Maschinelle Sprachverarbeitung

Assignment 5: Gene NER using Conditional Random Fields

Ulf Leser

Assignment

- Perform gene NER with a (linear chain) CRF
- You must use a tool: BANNER or Mallet or ...
  - http://sourceforge.net/projects/banner/
  - https://sites.google.com/site/bannerintrotutorial/
  - http://mallet.cs.umass.edu/index.php
  - Banner has predefined features, Mallet is a “raw” CRF
  - If you prefer another CRF implementation – fine
  - Must be shippable as executable on GRUENAU2
- You may use whatever trick you like
  - Dictionary as feature or post filter, POS tags as feature, lemmatization (BioLexicon), ...
- Setting same as for task 4

Same as Assignment 4: We Provide

- “dictionary_genenames_multitoken.txt”: Dict. gene names
  - Now with multi token entities
- “english_stop_words.txt” ~500 stop words
- “training_annotated.iob”: A gold standard corpus
  - Now with B-Protein and I-Protein
- “training_not_annotated.iob”: For convenience
- “test_not_annotated.iob”: For evaluation
- “eval.scala”: Evaluation script
  - Run with
    <scala eval.scala goldstandard.iob predict.iob>

Competition

- Best F-measure on strict comparison wins
  - See evaluation script
  - scala eval.scala goldstandard.iob goldstandard.predict
    - Precision: 0,40
    - Recall: 0,44
    - F1 Score: 0,42

Submission by Mail to Ulf Leser

- Results due on 7.2.2016
- Must run on gruenau2
- Performance (F1) must be better than 35% on test data
- Submit one JAR file called groupX.jar
  - java –jar groupX.jar test_file_name new_file
  - new_file is the IOB-tagged version of test_file_name
  - Include source code and results of 10-fold CV on training data
    - Use our evaluation script
    - Precision, Recall, F1