The Impact of Topic Bias on Quality Flaw Prediction in Wikipedia

Oliver Ferschke
Motivation

We are drowning in information and starving for knowledge.

–John Naisbitt
Motivation

A classifier is only as good as the data it is trained on.
Topic Bias

If ML features are topic-dependent variables, classifiers trained on multi-topic corpora will be topically biased.

- Gender and age identification
- Authorship attribution
- Native language detection
- Genre detection

⇒ Wikipedia Quality Flaw Detection
Quality Flaws in Wikipedia

- Cleanup templates are TODO-markers for authors

This article appears to be written like an advertisement. Please help improve it by rewriting promotional content from a neutral point of view and removing any inappropriate external links. (May 2013)

- Assumption: cleanup templates = quality flaw labels

- Overall ~400 different cleanup templates

# Neutrality and Style Flaws

## Neutrality

| Advert-like | NPOV | Globalize | Peacock | Weasel |

## Style

| Tone | In-universe | Copy-edit | Trivia | Essay-like | Confusing | Technical |

---

*Some people say that weasel words are great!*

*I am the greatest bird ever!*
Biased Template Distribution

Cleanup templates are not equally distributed over all articles

- Topical preference
  - Templates are primarily assigned to articles of certain topics

- Topical restriction
  - Templates can only be applied to articles of certain topics
Biased Template Distribution: Examples

- **in-universe**
  - restricted by definition
  - only applies to articles about fiction

- **advert**
  - topical preference by definition
  - more frequent than average on company and biography articles

- **copy-edit**
  - topical preference due to usage
  - no inherent reason for topic bias
  - but: used more often than average in specific WikiProjects, e.g. „linguistics“ and „law“
Consequences of Biased Distribution (1)

Data sampling for training corpora
Consequences of Biased Distribution (1)

Data sampling for training corpora

average cosine similarity of category frequency vectors

\[ \text{sim} \ (A_f, A_{md}) = 0.151 \]
Consequences of Biased Distribution (2)

average cosine similarity of category frequency vectors

\[
sim(A_f, A_{rand}) = 0.151
\]

\[
sim(A_f, A_{rel}) = 0.960
\]

\[\Rightarrow\] rule out the topic as the most discriminative factor
How should we sample the data?

http://www.sparkred.com
How to sample $A_{rel}$?

problem:
no articles are marked not to suffer from a certain flaw

solution:
- process all adjacent revision pairs of each article
  (WikiHadoop)
- find occurrences where cleanup templates have been removed
- assumption:
  if a cleanup template has been removed,
  the problem has been fixed

$\Rightarrow$ ignore unstable edits (vandalism, edit wars)
$\Rightarrow$ topic bias is automatically eradicated
Reliable Positive Training Instances \((A_f)\)

**simple approach**

extract all articles that are tagged with any template from the template cluster

\(\Rightarrow\) **problem**: outdated templates

**revision-based approach**

use the article revision in which the template has first been assigned
### Corpora

Three datasets for each flaw

- **BASE** positives (latest) + random negatives
- **RELP** reliable positives + random negatives
- **RELALL** reliable positives + reliable negatives

<table>
<thead>
<tr>
<th>Flaw</th>
<th>Positives</th>
<th>Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advert</td>
<td>7,332</td>
<td>39,133</td>
</tr>
<tr>
<td>Globalize</td>
<td>1,609</td>
<td>8,196</td>
</tr>
<tr>
<td>Peacock</td>
<td>1,195</td>
<td>7,022</td>
</tr>
<tr>
<td>POV</td>
<td>5,086</td>
<td>105,066</td>
</tr>
<tr>
<td>Weasel</td>
<td>704</td>
<td>12,710</td>
</tr>
<tr>
<td>Confusing</td>
<td>1,084</td>
<td>6,225</td>
</tr>
<tr>
<td>Copy-edit</td>
<td>1,954</td>
<td>2,878</td>
</tr>
<tr>
<td>Essay-like</td>
<td>1,244</td>
<td>3,898</td>
</tr>
<tr>
<td>In-Universe</td>
<td>2,227</td>
<td>5,270</td>
</tr>
<tr>
<td>Technical</td>
<td>690</td>
<td>2,056</td>
</tr>
<tr>
<td>Tone</td>
<td>4,563</td>
<td>20,166</td>
</tr>
<tr>
<td>Trivia</td>
<td>1,282</td>
<td>70,304</td>
</tr>
<tr>
<td>Σ</td>
<td>32,447</td>
<td>282,924</td>
</tr>
</tbody>
</table>

Experiments

- binary classification
- 2,000 documents per flaw
- 10-fold cross validation
- only n-gram features

- SVM with RBF kernel performed best among all tested algorithms

<table>
<thead>
<tr>
<th>Flaw</th>
<th>BASE</th>
<th>RELP</th>
<th>RELALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advert</td>
<td>.86</td>
<td>.88</td>
<td>.75</td>
</tr>
<tr>
<td>POV</td>
<td>.75</td>
<td>.80</td>
<td>.71</td>
</tr>
<tr>
<td>Globalize</td>
<td>.85</td>
<td>.87</td>
<td>.69</td>
</tr>
<tr>
<td>Peacock</td>
<td>.77</td>
<td>.82</td>
<td>.69</td>
</tr>
<tr>
<td>Weasel</td>
<td>.69</td>
<td>.77</td>
<td>.72</td>
</tr>
<tr>
<td>Tone</td>
<td>.70</td>
<td>.79</td>
<td>.69</td>
</tr>
<tr>
<td>In-universe</td>
<td>.96</td>
<td>.96</td>
<td>.69</td>
</tr>
<tr>
<td>Copy-edit</td>
<td>.81</td>
<td>.73</td>
<td>.72</td>
</tr>
<tr>
<td>Trivia</td>
<td>.72</td>
<td>.77</td>
<td>.70</td>
</tr>
<tr>
<td>Essay-like</td>
<td>.79</td>
<td>.83</td>
<td>.64</td>
</tr>
<tr>
<td>Confusing</td>
<td>.76</td>
<td>.80</td>
<td>.70</td>
</tr>
<tr>
<td>Technical</td>
<td>.87</td>
<td>.88</td>
<td>.67</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>.79</strong></td>
<td><strong>.83</strong></td>
<td><strong>.70</strong></td>
</tr>
</tbody>
</table>
Cross Corpus Analysis

- Highly ranked ngrams in the biased setup (BASE, RELP) are mainly topic related ngrams

- Cross corpus experiments
  - train on biased data, test on reliable data
    → classifier fails, because topic related ngrams are useless
  - train on reliable data, test on biased data
    → flaws with lexical cues: better performance, near cross validation
    → other flaws: no improvement
Conclusions and Future Work

- Topic bias causes overly optimistic evaluation results
- Classifiers are likely to identify articles prone to certain flaws, but not necessarily flawed articles
- Careful sampling can eradicate the topic bias

Directions for future improvement

- Break down quality flaw detection to section/sentence level
- Use commit comment as additional information source for determining reliable negatives („fixed issue“)
Thank you for your attention!

Ubiquitous Knowledge Processing Lab

Additional Online Material: http://www.ukp.tu-darmstadt.de/data/wiki-flaws/