Overview of Natural Language Processing
Part II & ACS L390

Assignment 3: Inducing Word Categories

Weiwei Sun and Yulong Chen

Department of Computer Science and Technology
University of Cambridge

Michaelmas 2024/25

Assessment

Your marks are based on three practicals:

- Assignment 1: 10%; submitted through Moodle by Thursday 31 October at 12:00
- Assignment 2: 25%; submitted through Moodle by Thursday 14 November at 12:00
- **Assignment 3: 65%**; submitted through Moodle by Thursday 5 December at 12:00

Assignment III: Unsupervised part-of-speech tagging

Goal: Build two computational models to induce word categories from raw text with unsupervised learning.

- Model 1: HMM
- Model 2: K-means over pre-trained word embeddings.

Data and tool

- The data is from Penn TreeBank.
- Training and evaluate your model using word sequences in `ptb-train.conllu`.
- `utils.py`: mainly for evaluation.

Subtask 1

- Implement and train an HMM tagger.

Subtask 2

- Conduct K-means clustering on word embeddings of the Penn TreeBank sentences.
- Use BERT-base-uncased (`https://huggingface.co/google-bert/bert-base-uncased`) to generate word embeddings.

Evaluation

- Key: compare automatically induced tags with the Penn TreeBank part-of-speech labels.
- We provide the implementation of two popular ways to numerically summarise the performance:
  - Variation of information
  - V-measure

Submission

You should submit the following:

- code with *readme*
- report (PDF). The word limit is 2000.