The Impact of Topic Bias on Quality Flaw Prediction in Wikipedia

Oliver Ferschke

Oliver Ferschke, Iryna Gurevych and Marc Rittberger

51st Annual Meeting of the Association for Computational Linguistics
Sofia, Bulgaria, August 4–9, 2013

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

TECHNISCHE UNIVERSITÄT DARMSTADT

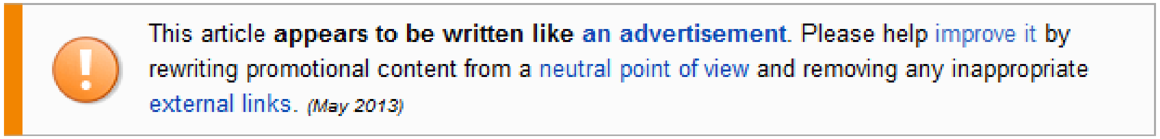
If ML features are topic-dependent variables, classifiers trained on multi-topic corpora will be topically biased.

- 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



- Assumption: cleanup templates = quality flaw labels

- Overall ~400 different cleanup templates

M. Anderka and B. Stein. A Breakdown of Quality Flaws in Wikipedia. Proc. of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality. P 11-18. 2012.

Neutrality and Style Flaws

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Neutrality	Style
Advert-like	Tone
NPOV	In-universe
Globalize	Copy-edit
Peacock	Trivia
Weasel	Essay-like
	Confusing
	Technical

Biased Template Distribution

Cleanup templates are not equally distributed over all articles

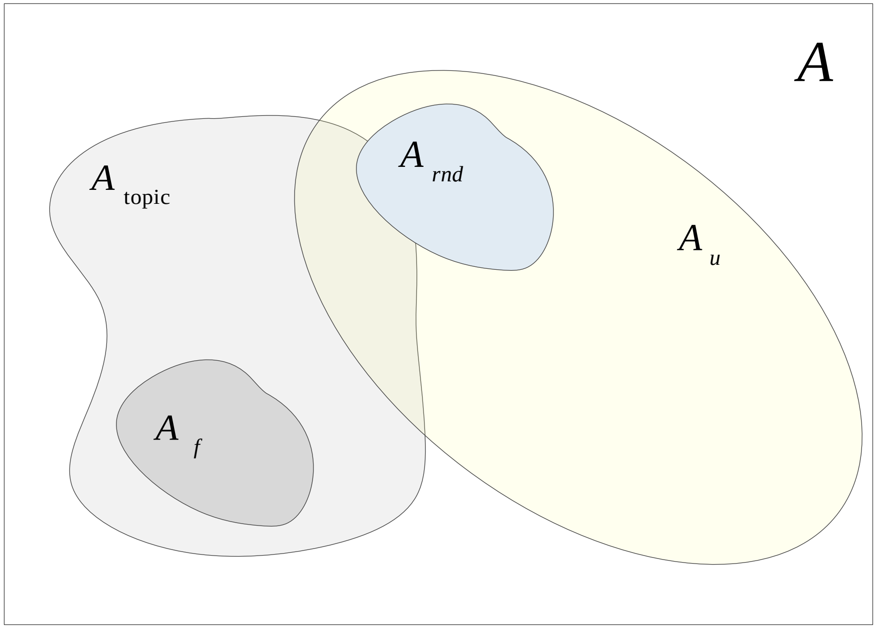
- Topical preference
  - Templates are primarily assigned to articles of certain topics
- Topical restriction
  - Templates can only be applied to articles of certain topics

Biased Template Distribution: Examples

TECHNISCHE UNIVERSITÄT DARMSTADT

- **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)



