

GDEX - New features in SketchEngine

Milos Husak

Faculty of Informatics, Masaryk University
Botanicka 68a, CZ-602 00 Brno, Czech Republic
E-mail: xhusak@fi.muni.cz

SKEW2 - 2011

Outline

- 1 Introduction
 - GDEX library
 - Configuration files
 - Classifiers
 - Output
- 2 Integration with Sketch Engine
 - Concordance
 - TickBox Lexicography
- 3 GDEX Tools
 - Cooperation with WEKA
- 4 Conclusion

What is it?

- a library for sentence evaluation used in SkE
 - sorting sentences in TickBoxLexicography
 - sorting sentences in Concordance view
 - example application at www.forbetterenglish.com
- goes with couple of GDEX tools
 - ranking of out-of-corpus sentences
 - evaluation of TbLex logs
 - cooperation with WEKA

Classifiers

- classifiers
 - small procedures that quantify measurable features of sentences or tokens
 - sentence classifiers: sentence length, keyword position, ...
 - token classifiers: token frequencies, matches to RE ...
- operators
 - perform some operation or function on input values
 - examples: maximum, minimum, average, normalization, equality ...
 - operate on results of token classifiers or on results of several classifiers

Configuration files

- describe how to evaluate the sentences
- specify the set of classifiers and their combination via operators
- something between an expression and a program

```
name smaller than 4
classifier {
  name smaller than
  classid op_smaller_than
  val 4.0
  subclassifiers [
    {
      name sentence length
      classid s_sentence_length
    }
  ]
}
```

The output

- For sorting of sentences we need a Single value describing the quality of the sentence
- A single number is the only result for each sentence
 - all operators and sentence classifiers return a single number
 - token operators return a list of values (and need to be processed)

Integration with Concordance

- sorting sentences in TickBoxLexicography, and Concordance view
- in case more configurations are available, user can select a custom configuration for every corpus
 - currently the single (English) configuration is used for all corpora
 - however new configurations for Slovene are being evaluated by Amebis

Activating GDEX

Page size (number of lines):

KWIC Context size (number of characters):

☒ Sort good dictionary examples.

Number of lines to be sorted:

GDEX configuration for current corpus:

☐ Icon for one-click sentence copying

☐ Allow multiple lines selection

XML template for one-click copying:

Comparing two configurations

Tickbox Lexicography - Select Examples

Lemma: **test**

Gramret: **a_modifier**

Template: **vanilla**

Alternative GDEX configuration:

Slovene2

GDEX: default configuration

losov

- ☐ Takoj za tem z močnim sunkom avto obrne v drugo smer - in že je tako imenovani losov **test** končan.
- ☐ Losov **test** (povprečno)?
- ☐ Losov **test** v Španiji?
- ☐ Pojem " losov **test** " je naredil skok v simboliko.
- ☐ Drugače zna biti zadek precej hitrejši od nosu, kar kvari losov **test** , slalomske čase pa krajša.
- ☐ Vse drugo, še posebej v zvezi z losovim **testom** , pa je zavajanje bralcev.

dopinški

- ☐ Nikdar nisem bil pozitiven na dopinškem **testu** .
- ☐ SNL uvedli dopinške **teste** .

GDEX: Slovene2

losov

- Takoj za tem z močnim sunkom avto obrne v drugo smer - in že je tako imenovani losov **test** končan.
- Pojem " losov **test** " je naredil skok v simboliko.
- Losov **test** , 50 m (povprečno)?
- Losov **test** je preteklost, 140 KM pa prihodnost.
- Lega je dobra in čeprav na našem » losovem **testu** « ravno ne najboljša, še vedno dovolj športna in adrenalinska.
- Drugače zna biti zadek precej hitrejši od nosu, kar kvari losov **test** , slalomske čase pa krajša.

dopinški

- In za vse so krivi » butasti « dopinški **testi** .
- A takrat je bilo lažje, saj dopinških **testov** ni bilo veliko.

GDEX Tools

- Ranking out-of-corpus sentences
- Evaluating TBLex logs

Both tools export the results into ARFF files that can be processed by WEKA.

ARFF files

- Attribute-Relation File Format
- data definition header
- data part contains results of each classifier (excluding token classifiers) for every sentence

```
@relation Logged_Sentences_from_TBLex
@attribute 'average freq' numeric
@attribute average_wordlen numeric
@attribute good {0,1}
@data
3104326.5,6.91666666667,0
9459885.69231,4.07692307692,0
5219375.6,5.4,0
6309800.1,5.6,0
```

Cooperation with WEKA

- WEKA
 - data mining software developed at The University of Waikato
 - <http://www.cs.waikato.ac.nz/ml/weka/>
 - implements host of Machine Learning algorithms (Bayesian methods, decision trees, neural networks and many more)
 - can visualize the data, results of the machine learning and some of the learned models
- GDEX Tools
 - can export GDEX measurements into ARFF file
 - allow to create configurations based on the data analysis of the ARFF file

- Thank you for your attention