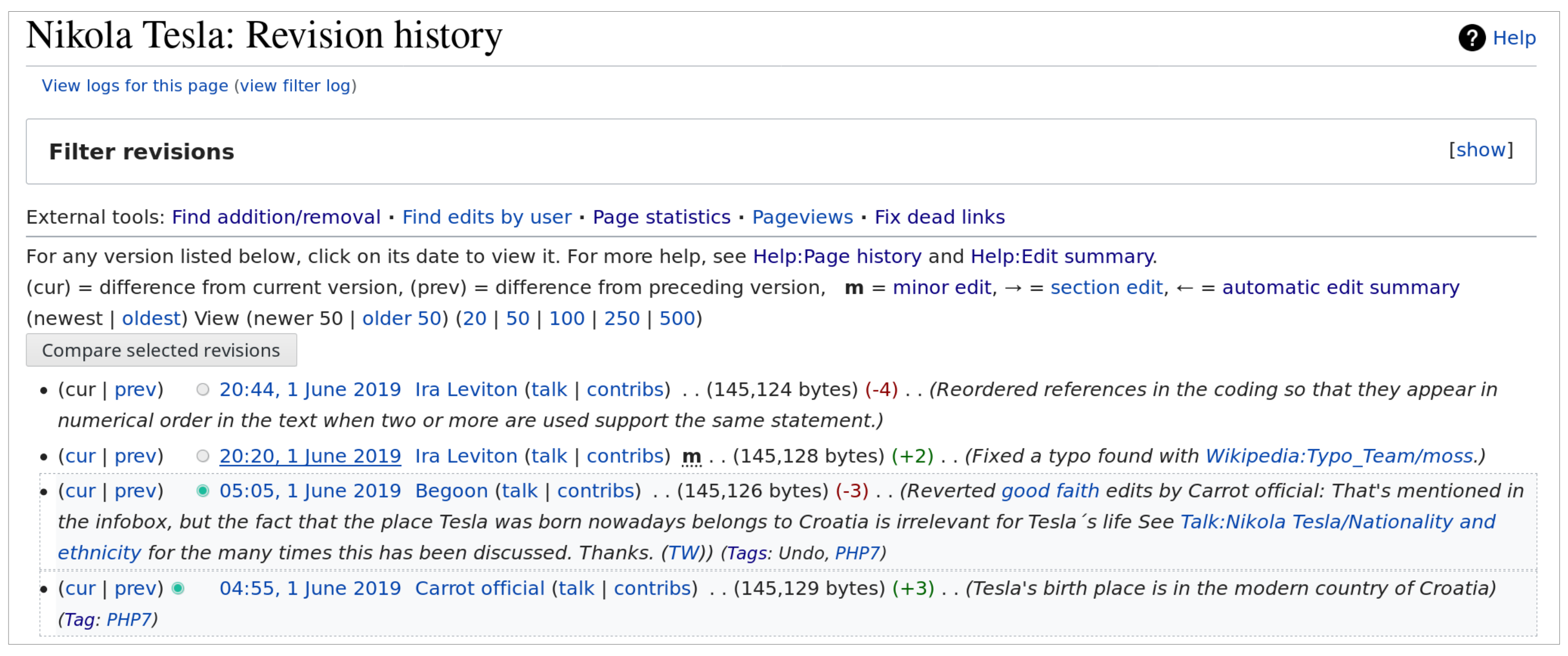
Automatic Semantic Categorization of Wikipedia Revisions

Taxonomy-Based Multi-Label Classification for Article Edits

Thorsten Ruprechter (ruprechter@tugraz.at)
Denis Helic (dhelic@tugraz.at)

Motivation

The online encyclopedia Wikipedia has been a rich resource for scientific research for many years. It supplies data for various application fields, such as natural language processing, time series analysis, or user interaction studies. For these topics, Wikipedia's article revision history provides access to a wide range of features (see Figure 1). However, many approaches require categorization of which goal a user tries to achieve when editing an article. Although algorithms considering syntactic changes of revisions are quite prevalent, only sparse research into automatically classifying semantic purposes of revisions exist. The lack of large, semantically labeled Wikipedia edit corpora is especially notable. Although some authors use small, manually labeled sets of revisions for their research, no automatic semantic classification of revision data seems to be feasible at the moment.



