FlawFinder

A Modular System for Predicting Quality Flaws in Wikipedia

TECHNISCHE UNIVERSITÄT DARMSTADT

Oliver Ferschke, Iryna Gurevych and Marc Rittberger

CLEF 2012 Labs and Workshop, Notebook Papers, September 2012. Rome, Italy., September 17–20, 2012

UKP

Introduction

TECHNISCHE UNIVERSITÄT DARMSTADT

UBIQUITOUS KNOWLEDGE PROCESSING

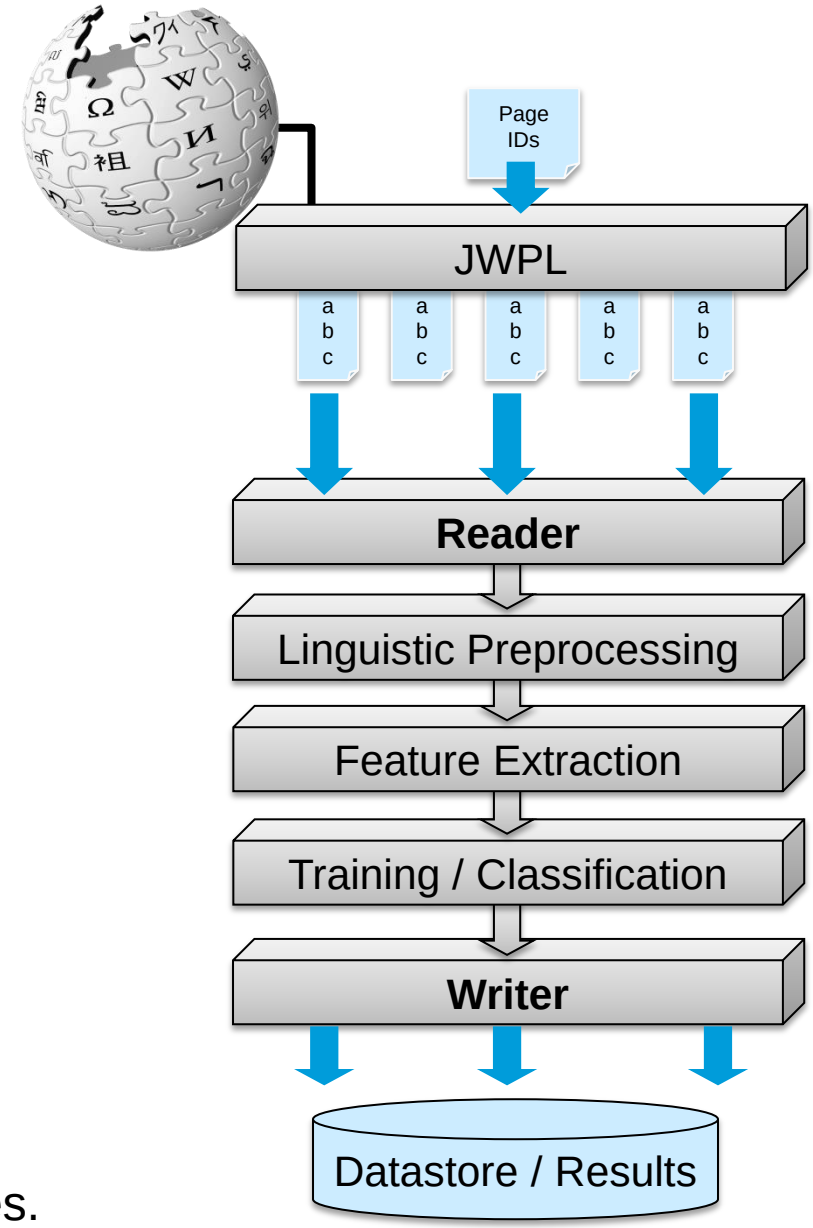
DIPF
Educational Research and Educational Information

