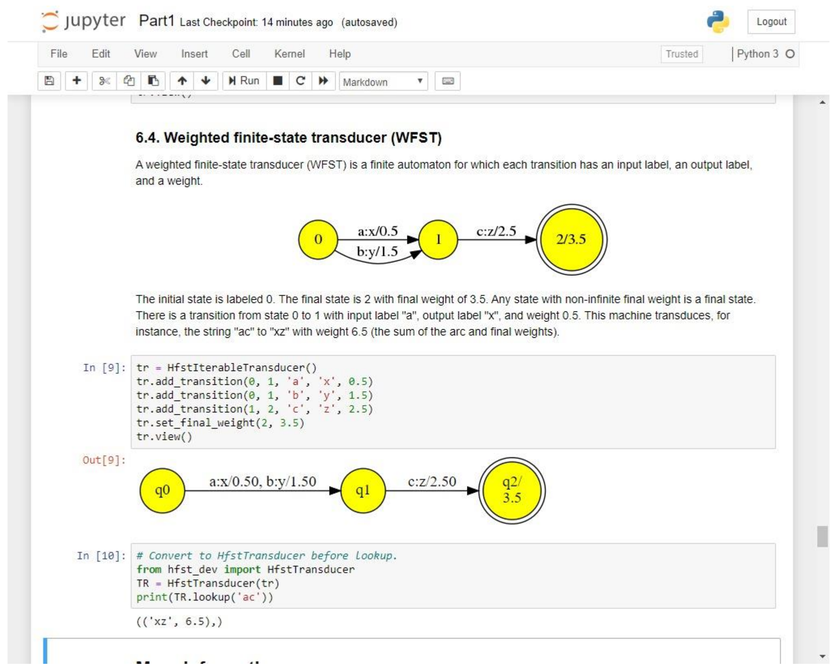
DELIVERABLE 2.2.2 TRANSFORMER ADAPTATION FOR SPECIALISED DATA

This deliverable will also be a self-study course implemented as Jupyter Notebooks. The course will deal with specialised data, such as **speech recognition of a Finnish dialect variant or a big language model adaptation for sentiment analysis**.

The previous deliverable (2.2.1) was about creating a hands-on tutorial on using existing resources. This deliverable (2.2.2) will have two purposes. Firstly, we will have a self-study course to offer to students and scholars. Secondly, when creating the course **we will develop new linguistic resources** that can be offered as a part of the resources of the Language Bank of Finland.

We also aim to write scientific publications based on the observations and results when implementing this deliverable.

Collaborators in the deliverable include CSC and Turku NLP.



A screenshot of an existing Jupyter Notebooks course 'Computational Morphology with HFST'