Nikola Tesla: Revision history

Help

View logs for this page (view filter log)

Filter revisions [show]

External tools: Find addition/removal · Find edits by user · Page statistics · Pageviews · Fix dead links

For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary.
(cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary
(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) ○ 20:44, 1 June 2019 Ira Leviton (talk | contribs) . . (145,124 bytes) (-4) . . *(Reordered references in the coding so that they appear in numerical order in the text when two or more are used support the same statement.)*
- (cur | prev) ○ 20:20, 1 June 2019 Ira Leviton (talk | contribs) **m** . . (145,128 bytes) (+2) . . *(Fixed a typo found with Wikipedia:Typo_Team/moss.)*
- (cur | prev) ● 05:05, 1 June 2019 Begoon (talk | contribs) . . (145,126 bytes) (-3) . . *(Reverted good faith edits by Carrot official: That's mentioned in the infobox, but the fact that the place Tesla was born nowadays belongs to Croatia is irrelevant for Tesla´s life See Talk:Nikola Tesla/Nationality and ethnicity for the many times this has been discussed. Thanks. (TW)) (Tags: Undo, PHP7)*
- (cur | prev) ● 04:55, 1 June 2019 Carrot official (talk | contribs) . . (145,129 bytes) (+3) . . *(Tesla's birth place is in the modern country of Croatia) (Tag: PHP7)*

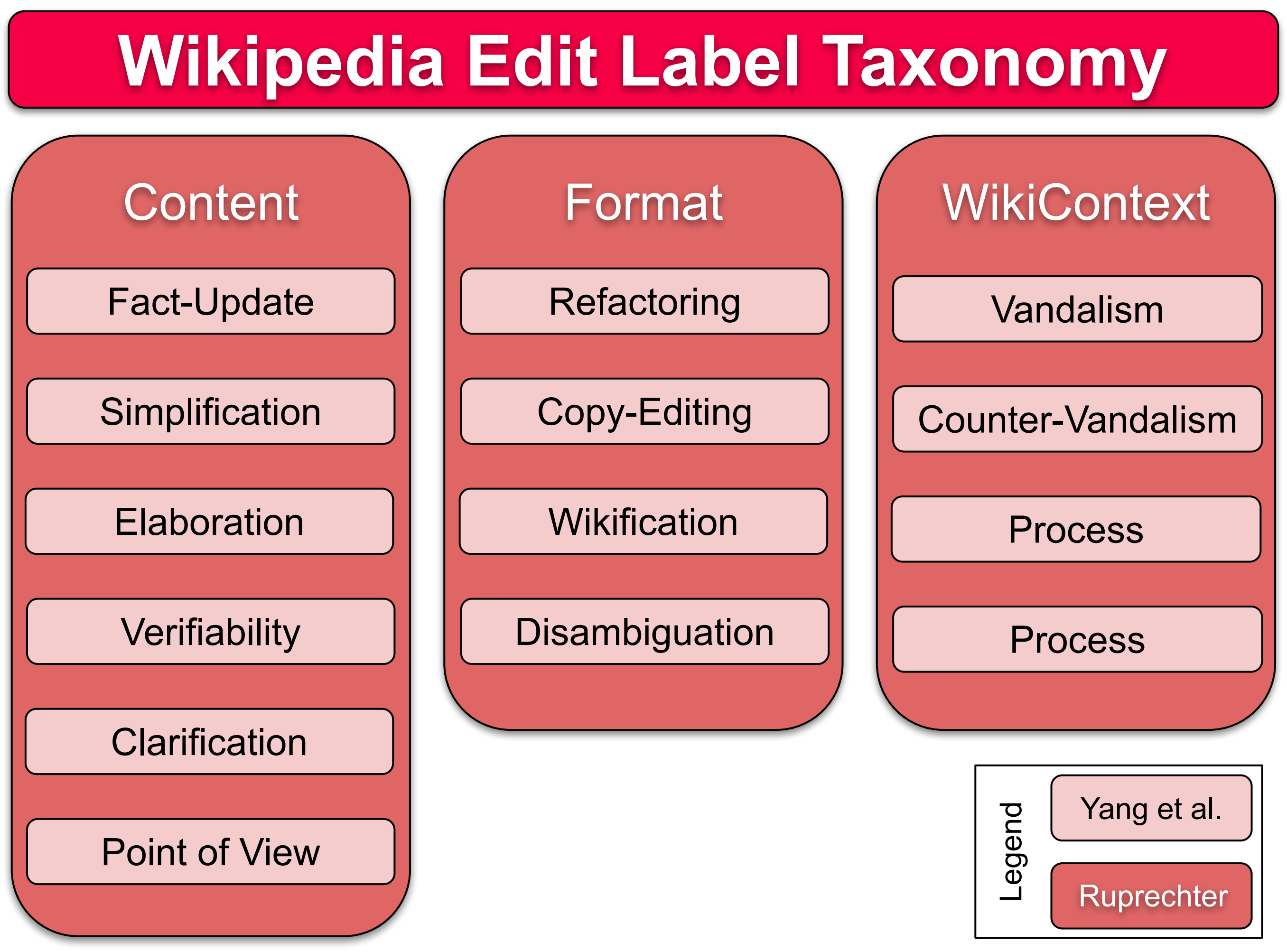
Figure 1: Excerpt of the revision history for the Wikipedia article about "Nikola Tesla". The highlighted revisions are part of an edit war about his nationality.

