NLP Assignment 2: Semantic frame extraction

Summary

In this second assignment, you will implement a Conditional Random Field (CRF) model that is able to extract a number of semantic frames from the CALOR corpus. The CALOR corpus is a collection of French encyclopedic documents, which was especially created for the task of semantic frame extraction (Marzinotto et al., 2018). CRFs are a sequence labeling technique and they have been used in many NLP tasks such as POS tagging, chunking or named entity recognition. The difference between these tasks and semantic frame extraction (or semantic role labeling in general) is that words can belong to multiple frames at the same time, which poses an additional challenge to the task.

Before starting this assignment, you should study the paper on semantic frame extraction by [Marzinotto et al.](https://canvas.vub.be/courses/1246/files/153952/download?verifier=jFjYIHPo5Z8vFVTJkttX8U26QX8v0GXPby29FLrj) (available on Canvas). As background material, you can consult the slides of Class 8.

Assignment

For this assignment you need to select three frames of your interest that are annotated in the corpus (for example Accomplishment or Hunting). Make sure that you select one frame with a high frequency, one with a medium frequency and one with a low frequency (cf. Figure 2 in the background paper). You need to train a CRF for each lexical unit involved in these frames, test their performance and report precision, recall and F1 score.

For training the CRFs (one per lexical unit) you follow the approach taken in the paper by Marzinotto et al. As features you use lemma, POS and dependency path. Make sure that you make random train and test splits for every CRF.

About the corpus

The [CALOR corpus](https://canvas.vub.be/courses/1246/files/154348/download?verifier=Txxar7XWMoUHfPeLaGrlQYMxt066t2L3XYo0SLqY) is a 1.3M words corpus that contains semantic-frame annotated documents from four different sources: Wikipedia’s Archeology portal (WA, 201 documents), Wikipedia’s World War I portal (WGM, 335 documents), Vikipedia’s Prehistory and Antiquity (VKH, 183 documents) and ClioTexte’s resources about World War I (CTGM, 16 documents). The frame semantic annotations of CALOR are limited to a subset of frames from FrameNet (53 frames). A total number of 145 lexical units (LUs), nouns and verbs, were selected as frame triggers.

**DISCLAIMER!!!** **The CALOR corpus has not been officially released! Therefore, you should handle the data confidentially and you cannot upload the code of this assignment to public code repositories such as Github or distribute the corpus to people outside of the NLP course.**

You can download it [here](https://canvas.vub.be/courses/1246/files/154348/download?verifier=Txxar7XWMoUHfPeLaGrlQYMxt066t2L3XYo0SLqY).

Handing in your assignment

The format of your assignment should be a Jupyter Notebook. This is a dynamic document that mixes Markdown and code blocks. Use the Markdown blocks to document your code and motivate the decisions you make! Make sure to describe the dataset that you have used, the pre-processing steps to clean up the data and your choices regarding the machine learning algorithm and features. You can see the Jupyter Notebook as a code file and report bundled into one.