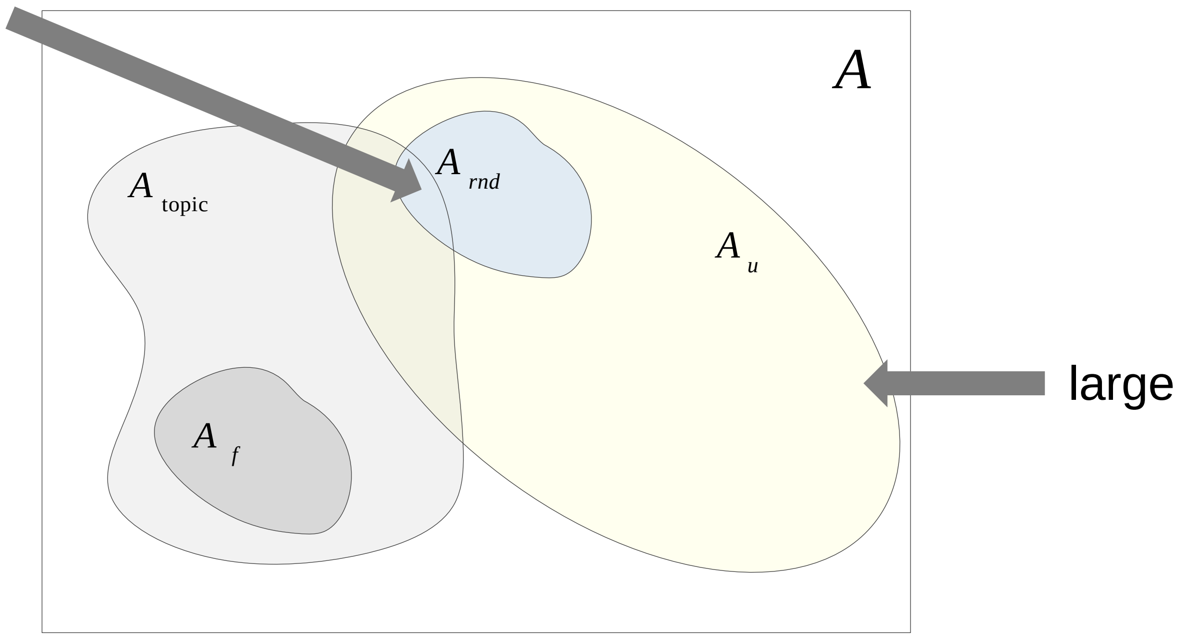
Data sampling for training corpora

Consequences of Biased Distribution (1)

small

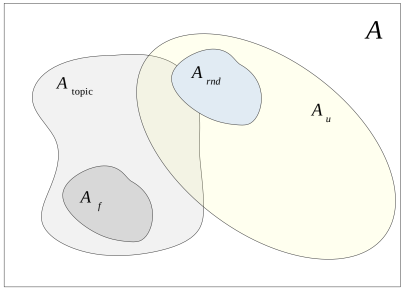
average cosine similarity
of category frequency vectors

$sim(A_f, A_{rnd}) = 0.151$



Data sampling for training corpora

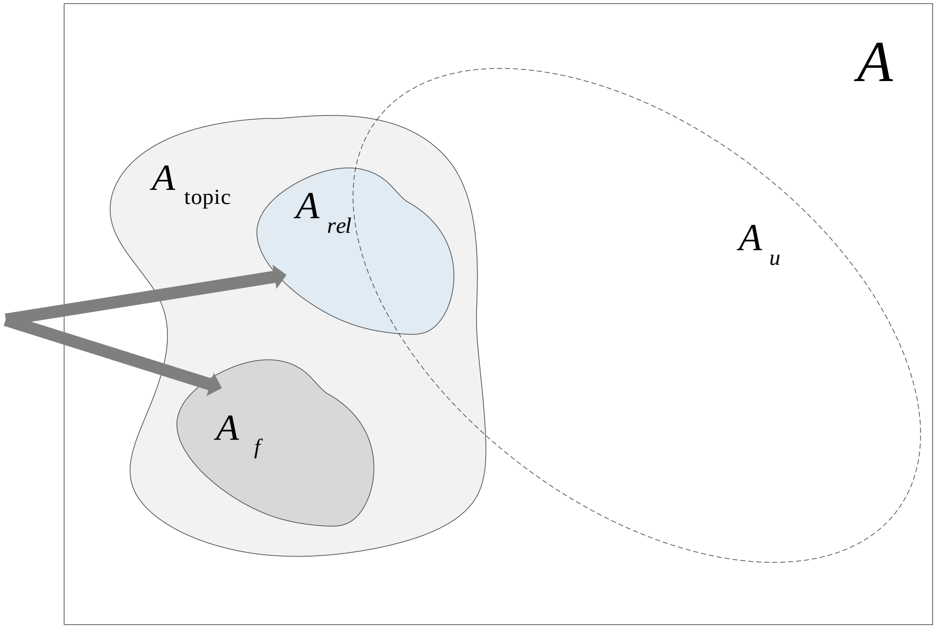
Consequences of Biased Distribution (2)



