Identifying Semantic Edit Intentions from Revisions in Wikipedia

Diyi Yang[♣], **Aaron Halfaker**[♢], **Robert Kraut**[♠], **Eduard Hovy**[♣]
[♣] Language Technologies Institute, Carnegie Mellon University {diyi,hovy}@cmu.edu
[♢] Wikimedia Foundation ahalfaker@wikimedia.org
[♠] Human-Computer Interaction Institute, Carnegie Mellon University robert.kraut@cmu.edu

Abstract

Most studies on human editing focus merely on syntactic revision operations, failing to capture the intentions behind revision changes, which are essential for facilitating the single and collaborative writing process. In this work, we develop in collaboration with Wikipedia editors a 13-category taxonomy of the semantic intention behind edits in Wikipedia articles. Using labeled article edits, we build a computational classifier of intentions that achieved a micro-averaged F1 score of 0.621. We use this model to investigate edit intention effectiveness: how different types of edits predict the retention of newcomers and changes in the quality of articles, two key concerns for Wikipedia today. Our analysis shows that the types of edits that users make in their first session predict their subsequent survival as Wikipedia editors, and articles in different stages need different types of edits.

1 Introduction

Many online text production communities, including Wikipedia, maintain a history of revisions made by millions of participants. As Wikipedia statistics as of January 2017 show, English Wikipedia has 5.3 million articles with an average of 162.89 revisions per article, with revisions growing at a rate of about 2 revisions per second. This provides an amazing corpus for studying the types and effectiveness of revisions. Specifically, differences between revisions contain valuable information for modeling document quality or extracting users' expertise, and can additionally support various natural language processing (NLP) tasks such as sentence compression (Yamangil and Nelken, 2008), lexical simplification (Yatskar et al., 2010), information retrieval (Aji et al., 2010), textual entailment recognition (Zanzotto and Pennacchiotti, 2010), language bias detection (Recasens et al., 2013), spelling errors and paraphrases (Zesch, 2012; Max and Wisniewski, 2010).

To avoid building different approaches to extract the information needed by different NLP tasks (Ferschke et al., 2013), a unified framework to recognize edits from revisions is needed. Prior research on revision editing primarily develop syntactic edit action categories, from which they try to understand the effects of edits on meaning (Faigley and Witte, 1981; Yang et al., 2016). For instance, Daxenberger and Gurevych (2012) categorized edits based on whether edits affect the text meaning, resulting in syntactic edit categories such as file deletion, reference modification, etc. However, simply understanding the syntactic revision operation types does not provide the information we seek: *why* do editors do what they do? *how effective* are their actions? For example, syntactic edit type taxonomies cannot tell the difference between simplifying a paragraph and maliciously damaging that paragraph, since both involve deleting a sentence.

In this work, we focus explicitly on revision intention. We introduce a fine-grained taxonomy of the reasons why an author in Wikipedia made an edit. Example edit intentions include copy editing, elaboration, verification, and simplification. Compared to taxonomies that either focus on low-level syntactic operations (Faigley and Witte, 1981) or that mix syntactic and semantic classes (Daxenberger and Gurevych, 2013), a clean higher-level semantic categorization enables us to easily identify textual meaning changes, and to connect revisions to "what happens in the mind of the revising author during the revision"

(Fitzgerald, 1987; Daxenberger, 2016). In order to capture the meaning behind edits, we worked with 13 Wikipedians to build a taxonomy that captured the meaning of an revision, which we term *edit intention*, and hand-labeled a corpus of 7,177 revisions with their edit intentions. We then developed an automated method to identify these edit intentions from differences between revisions of Wikipedia articles. To explore the utility of this taxonomy, we applied this model to better understand two important issues for Wikipedia: new editor retention and article quality. Specifically, we examined whether edit intentions in newcomers' first editing sessions predict their retention, and examined how edits with different intentions lead to changes in article quality. These analyses showed that specific types of editing work were positively correlated with newcomer survival and articles in different stages of development benefited differently from different types of edits.

