FlawFinder
A Modular System for Predicting Quality Flaws in Wikipedia

Oliver Ferschke, Iryna Gurevych and Marc Rittberger
Introduction
Task-based system with Multiple processing pipelines.
Data Import

- 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)


[http://jwpl.googlecode.com](http://jwpl.googlecode.com)
Preprocessing

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

![Diagram of Linguistic Preprocessing]

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

- 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
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

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- Outdated labels (positive instances)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
Conclusions & Outlook

- 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  - Solves outdated label problem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!

Ubiquitous Knowledge Processing Lab

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