Oliver Ferschke

Iryna Gurevych

Marc Rittberger

FlawFinder

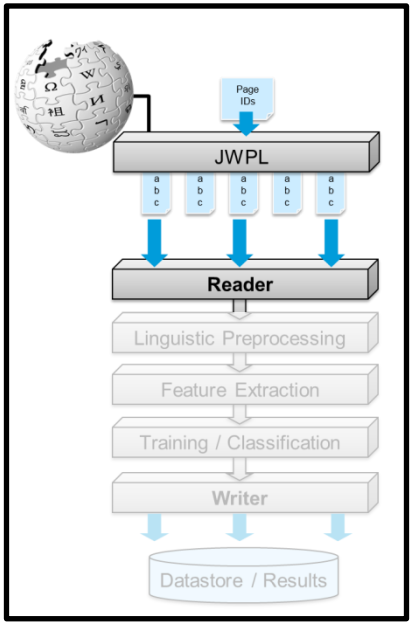


Task-based system with
Multiple processing pipelines.

Data Import

TECHNISCHE UNIVERSITÄT DARMSTADT

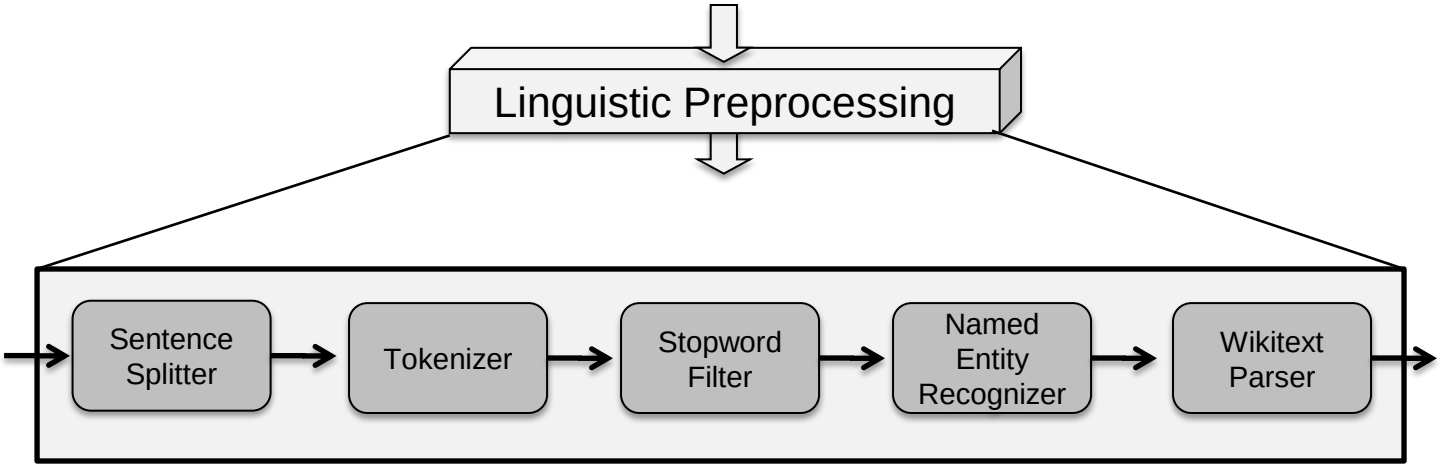
- Document retrieval via **Java Wikipedia Library** and **Wikipedia Revision Toolkit**
  - article text
  - revision history
  - revision meta data (authors, edit comment, timestamps)
  - links (in/out, internal/external)
- JWPL database based on Wikipedia data dump from January 4th, 2012.