2 Related Work

Wikipedia revision histories have been used for a wide range of NLP tasks (Yamangil and Nelken, 2008; Aji et al., 2010; Zanzotto and Pennacchiotti, 2010; Ganter and Strube, 2009; Nelken and Yamangil, 2008). For instance, Yatskar et al. (2010) used Wikipedia comments associated with revisions to collect relevant edits for sentence simplification. Max and Wisniewski (2010) constructed a corpus of rewritings that can be used for spelling errors and paraphrases (Zesch, 2012). Similarly, Zanzotto and Pennacchiotti (2010) used edits as training data for textual entailment recognition, and Recasens et al. (2013) analyzed real instances of human edits designed to remove bias from Wikipedia articles. Most of these work employed manually defined rules or filters to collect relevant edits to the NLP task at hand.

Towards analyzing revisions and developing unified revision taxonomies (Bronner and Monz, 2012; Liu and Ram, 2011), Fong and Biuk-Aghai (2010) built machine learning models to distinguish between factual and fluency edits in revision histories. Faigley and Witte (1981) made a distinction between changes that affect meaning, called *text-base changes* and changes which do not affect meaning, called *surface changes*. The two categories are further divided into formal changes, meaning-preserving changes, microstructure changes and macro-structure changes. This taxonomy was later extended by Jones (2008) to take into account edit categories such as significant deletion, style, image insertion, revert, etc. Pfeil et al. (2006) proposed a 13-category taxonomy based on the data and performed manual annotation to compare cultural differences in the writing process in different versions of Wikipedia. Daxenberger and Gurevych (2013) introduced a finer-grained edit taxonomy, and performed multi-label classification to extract edit categories based on unparsed source text (Daxenberger and Gurevych, 2012). However, most taxonomies of edit categories contain only syntactic actions or a mixture of syntactic and semantic actions, failing to capturing the intention of revisions.

In terms of revision intentions, Zhang and Litman (2016) incorporated both argumentative writing features and surface changes from Faigley and Witte (1981) and constructed eight categories of revision *purposes*, such as claims/ideas, warrant/reasoning/backing, rebuttal/reservation, organization, clarify, etc. Tan and Lee (2014) used revisions to understand statement strength in academic writings. There are multiple works on the detection of specific subsets of revision intentions in Wikipedia, such as vandalism detection where the goal is to classify revisions as vandalized or non-vandalized (Harpalani et al., 2011; Adler et al., 2011) and language bias/neutral point of view detection (Recasens et al., 2013). Instead of recognizing a specific type of revision intention each time, our work aims at designing a systematic and comprehensive edit intention taxonomy to capture intentions behind textual changes.

Prior work also used edit types and intentions to better understand the process of collaborative writing, such as article quality improvement (Kittur and Kraut, 2008). For example, Liu and Ram (2011) found that Wikipedia article quality correlates with different types of contributors; similarly Yang et al. (2016) pointed out articles in different quality stages need different types of editors. However, there are few studies examining the specific types of edits that are predictive of article quality. Recent research shows that the number of active contributors in Wikipedia has been steadily declining since 2007, and Halfaker et al. (2012) suggested that the semi-automated rejection of new editors' contributions is a key cause, but they did not explore whether or not specific types of newcomers' work got rejected at different rates

