Maschinelle Sprachverarbeitung

Assignment 4: Rule-Based Dictionary gene NER

Ulf Leser

Assignment

- Perform dictionary-based gene named entity recognition
- Input
  - Training corpus (with gene names tagged) and a test corpus (only text, no annotations)
    - IOB format
    - All multi-token entities have been removed
  - Dictionary (processing is allowed)
- Output: Annotated test corpus
- You must not apply ML (SVM, HMM, CRF, ...)
- Feel free to use a IE-framework
  - Or write your own fuzzy dictionary matching algorithm

We Provide

- “dictionary_genenames.txt”: ~100.000 human gene names
  - Excerpt from Entrez Gene, all single token
  - All lower case, no duplicates
- “english_stop_words.txt” ~500 stop words
- “training_annotated.iob”: A gold standard corpus
  - Only in B-Protein (single token)
- “test_no_annotation.iob”: Evaluation texts
  - Only B-Protein (single token)
- “eval.scala”: Evaluation script
  - Run with
    <scala eval.scala goldstandard.iob predict.iob>

Your Task: Tag all genes in the test corpus

- Only rule-based / dictionary methods allowed
  - Edit-distance matching, n-gram overlap, stemming, regex, ...
  - No classification: CRF, HMM, SVM, Naïve Bayes, ...
- Rules may be derived from the training data
  - OK: Count POS-n-Grams around matches and turn into rule
  - Not OK: Count POS-n-Grams and turn frequencies into features
- If you want to do something fancy, ask for approval first
- Test method using our evaluation script on the test data
- May use IE-framework: LingPipe, OpenNLP, NLTK, GATE
  - Process the corpus: Load corpus and remove stop words
  - Tag all occurrences of terms from the gene list [in the corpus]
  - Do whatever is necessary with the tool you have chosen

Example

```
Number                  O
of                      O
glucocorticoid          B-protein
receptors               O
in                      O
lymphocytes             O
and                     O
their                   O
sensitivity             O
to                      O
hormone                 O
action                  O
.                       O

The                     O
study                   O
demonstrated            O
a                       O
```

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 24.1.2016
- 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 evaluation on training data
    - Use our evaluation script
    - Precision, Recall, F1