Linguistics Data 2 Assignment 4

Zubair Abid (20171076)

2020-04-21

Problem

We are trying to predict AnnCorra tags to be assigned across a sentence, given previously annotated data that also contains feature information of each word, including POS tags, gender, number, case, vibhakti, TAM, and chunk type.

Solution

We build a CRF classifier that takes the features of the word as input, and returns as output the AnnCorra tag for the specific word.

As the tagging depends on the features provided, we change that to check for differences in the result. Starting off with a baseline of including all features for consideration, we move on to include and exclude features to then compare their relative performance.

Results

Features	Weighted Precision	Weighted Recall	Weighted F1
all (baseline)	0.830	0.840	0.830
all but POS	0.798	0.807	0.795
all but chunkType	0.821	0.831	0.830
all but chunkId	0.812	0.821	0.811
all but chunkId, POS	0.785	0.781	0.768
only POS	0.660	0.685	0.656
only POS, Chunk, voice	0.737	0.735	0.719
all but Chunk, voice	0.793	0.805	0.791
all+word but Chunk, voice	0.820	0.833	0.824
all+word	0.853	0.862	0.855

Evaluating the model, and notes on running the code

For ease of evaluation, a script and a test file have been provided. You can run it by typing `python eval.py crf_model.pkl testfile.pkl`, where `crf_model.pkl` is the downloaded model file, and `testfile.pkl` is the provided test file. It will output the F-Score, and a class-wise breakdown of the result.

For running the python notebooks, we need to download the provided `AnnCorra.zip` provided and extract it into a folder. Then, create a `Data` folder. Now all that needs to be run is `dataformatter.ipynb` and `CRFTagger.ipynb`, in succession. The model takes about 10 minutes to train on an i7-5600U.