Maschinelle Sprachverarbeitung

Assignment 2: Part-of-Speech Tagging

Ulf Leser

Assignment

- Implement a HMM for POS tagging with the Brown tag set
- We provide a large set of tagged texts for learning
- You learn a MLE-model from this training data
- You implement a tagger using this model (Viterbi)
- We will evaluate your tagger+model on held-back data

Where you can Excel

- Take care of numerical stability
- For transition probabilities, use 1-grams, 2-grams ...
- Take measures regarding unseen words
  - The corpus is “standard English”
- Do not forget about start probabilities
- Smoothing is essential
- Test your methods using 10-fold cross-validation

- Submissions with very bad performance (<50%) are not considered valid

Tag Set Format

- The corpus not only uses the standard tags, but also several modifiers
  - “Fused” words get two tags (“wanna” -> WB+TO), Negation (*), Format hints (titles -TI, headlines -HL) ...
- All this should be ignored
- Consider as POS tag only the prefix of letters, and possibly a closing “$”; and the tags for the special characters like parenthesis (see Brown corpus handbook for details)

Submission by Mail to Ulf Leser

- Results due on 6.12.2015 (three weeks)
- Submit one JAR file called groupX.jar
  - Must start with „java -jar groupX.jar directory
    - JAR file must include your model
  - Must tag all files in „directory/“
  - Write to “directory/ORG_FILE.pos” in same format as training data
  - Include source code
  - Test your JAR before submission
    - There is one test file in the format of the test data
- From the training data
  - Submit accuracy results from 10-fold cross validation plus mean
  - Also submit wall-clock time on GRUENAU4

Competition

- We will evaluate your tagger using a held-back text collection
- Good taggers must be fast and good
- We expect the quality to not vary too much, therefore focus on speed
- The competition is about having a fast tagger with reasonable performance
- Every tagger with accuracy among the top-6 results on the test data participates in competition
- Fastest tagger wins