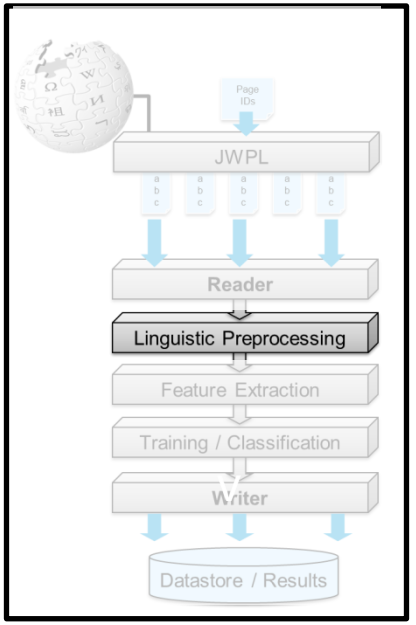


http://jwpl.googlecode.com

Preprocessing

- UIMA-based NLP components for preprocessing from the Darmstadt Knowledge Processing Repository



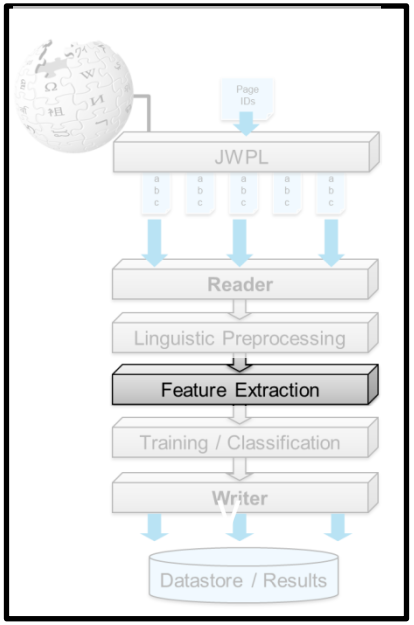


http://dkpro-core.googlecode.com

Features

- NGram features
- Structural features
- Reference features
- Network features
- Named entity features
- Revision-based features
- Other features

- 32 feature types in 7 categories
- ClearTK framework
  - „plug and play“ feature extractors
  - independent from utilized ML toolkit
- Information Gain approach for feature selection
- Unsupervised discretization of numeric features