Research Problem and Contributions

- Define a simplified semantic taxonomy for Wikipedia edit classification
- Utilize supervised learning to train a multi-label classifier
- Classify revisions of articles belonging to different quality categories
- Evaluate classification by anaylzing transitions between specific edit labels

Edit Label Taxonomy

To enable automatic labeling of article edits, we derive 3 superlabels from the 14-label taxonomy introduced by Yang et al[1]: *Content*, *Format*, and *WikiContext*. First, *Content* captures all user revisions aiming towards extending or reducing actual information on the article page. Secondly, *Format* describes all edits modifying text without changing its meaning, facts, or contained information. Lastly, *WikiContext* includes all meta-interactions possible in Wikipedia articles, such as tagging pages, vandalism, counter-vandalism, or others. Figure 2 shows how we aggregated Yang's taxonomy into three superlabels for this work.



Wikipedia Edit Label Taxonomy

Content: Fact-Update, Simplification, Elaboration, Verifiability, Clarification, Point of View

Format: Refactoring, Copy-Editing, Wikification, Disambiguation

WikiContext: Vandalism, Counter-Vandalism, Process, Process

Legend: Yang et al., Ruprechter

Figure 2: New Wikipedia edit label taxonomy based on taxonomy by Yang et al[1].

Dataset and Experimental Setup

- Retrieve Dataset provided by Yang et al.[1] (5777 revisions, 14 labels)
- Upsample *WikiContext* category by manually pulling *Vandalism* and *Counter-Vandalism* revisions using *Wikimedia API*[2] ⇒ ~7000 revisions
- Convert upsampled dataset from 14 to 3 categories
- Utilize feature set provided by Yang[1], retrieve features using *RevScoring*[3]
- Train Random Forest utilizing *scikit-learn*[4] (weighted F1-score: 0.812)
- Process revision history of 100 articles per article quality (see Figure 3): *Featured*, *A-class*, *Good*, *B-class*, *C-class*, *Edit Wars/Conflicted*[5,6,7]
- Automatic multi-label classification of revisions for all retrieved articles
- Investigate transitions between labels to detect patterns and evaluate classifier according to basic assumptions

Quality	Importance					
	Top	Mid	High	Low	None	Total
Featured	1,303	2,001	1,927	1,294	180	**6,705**
Featured Lists	146	556	620	529	105	**1,956**
A-class	253	476	635	422	88	**1,874**
Good Articles	2,432	5,449	10,772	12,315	1,791	**32,759**
B-class	13,171	25,172	38,952	34,222	15,590	**127,107**
C-class	12,118	36,299	82,764	124,305	53,135	**308,621**
Start	18,249	83,129	342,757	994,293	354,500	**1,792,928**
Stub	4,345	32,029	247,172	2,184,242	879,005	**3,346,793**
List	3,571	12,950	40,284	117,978	74,225	**249,008**
Assessed	55,588	198,061	765,883	3,469,600	1,378,619	**5,867,751**
Unassessed	125	535	1,745	17,039	490,496	**509,940**
Total	**55,713**	**198,596**	**767,628**	**3,486,639**	**1,869,115**	**6,377,691**

Table 1: All rated articles by quality and importance as of 7th June 2019[8].

Results and Discussion

Figure 2 shows the label transition frequencies for all investigated article categories. For higher-quality content such as *Featured* or *A-class* articles, it seems that most edit sequences contain multiple consecutive *Format* actions. These results support the hypothesis that well-formatted and structured information is necessary for top-quality Wikipedia articles. On the contrary, articles which are subject to controversy and edit wars more commonly iterate between *Content* and *WikiContext* respectively. Information and facts in these articles appear to be more disputed than in well-polished ones.