average cosine similarity
of category frequency vectors

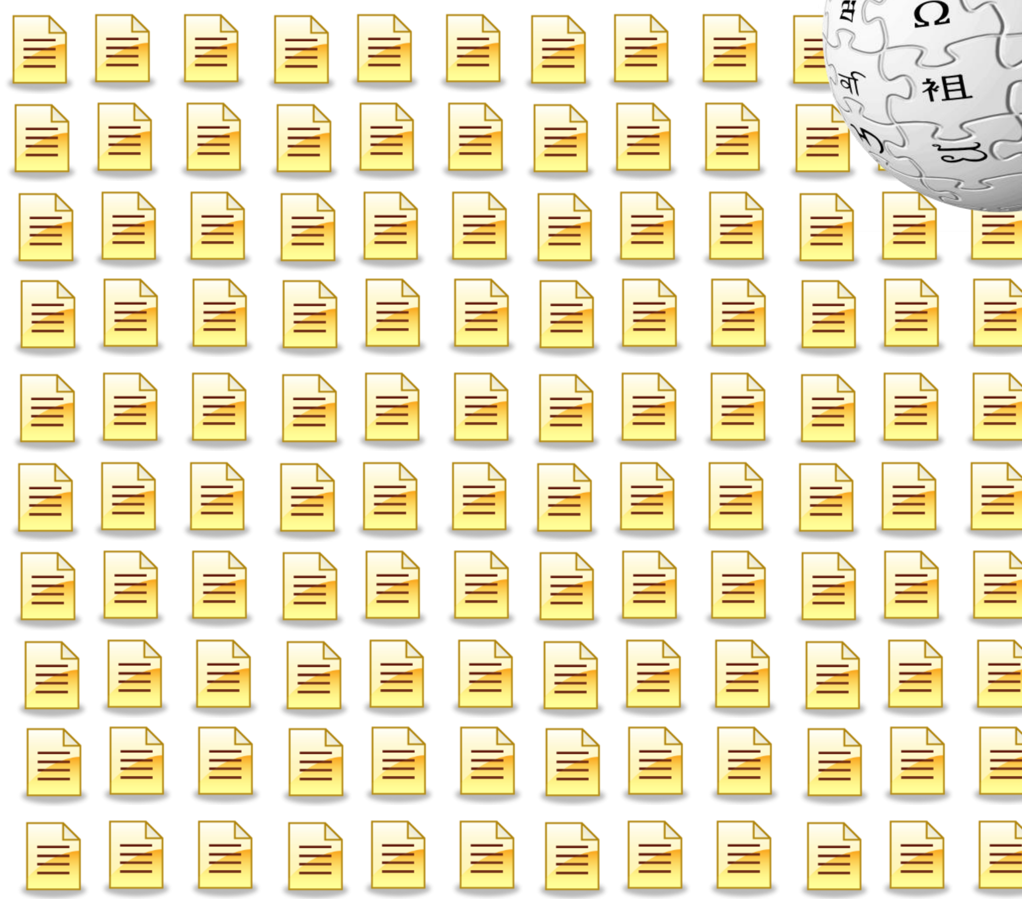
$sim(A_f, A_{rnd}) = 0.151$

$sim(A_f, A_{rel}) = 0.960$



➔ rule out the topic as as the most discriminative factor

How should we sample the data?



http://www.sparkred.com

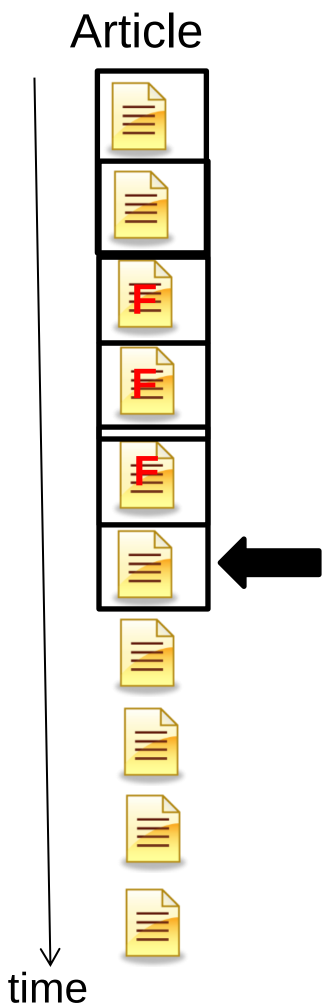
How to sample A_{rel} ?

problem:

no articles are marked not to suffer from a certain flaw

solution:

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



➔ignore unstable edits (vandalism, edit wars)

➔topic bias is automatically eradicated

Reliable Positive Training Instances (A_f)

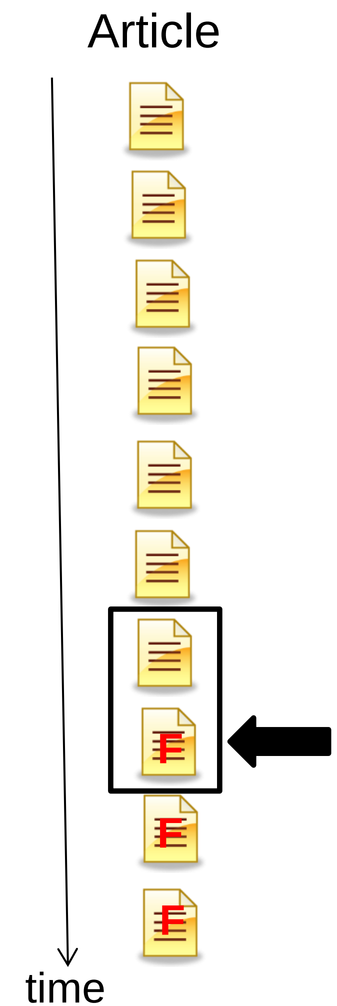
simple approach

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

➔ *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

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

Download: `http://www.ukp.tu-darmstadt.de/data/wiki-flaws/`

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

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	.70	.79	.69
In-universe	.96	.96	.69
Copy-edit	.81	.73	.72
Trivia	.72	.77	.70
Essay-like	.79	.83	.64
Confusing	.76	.80	.70
Technical	.87	.88	.67
Average	**.79**	**.83**	**.70**

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!

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Ubiquitous Knowledge Processing Lab

LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz

VolkswagenStiftung

DIPF

Educational Research
and Educational Information

Emmy Noether-Programm

Deutsche Forschungsgemeinschaft
DFG

Additional Online Material:

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