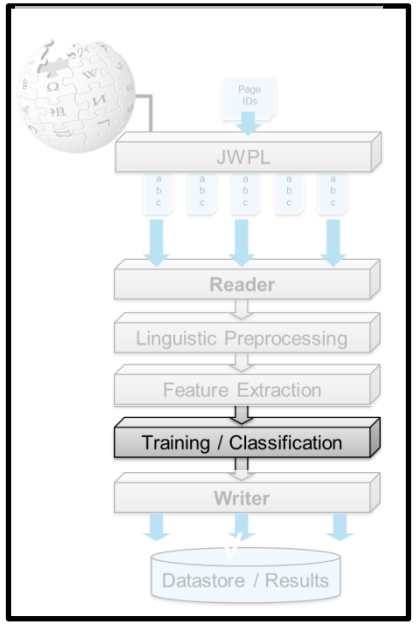


http://cleartk.googlecode.com

Classification Approach

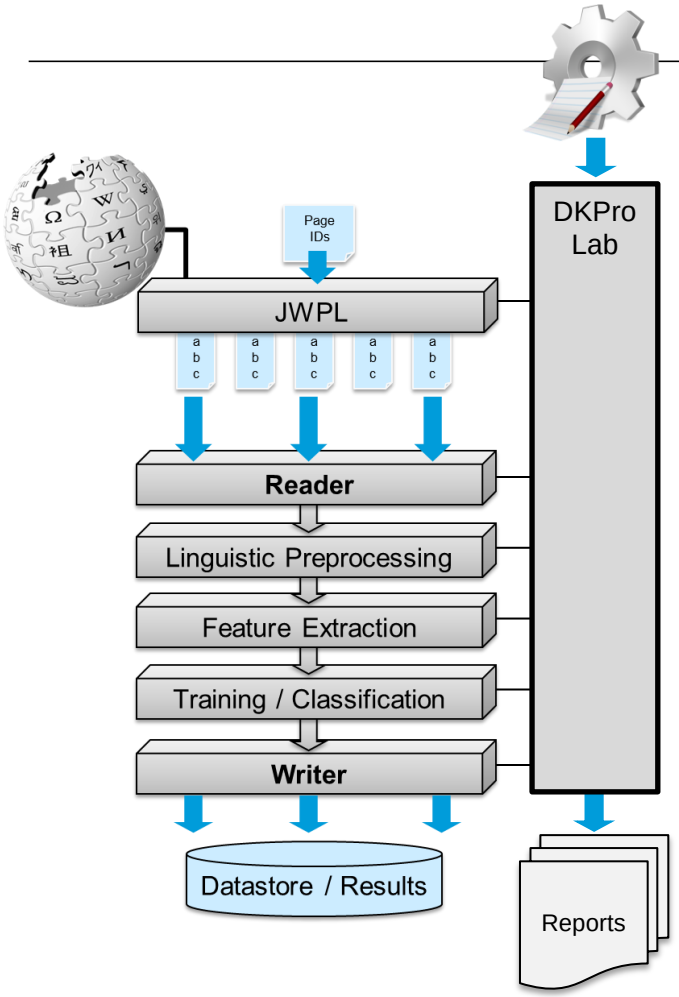
TECHNISCHE UNIVERSITÄT DARMSTADT

- Binary classification
  - Naive Bayes
  - AdaBoost with depth-limited C4.5 decision trees as weak classifiers
- Negative instances
  - Random sample of untagged articles
- Evaluation
  - 10-fold cross validation on 1000 documents
  - Stable sampling of negative instances in one evaluation run



MALLET
machine learning for language toolkit

Parameter Optimization



- The overall system is a „pipeline of pipelines“.
- Individual pipelines can be parameterized

Parameter optimization:

- Find best parameter setting across all pipelines
- Report on performance for pipeline configurations

DKPro Lab:

- Task based processing
- Parameter injection
- Global configuration
- Report probes gather statistics for global report

http://dkpro-lab.googlecode.com

Error Analysis and Evaluation

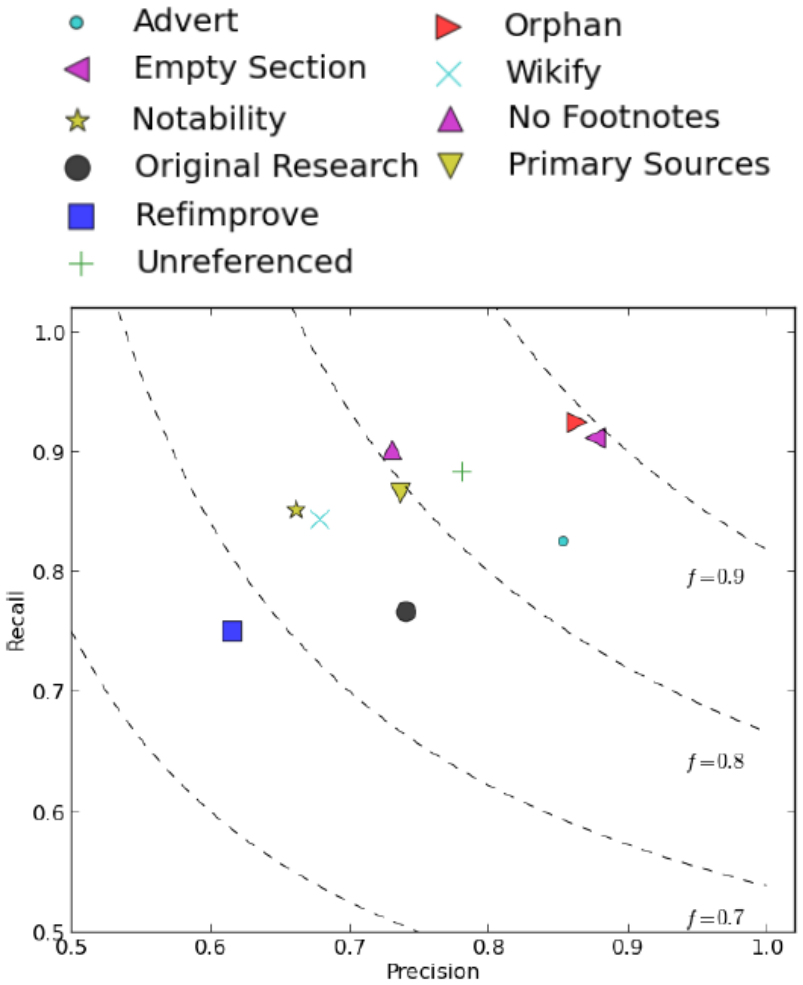
TECHNISCHE UNIVERSITÄT DARMSTADT

Common error sources

- Outdated labels (positive instances)
- Missing labels (negative instances)