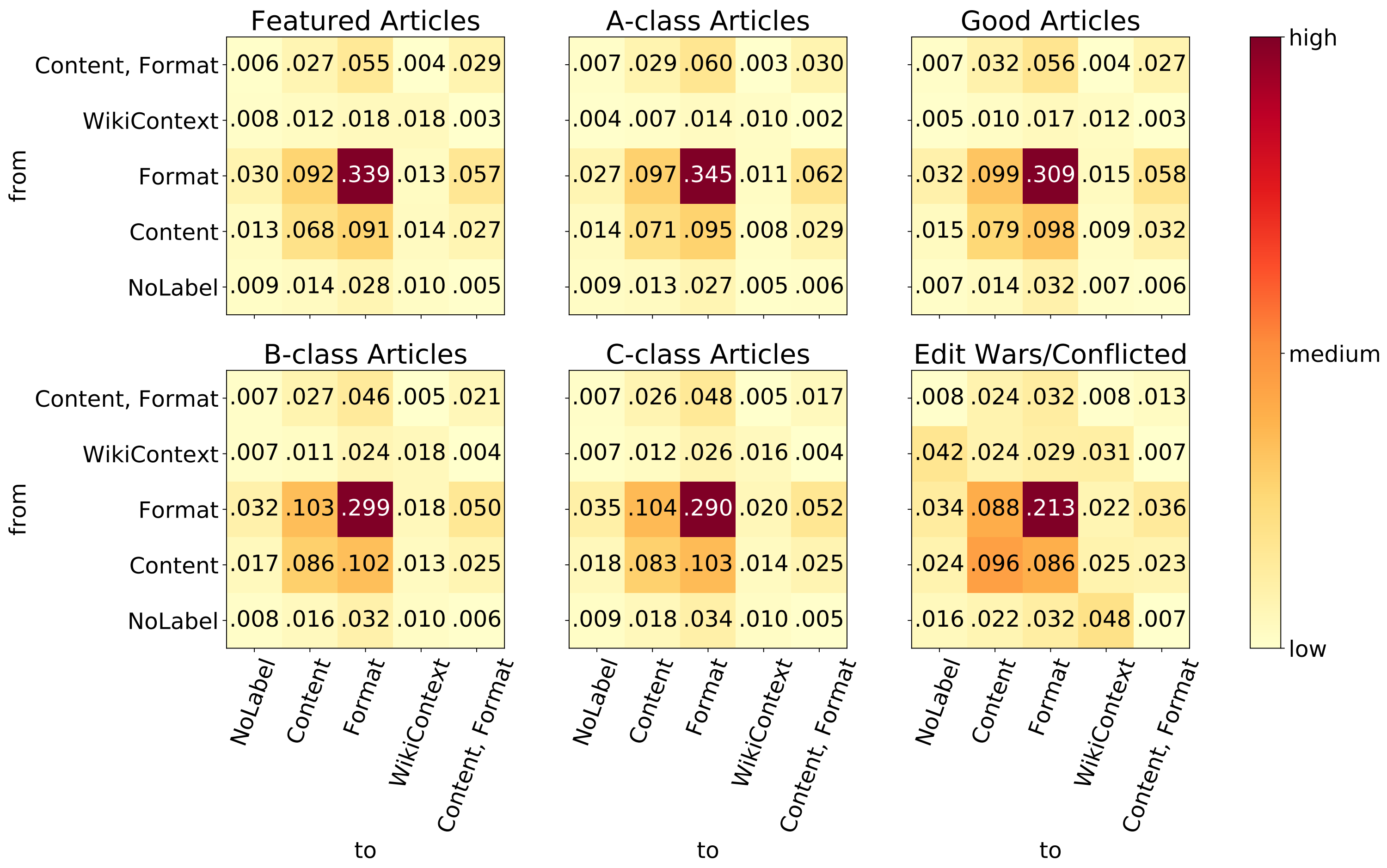


Figure 3: Heatmaps visualizing edit label transition frequencies by article category.

References

1 Yang, D., Halfaker, et al. (2017). Identifying semantic edit intentions from revisions in wikipedia.
2 mediawiki.org/wiki/API
3 pythonhosted.org/revscoring
4 Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python
5 en.wikipedia.org/wiki/Wikipedia:Lamest_edit_wars
6 Yasseri, T., Spoerri, A., Graham, M., & Kertész, J. (2014). The Most Controversial Topics in Wikipedia.
7 Flöck, F., Erdogan, K., & Acosta, M. (2017). TokTrack: A complete token provenance and change tracking dataset for the english Wikipedia.
8 en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Statistics

Graz University of Technology
Institute of Interactive Systems and Data Science
Inffeldgasse 16c/I
8010 Graz
Austria