Crowdsourcing a Wikipedia Vandalism Corpus

Martin Potthast

Bauhaus-Universität Weimar
99421 Weimar, Germany

martin.potthast@uni-weimar.de

ABSTRACT

We report on the construction of the PAN Wikipedia vandalism corpus, PAN-WVC-10, using Amazon's Mechanical Turk. The corpus compiles 32 452 edits on 28 468 Wikipedia articles, among which 2 391 vandalism edits have been identified. 753 human annotators cast a total of 193 022 votes on the edits, so that each edit was reviewed by at least 3 annotators, whereas the achieved level of agreement was analyzed in order to label an edit as "regular" or "vandalism." The corpus is available free of charge.[1]

Categories and Subject Descriptors: H.3.4 [Information Storage and Retrieval]: Systems and Software—*Performance Evaluation*

General Terms: Experimentation

Keywords: Wikipedia, Vandalism Detection, Evaluation, Corpus

1. INTRODUCTION

Wikipedia is an encyclopedia written by the crowd. The key to Wikipedia's success is a collaborative writing process, where everybody can edit every article. Ideally, the reader of an article also revises it to the best of her abilities, e.g. by correcting errors, by improving the writing style, by adding missing information, or by removing redundancy. In this way Wikipedia's articles get continuously improved and updated. This "freedom of editing" gave the lie to those who suggested that the resulting articles would be characterized by poor quality and instability. Wikipedia thrives. There is no free lunch, however, and Wikipedia faces problems that limit its growth, such as vandalism, edit wars, and lobbyism. Our concern is the automatic detection of vandalism in Wikipedia, i.e., the detection of edits that were made with bad intentions. We contribute to this research field by developing a large corpus of human-annotated edits, which is a prerequisite for the meaningful evaluation of vandalism detection algorithms. In particular, we report on our efforts to use Amazon's Mechanical Turk as a possibility to drive the corpus size to the necessary order of magnitude without compromising the corpus quality.

Related Work. Although vandalism has been observed in Wikipedia right from the start, and, although vandalism is often deemed one of Wikipedia's biggest problems, research has addressed automatic vandalism detection only recently—for the first time in [3, 5, 7]. Vandalized articles often get restored rather quickly by other editors, but still, the authors of [6] find that the number of times vandalized articles get viewed amounts up to hundreds of millions, and that the probability of encountering vandalism grew exponentially between 2003 and 2006. In reaction to this development, the Wikipedia community has developed a number of rule-based robots that are capable of restoring the most obvious cases of vandalism, or that aid editors to do so [2]. However, the performance of the robots is surpassed, for instance, by an approach based on machine learning [5]. Other reactions include the temporary suspension of the freedom of editing for articles that are often vandalized, which threatens the very idea of Wikipedia.

The first vandalism corpus was the Webis-WVC-07, which consists of 940 human-annotated edits of which 301 are vandalism [4]. The PAN-WVC-10 is two orders of magnitude larger and has been annotated by many different people; it thus forms a more representative sample of vandalism and allows for better estimates of whether a vandalism retrieval model will actually work in practice. In this respect, the Mechanical Turk provides an exciting new way to scale up corpus construction, which has also been applied successfully, e.g., to recreate TREC assessments [1].

2. CORPUS DESIGN

Corpus Layout. An edit marks the transition from one article revision to another. On Wikipedia, each revision of every article is accessible by means of a permanent identifier, so that an edit is described uniquely by a pair of revision IDs referencing the old article revision and the new revision.[2] Basically, our corpus is a list of revision ID pairs along with labels whether or not the respective edit is vandalism. Moreover, for each edit meta information is given as well as the plain texts of both the old and the new article revision.

Corpus Acquisition. Our sample of edits is drawn from the revision histories of Wikipedia articles by means of probability proportional to size sampling, where in our case, the "size" of an article is the average number of times it gets edited in a given time frame. We hypothesize that the average edit ratio of an article correlates with the number of times it gets viewed. In that case, our edit sample resembles well the distribution of article importance at the time of sampling, which presumably also influences the articles chosen by vandals. By contrast, the edits of the Webis-WVC-07 were chosen in search for vandalism from articles whose topics, per se, have a high conflict potential, which reveals a sample bias of that corpus.

Corpus Annotation. Amazon's Mechanical Turk is a platform for paid crowdsourcing. It acts as an intermediary between workers and so-called requesters who offer tasks and a reward for each task being solved. Typically, task assignment and result submission is handled double-blind. This sense of anonymity and the fact that real money can be earned tempts some workers to fake results in

[1] Download the corpus from http://www.webis.de/research/corpora

[2] Here is an example for a vandalism edit, shown as Wikipedia Diff page: http://en.wikipedia.org/w/index.php?diff=327907617&oldid=327774745

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

Table 1: Re-annotation of the Webis-WVC-07 corpus.

	3 Annotators / Edit		16 Annotators / Edit	
Agreement with	3 agree	56 %	more than 2/3 agree	93 %
Webis-WVC-07	3 disagree	2 %	more than 2/3 disagree	1 %
(Gold Standard)	2 agree	36 %	tie majority agrees	0 %
	2 disagree	6 %	tie majority disagrees	6 %
Accuracy	if 3 agree	96 %	if more than 2/3 agree	99 %
Baseline (all edits regular)		68 %		68 %

order to get paid without working. Requesters therefore may approve or reject results while workers are paid only if their results are approved—which in turn of course tempts requesters to reject acceptable results to save the money. From this it becomes clear that requesters need to analyze the results obtained via the Mechanical Turk to sort out bad workers, while the type, design, and reward of a task may influence their amount significantly. Hence, a task should be designed so as to make faithful work more worthwhile than deception. In our case we first presented workers with a list of links to edits, and along each link, a form to select whether the linked edit is regular or vandalism. This simple and straightforward design led 80 % of the workers to quickly select options at random without clicking on the associated link. We therefore redesigned our task as a dialog that shows the worker one edit at a time, along the aforementioned form. This lowers the bar for faithful work since no additional interactions are necessary, and at the same, faking results requires the same amount of interaction.