- Unclear label definitions
  ➔ esp. reference flaws are often confused

- Section-scope and article-scope flaws mixed



UKP

Conclusions & Outlook

TECHNISCHE UNIVERSITÄT DARMSTADT

- Use article revision in which tag was first inserted
  - Solves outdated label problem
- Use revision history for identifying negative instances
  - Solves missing label problem
- Separate treatment of section- and article-scope templates
- Real world application: multi-flaw classification
  - problems with overlaps in flaw definitions

Thank you for your attention!

TECHNISCHE UNIVERSITÄT DARMSTADT

Ubiquitous Knowledge Processing Lab

LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz

DIPF
Educational Research and Educational Information

VolkswagenStiftung

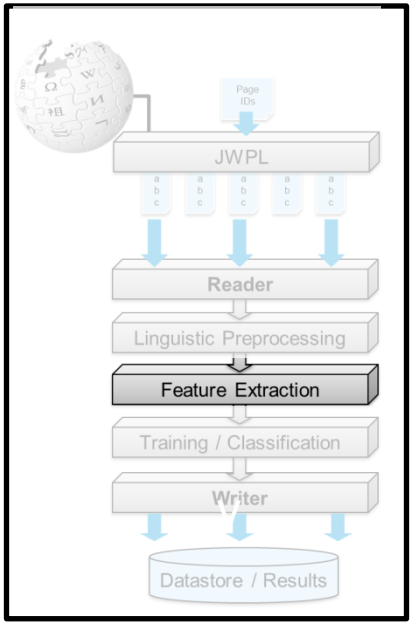
Emmy Noether-Programm
Deutsche Forschungsgemeinschaft DFG

http://www.ukp.tu-darmstadt.de

Features

- NGram features
- Structural features
- Reference features
- Network features
- Named entity features
- Revision-based features
- Other features

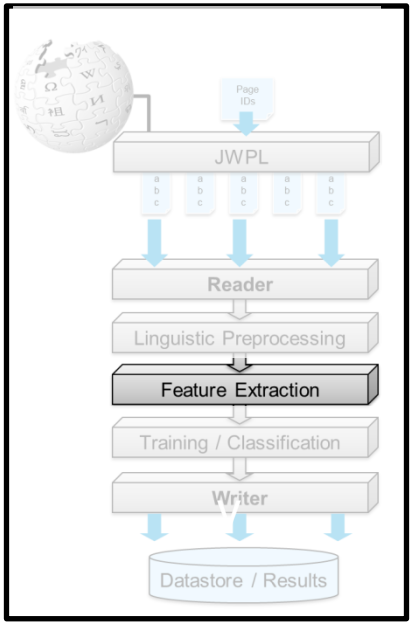
- Token-unigrams, bigrams, trigrams
- Extracted from article text w/o markup
- Min. frequency (5)
- Stopword filtered



Features

- NGram features
- Structural features
- Reference features
- Network features
- Named entity features
- Revision-based features
- Other features

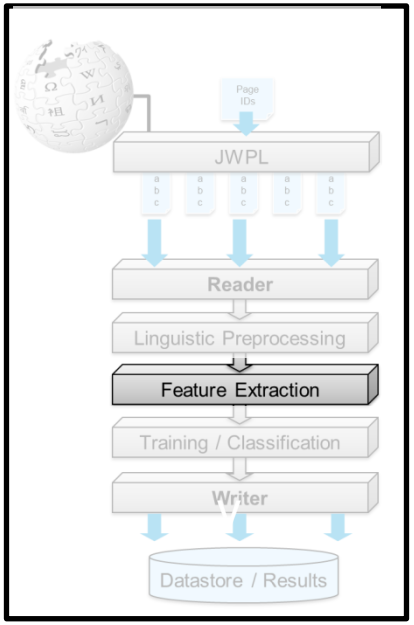
- Empty sections
- Number of sections
- Mean section length
- Markup to text ratio



