The Quality of Massive Open Online Collaboration

Oliver Ferschke

Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

UBIQUITOUS KNOWLEDGE PROCESSING

In a Nutshell

Motivation
- The growing amount of user generated content makes automatic quality assessment increasingly important
- Wikipedia is one of the largest open collaborative efforts with information quality being a key aspect

Goals
- Develop an information quality model for Wikipedia articles
- Develop techniques for automatically measuring and managing quality
- Account for the collaborative and dynamic nature of Wikipedia both in the model and in the methods

Contributions
- Adaptation of a generic information quality model to Wikipedia based on an analysis of Wikipedia's collaboratively defined quality assessment policies
- Discourse analysis of Wikipedia article Talk pages to identify quality problems, proposed solutions and key personnel
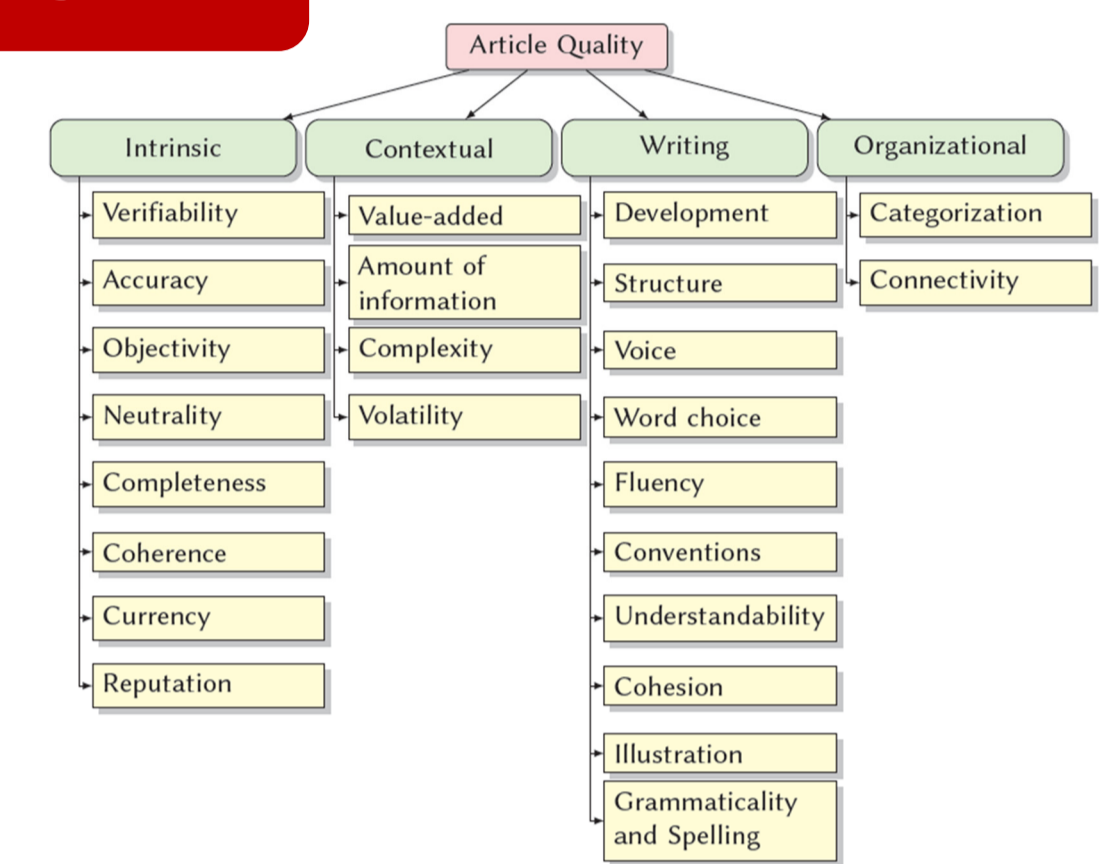- Quality flaw detection in articles based on cleanup template prediction