Label	Description	α	Before	After
Clarification	Specify or explain an existing fact or meaning by example or discussion without adding new information	0.394	0.7%	4.1%
Copy Editing	Rephrase; improve grammar, spelling, tone, or punctuation	0.800	11.8%	14.8%
Counter Vandalism	Revert or otherwise; remove vandalism	0.879	1.9%	1.5%
Disambiguation	Relink from a disambiguation page to a specific page	0.401	0.3%	1.8%
Elaboration	Extend/add substantive new content; insert a fact or new meaningful assertion	0.733	12.0%	12.0%
Fact Update	Update numbers, dates, scores, episodes, status, etc. based on newly available information	0.744	5.5%	5.2%
Point of View	Rewrite using encyclopedic, neutral tone; remove bias; apply due weight	0.629	0.3%	2.2%
Process	Start/continue a wiki process workflow such as tagging an article with cleanup, merge or deletion notices	0.786	4.4%	5.8%
Refactoring	Restructure the article; move and rewrite content, without changing the meaning of it	0.737	1.9%	2.9%
Simplification	Reduce the complexity or breadth of discussion; may remove information	0.528	1.6%	4.6%
Vandalism	Deliberately attempt to damage the article	0.894	2.5%	2.0%
Verification	Add/modify references/citations; remove unverified text	0.797	5.4%	9.8%
Wikification	Format text to meet style guidelines, e.g. add links or remove them where necessary	0.664	33.1%	33.6%
Other	None of the above.	0.952	1.2%	-
Corpus Size			4,977	4,977	7,177

Table 1: A taxonomy of edit intentions in Wikipedia revisions, Cronbach's α agreement and the distributions of edit intention before and after corpus expansion. The percentage in each row represents what percentage of revisions are labeled with this edit intention. The percentages do not sum up to 100% because one revision could belong to multiple categories. The *After* corpus is used for all our analyses.

and how that affects retention. In this paper, we take advantage of this new taxonomy to explore correlations between edit intentions, newcomers' retention, and article quality.

3 Semantic Taxonomy of Edit Intentions

A *revision* is created whenever an editor saves changes to a Wikipedia page. As one revision could contain multiple local changes, each revision can be labeled with one or more edit intentions, representing the purposes of why an editor made that change. Different from prior research (Daxenberger, 2016; Yang et al., 2016), we do not distinguish between revisions and edits. Although an edit is a coherent local change and might belong to any edit categories, it cannot be used to represent the intentions of editors during the revision. For example, it might be difficult to recognize *Refactoring* if only one single edit is present. Since relocation or reorganization might involve several changes in the article, looking at one might lose the whole picture and lead to information loss. Moreover, edit types simply extracted from an edit is inadequate in outlining the correct intentions, for instance, adding a sentence could be *Clarification*, *Elaboration*, or *Vandalism*.

3.1 Taxonomy of Edit Intentions

Our semantic taxonomy of edit intentions builds on prior literature on collaborative writing (Faigley and Witte, 1981; Fitzgerald, 1987), research on document revision analyses (Bronner and Monz, 2012), studies on edit categories (Daxenberger and Gurevych, 2012; Fong and Biuk-Aghai, 2010), and work on purpose/intention classification (Zhang and Litman, 2016). In order to

ensure that our taxonomy captured the *intentions* that Wikipedians would find meaningful, we set up discussions with a group of 12 interested editors on a Wikipedia project talk page, and iteratively refined our taxonomy based on their feedback. Our discussion with Wikipedia editors is in this page[1]. We also analyzed which intentions get more confused with which and used that to guide the refinement.

We define a top level layer for the revision intention taxonomy: intentions that are common in general revisions: **General Revision Intentions**, and intentions that are specific in Wikipedia: **Wikipedia Specific Intentions**. This categorization leads to 13 distinct semantic intentions, and Table 1 provides detailed descriptions. Specifically, general revision intentions include: *Clarification*, *Copy Editing*, *Elaboration*, *Fact Update*, *Point of View*, *Refactoring*, *Simplification* and *verification*, and can be applicable to other contexts. *Counter Vandalism*, *Disambiguation*, *Process*, *Vandalism*, and *Wikification* are edit intentions related to Wikipeida. We also propose an *Other* category, intended for edits that cannot be labeled using the above taxonomy.

As the first work to model intentions of revisions, our taxonomy distills and extends existing edit type taxonomies. For instance, our intentions of "elaboration" and "verification" are extensions of "evidence" type proposed by (Zhang and Litman, 2016), and a syntactic category of "information deletion" in (Daxenberger and Gurevych, 2013) could be an instance of our "vandalism" or "simplification" depending on the context.