3. PILOT EVALUATION

Before compiling our own corpus, we have first evaluated the quality of the annotations obtained via the Mechanical Turk by re-annotating the Webis-WVC-07 edits. This allows to determine whether and how scaling up vandalism annotation works. Additionally, we surveyed how often the workers use Wikipedia.

Corpus Annotation Accuracy. Table 1 shows the results of two rounds of re-annotating the Webis-WVC-07, each with a different number of annotators per edit. When considering three annotators per edit, two things can happen: all annotators agree with each other, or it is two against one. Moreover, in each of these cases the annotators either agree or disagree with the gold standard. In 56 % of the cases the annotators achieve perfect agreement with the gold standard, while in 2 % of the cases 3 annotators disagree completely. We have analyzed the latter edits and found that in half of these cases the annotations of the Webis-WVC-07 are wrong. In the other half of the disagreement cases we found that, with a ratio of about 3:1, more vandalism edits were considered regular than the other way around. We have conducted the same analysis for a round of 16 annotators per edit, only this time we consider more than 2/3 agreement among annotators as sufficient, while less agreement is considered a tie. Again, the majority of annotators either agree or disagree with the gold standard. We find that 93 % of the edits are annotated in accordance with the Webis-WVC-07. The remainder of the edits either correspond to the erroneous cases mentioned above or they are truly tough calls, even for an expert. Altogether, when considering only edits on which more than 2/3 of the annotators agree, the classification accuracies are 96 % and 99 %, which increase significantly over the baseline.

Worker Survey. The results of the survey are summarized in Table 2: while the majority of workers read Wikipedia daily to weekly a much smaller proportion also edit articles. 2 % of the workers admittedly vandalized Wikipedia. Interestingly, most of the workers do not often notice vandalism, however, when considering only workers who edit daily to monthly, the picture is turned upside down: Wikipedia's editors often have to restore vandalism. In any case, these numbers have to be taken with a grain of salt: they are not representative of all Wikipedia users, and there is no way of knowing whether the workers answered truthfully. In an attempt to minimize false answers, filling out the questionnaire was optional.

Table 2: Wikipedia usage of 753 Mechanical Turk workers.

Wikipedia Usage						Noticing Vandalism		
Reading		Editing		Vandalizing		(if editing daily-monthly)		
daily	27 %	daily	2 %	no	54 %	daily	3 %	(22 %)
weekly	23 %	weekly	3 %	yes	2 %	weekly	7 %	(34 %)
monthly	4 %	monthly	6 %			monthly	15 %	(33 %)
less	2 %	less	16 %			less	26 %	(10 %)
never	0 %	never	29 %			never	5 %	(1 %)
n/a	44 %	n/a	44 %	n/a	44 %	n/a	44 %	–

4. CORPUS CONSTRUCTION

In sum, we pursued the following strategy to generate our corpus: 33 000 edits were sampled from Wikipedia and annotated by 3 annotators each. All edits on which no more than 2/3 of its annotators agreed were re-annotated by another 3 annotators, and again, until ties were resolved or their number was small enough to be reviewed manually. We observe that the number of tie edits decreases exponentially with each iteration:

Iteration	0	1	2	3	4	5	6	7	8
Tie Edits	33 000	22 834	9 776	3 880	2 138	1 315	815	288	70

In order to check up on the worker's success in annotating edits, every 5th edit to be classified was in fact a vandalism edit chosen at random from the Webis-WVC-07. From iteration 3 onwards, however, the check edits were chosen from the vandalism edits already identified. This way, these edits received more votes than necessary, but in the long run, false positives may have been retracted. The 70 edits that were still tied after the 8th iteration have been reviewed by two experts who made a decision about them to the best of their knowledge. A handful of edits turned out to be undecidable, and those were given the benefit of the doubt. Finally, some of the edits became inaccessible along the way due to errors or administrative removal on the side of Wikipedia. A total of 32 452 edits were successfully annotated of which 2 391 are vandalism.

5. REFERENCES

[1] O. Alonso and S. Mizzaro. Can We Get Rid of TREC Assessors? Using Mechanical Turk for Relevance Assessment. In *Proc. of SIGIR'09.*

[2] R. S. Geiger and D. Ribes. The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. In *Proc. of CSCW'10.*

[3] K. Y. Itakura and C. L. A. Clarke. Using Dynamic Markov Compression to Detect Vandalism in the Wikipedia. In *Proc. of SIGIR'09.*

[4] M. Potthast and R. Gerling. Webis Wikipedia Vandalism Corpus Webis-WVC-07. http://www.webis.de/research/corpora, 2007.

[5] M. Potthast, B. Stein, and R. Gerling. Automatic Vandalism Detection in Wikipedia. In *Proc. of ECIR'08.*

[6] R. Priedhorsky, J. Chen, S. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, Destroying, and Restoring Value in Wikipedia. In *Proc. of Group'07.*

[7] K. Smets, B. Goethals, and B. Verdonk. Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. In *Proc. of WikiAI at AAAI'08.*