Quality of Wikipedia

Quality Standards in Wikipedia
- collaboratively defined
- change over time
- contextualization over formalization
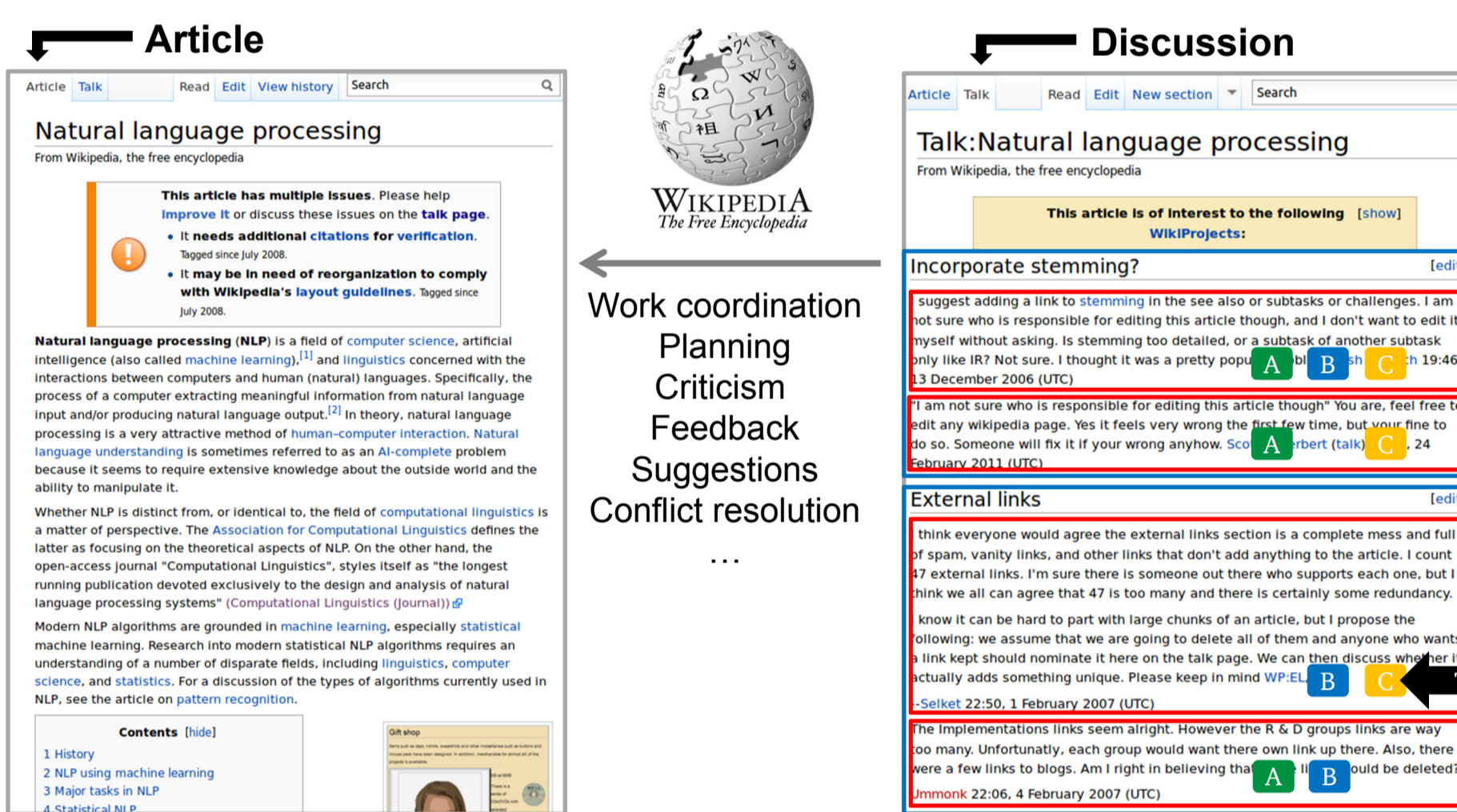- fragmented and maintained by individual subgroups

Quality Guidelines and Mechanisms
- Featured/good article criteria
- Wikipedia Manual of Style
- User Article Ratings
- WikiProject Quality Assessments
- Talk Pages

An Information Quality Model for Wikipedia Articles

User Discourse Analysis

Article

Discussion

Work coordination
Planning
Criticism
Feedback
Suggestions
Conflict resolution
...