3.2 Corpus Construction

To construct a reliable, hand-coded dataset to serve as ground truth for automatic recognition of edit intentions, we employed four undergraduate students who had basic Wikipedia editing experience to label edits using our intention taxonomy, based on written annotation guidelines[2] vetted by Wikipedia editors and provided examples[3]. Moreover, to expose annotators to more working knowledge of Wikipedia, we provided three one-hour training sessions where annotators were asked to label a small set of revisions (around 50 each time) and to discuss their disagreements until consensus.

We randomly sampled 5,000 revisions from Jan, 2016 to June 2016 from the recent changes table[4] in the Wikipedia database. For each revision, we displayed the content difference[5] before and after the change to annotators, via a labeling interface that we developed. Because an editor could make several different types of edits within a single revision, we asked four RAs to label each revision with one or more of the possible semantic intentions. We collected four valid annotations for 4,977 revisions. We used Cronbach's α, a measure of internal consistency, to evaluate agreement among the annotators. The overall agreement α score was 0.782, indicating substantial agreement between different annotators; The rule of thumb 1993 suggests that Cronbachs alpha scores larger than 0.7 are considered as acceptable. The inter-annotator agreement per semantic intention is described in column α in Table 1.

3.3 Corpus Expansion

As shown in column *Before* in Table 1, some types of edit intentions, such as *disambiguation* and *clarification*, were very rare in the random-sample corpus. To address this under-representation problem, we used the text of editors' comments to expand the corpus by retrieving 200 more revisions for each edit intention except Vandalism and Counter-Vandalism, resulting in 2,200 revisions[6]. More precisely, as a common practice (Zanzotto and Pennacchiotti, 2010; Recasens et al., 2013), we utilized regular expressions to match the text from the comments, which editors often wrote when saving their revisions, to the edit intentions. For example, editors might be signalling that they were intending to fix problems of *Point of View* when their comments contained keywords such as "npov" or "neutral". Even though the comments sometimes signal the editors' intents, they are not infallible, editors may fail to complete the comment field, may only label one of the multiple edit intentions for a single revision, or write comments that are inaccurate, irrelevant, or incomplete. Thus the first author annotated the 2,200 revisions from the expanded corpus and

[1] https://en.wikipedia.org/wiki/Wikipedia_talk:Labels/Edit_types/Taxonomy

[2] http://www.cs.cmu.edu/~diyiy/data/edit_intention_annotation_doc.pdf

[3] https://en.wikipedia.org/wiki/Wikipedia:Labels/Edit_types/Examples

[4] https://www.mediawiki.org/wiki/Manual:Recentchanges_table

[5] en.wikipedia.org/wiki/?diff=712140761

[6] We used a practical and economic way to expand the corpus, and this made the intention distribution skewed away. We acknowledge this expansion as a limitation.

merged it with the randomly sampled corpus. The frequency of the edit intentions before and after the expansion is in Table 1. We used the majority voting to resolve the disagreement. That is, if at least 3 out of 4 annotators picked an intention for a revision, it will be selected as the ground-truth. The final corpus contains 5,777 revisions, and can be downloaded from here[7].

4 Identification of Edit Intentions

We frame automated identification of edit intentions as a multi-label classification task. We designed four sets of features for identifying edit intentions from revisions. Set I comprised two features associated with the **Editor**: *user registration* indicating whether the editor of a particular revision was registered or anonymous and *tenure*, which refers to the elapsed months between the current revision and editors' registration date. Set II comprised 16 features associated with the **Comment** written by the editor to describe the revision, including *comment length* and a set of regular expressions to match intentions such as `*pov*`, `*clarify*`, `*simplif*`, `*add link*`, etc. Set III comprised 198 features associated with the **Revision Diff**, based on content differences between current revision and the previous one. They are similar to textual features defined in Daxenberger and Gurevych (2013), but we considered a wider range of objects being modified. In particular, we computed the difference in the number of characters, uppercase words, numeric chars, white-spaces, markups, Chinese/Japanese/Korean characters, HTML entity characters, URLs, punctuations, break characters, etc. We also considered languages features, such as the use of stop words, obscene words and informal words. Set IV comprises two features associated with **Vandalism** and **Revert**. We utilized the Wikipedia API to extract whether a revision was likely to be vandalism[8] or reverting revisions[9].