Features

- NGram features
- Structural features
- Reference features
- Network features
- Named entity features
- Revision-based features
- Other features

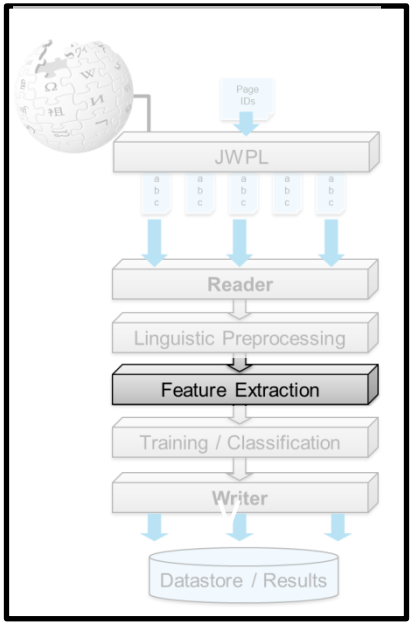
- Number of references
- Reference lists
- Reference to text ratio
- References per sentence



Features

- NGram features
- Structural features
- Reference features
- Network features
- Named entity features
- Revision-based features
- Other features

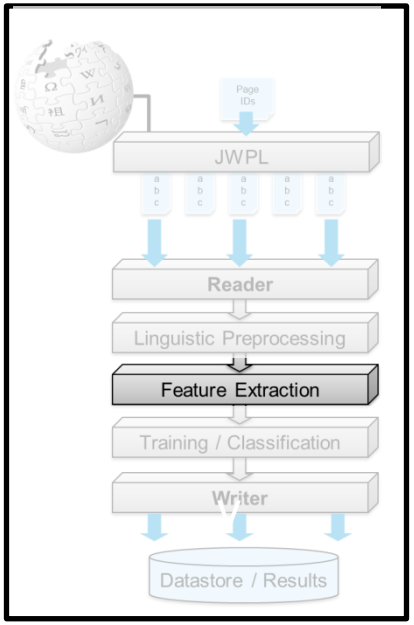
- External links
- Inlinks
- Outlinks



Features

- NGram features
- Structural features
- Reference features
- Network features
- Named entity features
- Revision-based features
- Other features

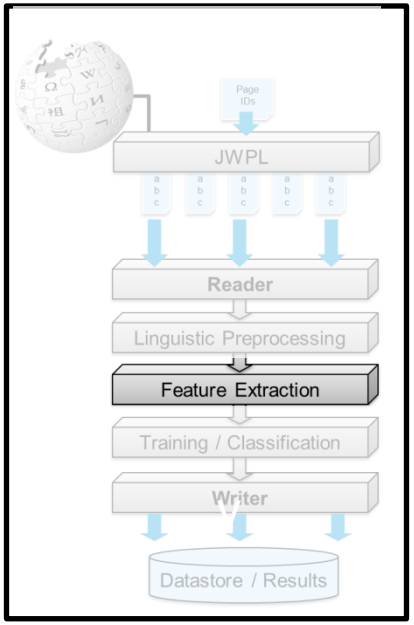
- NER types
  - Organization
  - Person
  - Location
- Absolute numbers and NER to text ratio



Features

- NGram features
- Structural features
- Reference features
- Network features
- Named entity features
- Revision-based features
- Other features

- Number of revisions
- Number of unique contributors
- Number of registered contributors
- Article age



Features

- NGram features
- Structural features
- Reference features
- Network features
- Named entity features
- Revision-based features
- Other features

- Number of discussions on Talk page
- Number of sentences, tokens and characters