Discussion/Topic

Turn

Turn

Turn-level dialog acts

Dialog Segmentation

Problem
- Unstructured dialog without explicit markup
- Free-edit policy causes entangled threads
- User signatures unreliable

Solution
- Backtrack revision history to reconstruct the course of the discussion
- Obtain user identity via revision meta data
- Entangled threads can be resolved

Classification scheme for capturing coordination efforts on article improvement
- **Article Criticism** (7 classes)
  quality problems addressed in discussion
- **Explicit Performatives** (4 classes)
  action proposed to solve quality problem
- **Information Content** (3 classes)
  information providing/seeking/changing
- **Interpersonal** (3 classes)
  sentiment expressed towards other participants of the discussion

Automatic Dialog Act Tagging

Goal: identify article quality problems and proposed solutions discussed on article Talk pages

Approach: text classification models trained on manually annotated discussions

Results
- Average human inter-rater agreement: κ=.67
- Ensemble classification system (SVM|NBayes|J48) with overall performance of F1=.74

Quality Flaw Detection

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 how an article can be improved

Quality Flaw Detection

Goal: identify whether an unseen article suffers from a particular quality flaw

Approach: text classification models trained on flawed (positive) and unflawed (negative) articles

Challenge: sampling of adequate positive and negative training instances

Sampling of Training Data

Problem:
1. no articles are marked as not to suffer from a particular flaw
2. we need negative instances from the same topic set as the positive examples to avoid a topic bias

Approach
- find adjacent article revision pairs in which the target cleanup template has been removed
- assumption: if a cleanup template has been removed, the problem has been fixed
- filter vandalism and edit wars

article #1 | article #2

untagged revision
current revision
flawed revision

reliable negative

reliable positive

latest negative

latest positive

time

	Flaw	Description
Neutrality	Advert	Appears to be written like an advertisement and is thus not neutral
	POV	The neutrality is disputed (non-neutral point of view)
	Globalize	May not represent a worldwide view of the subject
	Peacock	May contain wording that promotes the subject without imparting verifiable information
	Weasel	Contains vague phrasing that often accompanies biased or unverifiable information
Style	Tone	The tone of the article is not encyclopedic according to the Wikipedia Manual of Style
	In-universe	Describes a work or element of fiction in a primarily in-universe style
	Copy-edit	Requires copy editing for grammar, style, cohesion, tone, or spelling
	Trivia	Contains lists of miscellaneous information
	Essay-like	The article is written like a personal reflection or essay
	Confusing	The article may be confusing or unclear to readers
	Technical	The article may be too technical for most readers to understand