4.1 Identification Result

We extracted the input features with the help of Revision Scoring package[10] and framed this task a multi-label classification problem. For multi-label classification, we considered solving them by using single-label classification algorithms and by transforming it into one or more single-label classification tasks. We used the multi-label classifiers implemented in Mulan (Tsoumakas et al., 2011), with 10-fold cross validation. We utilized Binary Relevance (**BR**) to convert our multi-label classification into 13 binary single-label problems. Similar to Daxenberger and Gurevych (2013); Yang et al. (2016), we used Random k-labelsets **RAKEL** method that randomly chooses l small subset with k categories from the overall set of categories. We set l as 26, twice the size of the categories, and set k as 3. **MLKNN** method that classifies edit intentions based on K (K=10) nearest neighbor method. We used C4.5 decision tree classifiers in BR and RAKEL, as recommended by prior work (Daxenberger and Gurevych, 2013; Potthast et al., 2013). Prior research shows that sophisticated neural network models for text-classification largely rely on factors such as dataset size (Zhang et al., 2015; Joulin et al., 2016). Due to the size of our corpus and the complexity of this task, we did not use them.

To evaluate the relative accuracy of the multi-label classifier, we compared it to several baselines. The random baseline, denoted as **Random** in Table 2, assigns labels randomly. The majority category baseline, denoted as **Majority**, assigns all edits the most frequent intention, elaboration. Since revision comments may be especially as informative in reflecting edit intentions, the comment baseline, denoted as **CMT**, is a Binary Relevance classifier that includes only the comments features from Set II. We also created a Binary Relevance classifier, denoted as **BR-**, which excludes comment features and only used features from Sets I, III and IV.

Table 2 shows the evaluation metrics for the baselines and our multi-label classifiers. The metrics include the Exact Match subset accuracy, which evaluates whether the predicted labels are the same as the actual labels. These classifiers are available upon request. Table 2 also shows example-based measures of Accuracy, Precision, Recall and F1 Score, weighting each edit equally. It also shows label-based measures of accuracy – the micro- and macro-averaged F1 scores– which weight each edit intention category equally. As a ranking based measure, we measured One Error,

[7] http://www.cs.cmu.edu/~diyiy/data/edit_intention_dataset.csv

[8] https://ores.wmflabs.org/v2/scores/enwiki/goodfaith/71076450

[9] http://pythonhosted.org/mwreverts/api.html

[10] http://pythonhosted.org/revscoring/

Metric		Random	Majority	CMT	BR-	BR	MLKNN	RAKEL
Example	Exact Match	0.052	0.284	0.352	0.391	0.426	**0.452**	0.292
	Accuracy	0.052	0.283	0.428	0.498	0.540	**0.542**	0.338
	Precision	0.084	0.417	0.479	0.626	0.586	**0.599**	0.381
	Recall	0.052	0.285	0.458	0.562	**0.611**	0.578	0.344
	F1 Score	0.052	0.285	0.455	0.536	**0.580**	0.574	0.354
Label	Macro F1	0.060	0.042	0.310	0.487	**0.597**	0.576	0.385
	Micro F1	0.074	0.370	0.528	0.583	**0.621**	0.613	0.441
Ranking	One Error	0.920	0.583	0.415	0.400	0.358	**0.320**	0.434

Table 2: Performance comparison for predicting edit intentions from revisions. Best results are bold.

