The Impact of Topic Bias on Quality Flaw Prediction in Wikipedia

Oliver Ferschke
Iryna Gurevych
Marc Rittberger

In a nutshell

Motivation
The growing amount of user generated content makes automatic quality assessment increasingly important.

In Wikipedia, cleanup templates are used to mark quality flaws in articles.
These templates can be used as training data for quality flaw prediction.

The distribution of these labels over all articles is not equal. It is biased towards particular topics.

Consequence: biased training data results in biased classifiers and overly optimistic cross-validated performance.

Contributions
1. We factor out the topic bias by mining “reliable training instances”
2. We compare the performance of biased and unbiased quality flaw classifiers.
3. We provide a set of free corpora of article revisions with and without neutrality and style flaws.

Quality Flaws in Wikipedia
- are identified by user-assigned cleanup templates.
- constitute violations of the quality standards defined in the Wikipedia Manual of Style and the featured/good article criteria.
- provide concrete insights into how an article can be improved.

Template Clusters
- sets of semantically similar or synonymous templates which represent the same quality flaw.
- e.g. copy-edit := {cleanup-english, copy edit, copy-edit, copyediting, gcheck, grammar}.

Template Scope
- article: marks whole article as flawed.
- section: marks single section as flawed.
- inline: marks flaw at a concrete position.

This work focuses on article scope templates.

Evaluation

Datasets
- BASE: latest positives + latest negatives.
- RELP: reliable positives + latest negatives.
- RELALL: reliable positives + reliable negatives.

Cross validation (best perf. setup)
- binary classification
- SVM with RBF kernel
- only n-gram features (n=1,2,3)

Flaw BASE RELP RELALL
Advert .86 .88 .75
POV .75 .80 .71
Globalize .85 .87 .69
Peacock .77 .82 .69
Weasel .69 .77 .72
Tone .76 .79 .69
In-universe .96 .96 .69
Copy-edit .83 .73 .72
Trivial .72 .77 .70
Essay-like .79 .83 .64
Confusing .76 .80 .70
Technical .67 .66 .67
Average .79 .83 .70

F scores, 10-fold CV, 2000 articles.

References

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