Table: Example set of neutrality and style flaws used in our experiments

Results
- Average cross-validated performance on topically biased data F1=0.83
- Reliable negatives reduce cross-validated performance (0.70) due to a reduction of the topic bias, but they help to better model the actual quality problems
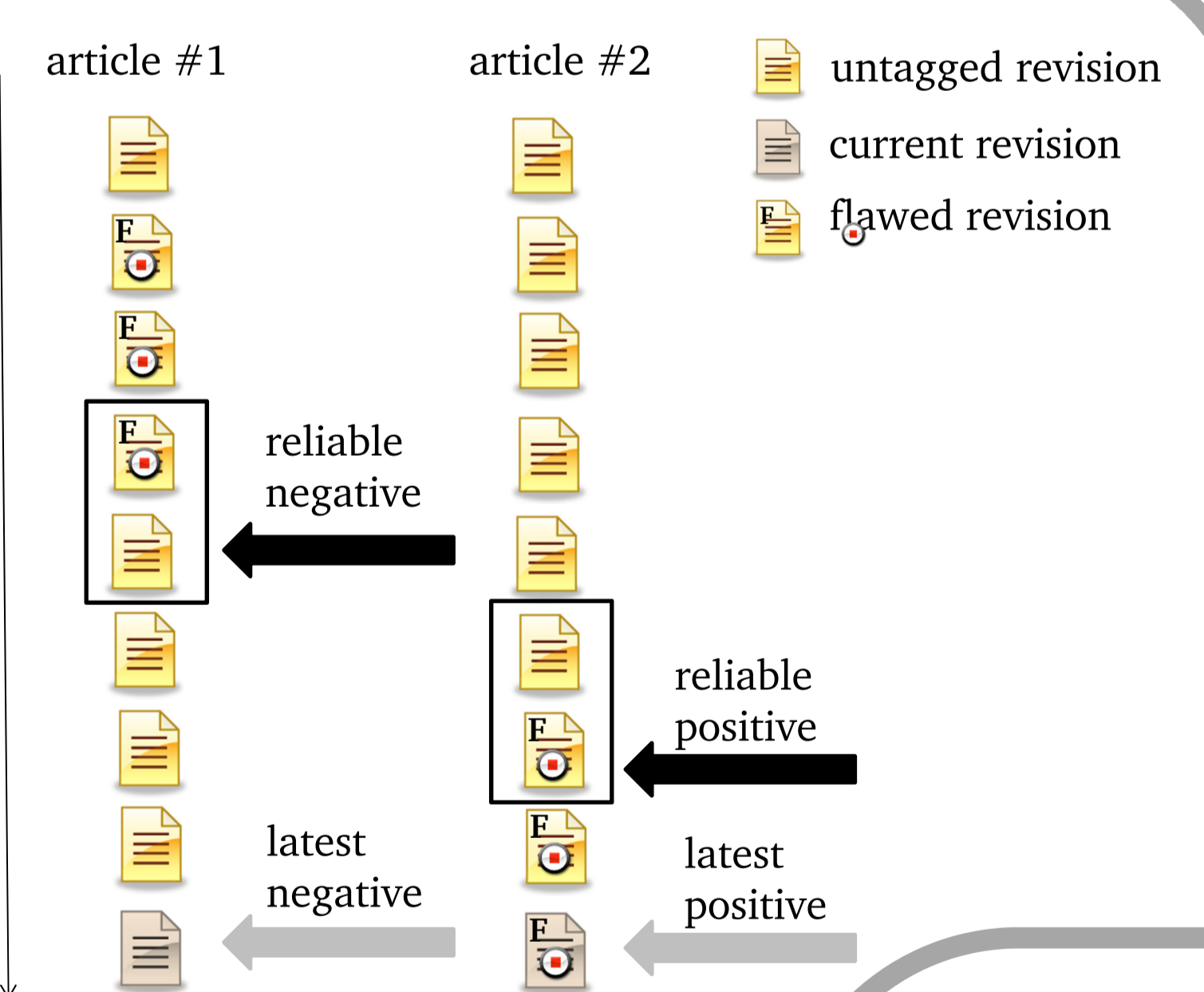
JWPL

jwpl.googlecode.com

References

- O. Ferschke, I. Gurevych, and M. Rittberger, 2013. **The Impact of Topic Bias on Quality Flaw Prediction in Wikipedia**. *In Proceedings of the 51st Annual Meeting of the Association of Computational Linguistics: Human Language Technologies*. Sofia, Bulgaria.
- O. Ferschke, J. Daxenberger, and I. Gurevych, 2013. **A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia**. *In Iryna Gurevych and Jungi Kim: The People's Web Meets NLP: Collaboratively Constructed Language Resources.*.
- O. Ferschke, I. Gurevych, and M. Rittberger, 2012. **FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia – Notebook for PAN at CLEF 2012**. *In CLEF 2012 Labs and Workshop, Notebook Papers*. Rome, Italy.
- O. Ferschke, I. Gurevych, and Y. Chebotar, 2012. **Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages**. *In Proceedings of the 13th Conference of the European Chapter of the ACL*. Avignon, France.
- O. Ferschke, T. Zesch, and I. Gurevych, 2011. **Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History**. *In Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies. System Demonstrations*. Portland, OR, USA.

LOEWE

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

DIPF
Educational Research and Educational Information