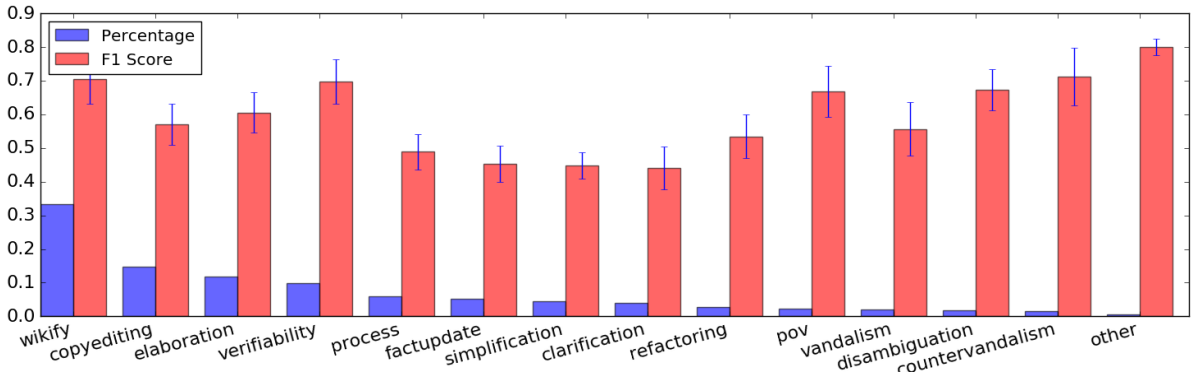

Figure 1: The relative frequency of each edit intention, and its F1 score provided by the **BR** model.

which evaluates how many times the top ranked predicted intention is not in the set of true labels of the instance.

Results show that the Binary Relevance (BR) and MLKNN classifiers, which used all our constructed features, outperformed Random and Majority baselines. Moreover, the BR and MLKNN methods show relatively similar best performances. Although multiple studies have utilized revisions' comments as "groundtruth" to collect desired edits, the CMT method, which includes only comment features, is less accurate than either the BR or MLKNN models. Note that predicting 14-category semantic intentions is more challenging compared to classifying low-level syntactic actions, such as inserting an image (Daxenberger and Gurevych, 2013).

5 Intentions, Survival and Quality

The automated measurement of edit intentions provides a general framework to analyze revisions and can facilitate a wide range of applications, such as collecting specific types of revisions (Yatskar et al., 2010; Recasens et al., 2013; Zanzotto and Pennacchiotti, 2010) and outlining the evolution of author roles (Arazy et al., 2015; Yang et al., 2016). In this section, we demonstrate two examples of how this intention taxonomy can be applied to better understand the success of online collaboration communities (Kraut et al., 2010), specifically the process of these sites to retain new contributors and create innovative products. To this end, we first investigate what newcomers are intended for in their first sessions and whether their edit intentions can account for their survival in Wikipedia. We then examine how edits carrying on different intentions at distinct times in an article's history influence changes in its quality.

5.1 How Edit Intentions Affect Survival

To explore newcomers' intentions during their first experience editing articles, we focus on users' first edit sessions in Wikipedia. Here, **Edit Session** is defined as a sequence of edits performed by a registered user with less than one hour's time gap between two adjacent edits (Halfaker et al., 2012). We then compare edit intentions of newcomers who survive - **Survivors**, and newcomers who do not - **Non-survivors**. Here, newcomers are defined as surviving if they performed an edit at

Edit Intention	*Intention Dist*		*Revert Ratio*	
	NS	SS	NS	SS
clarification	**0.2%**	**0.4%**	0.1%	0.1%
copy editing	**12.1%**	**14.4%**	**6.9%**	**3.8%**
counter vandalism	0.1%	0.0%	0.1%	0.0%
disambiguation	0.0%	0.0%	0.0%	0.0%
elaboration	27.7%	26.5%	**16.5%**	**6.9%**
fact update	4.2%	3.8%	**3.4%**	**1.7%**
point of view	**0.1%**	**0.2%**	0.0%	0.1%
process	2.0%	2.3%	**1.9%**	**0.7%**
refactoring	1.1%	1.3%	**0.9%**	**0.5%**
simplification	**3.7%**	**3.1%**	**3.1%**	**1.4%**
vandalism	**13.8%**	**6.1%**	**16.0%**	**4.7%**
verification	7.0%	7.4%	**3.8%**	**2.7%**
wikification	**25.8%**	**32.3%**	**14.0%**	**6.9%**

Table 3: The edit intention distribution in the first sessions (Intention Dist) and the revert ratio comparison (Revert Ratio), among non-survivors (NS) and survivors (SS). The numbers are bolded if 1-way ANOVA tests for difference between two groups are significant, with $p<0.05$.

least two months after their first edit session.

5.1.1 Intention Comparison

Among 100,000 randomly sampled Wikipedia users, 21,096 made revisions in the Main/Article namespace during their first editing session. Among these 4,407 were survivors (i.e., made an edit two months after registering) and 16,689 were non-survivors. We applied our edit intention model to 53,248 revisions in users' first sessions, and compared the percentages of different types of edit intentions between survivors and non-survivors, as shown in ***Intention Dist*** column in Table 3. We also performed 1-way ANOVA to test whether survivors and non-survivors have the same mean for each edit intention. We observed that, survivors tend to do more copy-editing (Δ_+=2.3%) and more wikification (Δ_+=6.5%), while non-survivors seem to perform more simplification and vandalism, which might provide signals for detecting vandals.

5.1.2 Revert Analysis

To explore the relationship between rejection of contributions and newcomer retention, we also visualized the revert ratios of different types of edit intentions for survivors and non-survivors in their first session. Here, **Revert** refers to whether an edit from the author was reverted or completely removed by another user, and we detect reverts using MediaWiki Reverts library[11]. We then measured the revert ratio for each edit intention by calculating the percentage of revisions belonging to a specific edit intention, among all reverted revisions in users' first sessions. As shown in the ***Revert Ratio*** column in Table 3, in general, non-survivors get reverted more compared to survivors, across all edit intentions. Interestingly, non-survivors compared to survivors get reverted more when performing *Wikification*, *verification* and *Refactoring*, suggesting that sophisticated types of work might not be suitable for beginners.

Edit Intention	Survival	Quality Changes
clarification	0.029	0.001
copy editing	0.033	$0.011^{\dagger}$
counter vandalism	0.004	$-0.020^{\dagger}$
disambiguation	-0.003	$-0.006^{\dagger}$
elaboration	-0.024	$0.061^{\dagger}$
fact update	-0.001	0.002
point of view	0.041	-0.003
process	$0.051^{\dagger}$	$-0.024^{\dagger}$
refactoring	-0.013	$0.011^{\dagger}$
simplification	-0.002	$-0.008^{\dagger}$
vandalism	$-0.211^{\dagger}$	$-0.005^{\dagger}$
verification	0.047	$0.068^{\dagger}$
wikification	$0.099^{\dagger}$	$-0.010^{\dagger}$

Table 4: Regression coefficients of different edit intentions for predicting Newcomer **Survival** and Article **Quality Changes**. † means the coefficient is statistically significant ($p<0.05$)

5.1.3 Newcomer Survival

As a further exploration of the relationship between edit intentions and newcomer survival, we performed a logistic regression using edits in survivors' and non-survivors' first sessions. To handle this imbalanced data (i.e., many more negative examples than positive examples in training), we performed majority-class under-sampling to make this dataset balanced. Similar to Halfaker et al. (2012), we controlled the number of revisions completed during the first session (a proxy for an editor's initial investment), and the number of revisions reverted in their first sessions. We

[11] http://pythonhosted.org/mwreverts/

described the regression coefficients of statistically significant edit intentions in the **Survival** column of Table 4. This logistic model achieves an Accuracy of 60.98%, Recall of 58.30%, Precision of 78.08% and F1-score of 66.76%. Editing articles for the purposes of *Process*, *Verification* and *Wikification* significantly predict the survival of newcomers, while performing vandalism is a strong negative predictor for survival.

5.2 How Intentions Affect Article Quality

Although there are over 5.5 million articles in the English Wikipedia, fewer than 0.2% have been evaluated by Wikipedians as good articles and around 92% have been evaluated as start or stub class articles, Wikipedia's two lowest quality categories. In this section, we examine how edits with different intentions at distinct times in an article's history influence changes in its quality.

This task is framed as a prediction task, i.e. using edits' intentions and a set of control variables to predict changes in article quality. We borrowed a Article Quality Prediction Dataset released in Yang et al. (2016), which consists of the quality ratings collected in January and June, 2015 of 151,452 articles. We collected 1,623,446 revisions made to these articles between January and June 2015, by randomly sampling 10% revisions that were made to these articles during that time periods. Specifically, the outcome *article quality change* is calculated by subtracting the previous quality score from the end quality score. The control variables include the previous article quality score, the total number of edits, the total number of editors, the changed bytes to an article, and the total number of edits to the article talk page during the six months. To construct edit-intention predictors, we summed the number of edits for each edit intention during the six months divided by the total number of revisions in this article.

Results of the linear regression model, shown in **Quality Changes** column of Table 4, show that our constructed regression model is significantly predictive of article quality changes (R^2 = 0.225). The results show that, keeping all control variables fixed, more *Copy Editing*, *Elaboration*, *Refactoring* and *Verification* are positively associated with improvements in article quality; in contrast, *Vandalism*, *Counter Vandalism*, *Disambiguation*, *Process* and *Simplification* predict declines in article quality. The first four of these edits types often

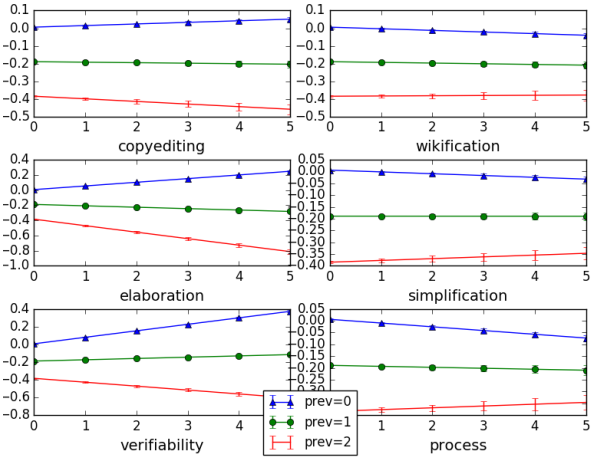

Figure 2: Interaction effect of different levels of edit intentions and different levels of previous article quality (**prev**) on article quality changes. All variables are standardized. The Y-axis measures the predictive margins and X-axis refers to different standardized levels of edit intention.

occur with reducing the article content, removing or redirecting pages. Improper use of them might be detrimental to article quality.

To determine if the effect of edit intentions on quality changes depends upon the initial quality of the article, we added the interaction terms between the previous quality score and edit percentages of different intentions (e.g., clarification x previous quality), and visualized interaction effects in Figure 2. When examining the interaction terms in more detail: the negative slope of *copy editing* (when prev=2) suggests that, as articles increase in quality, copy editing is needed less. We found similar trends for interactions between previous quality and *elaboration* and *verification*, which are essential for articles in the starting stages. In contrast, the positive slopes for *simplification*, *wikification* and *process* suggest that, as articles increase in quality, simplifying articles' content, adding proper links or reorganizing their structure becomes more important. Overall, these results reveal that different types of edit intentions are needed at different quality stages of articles.

6 Discussion and Conclusion

In this work, we proposed 13 semantic intentions that motivate editors' revisions in English Wikipedia. Example edit intentions include copy editing, elaboration, simplification, etc. Based

in a labeled corpus of revisions, we developed machine-learning models to automatically identify these edit intentions. We then examine the relations between edit intentions, newcomers survival, and article quality improvement. We found that (1) survivors tend to do more copy editing and wikification; non-survivors seem to perform more vandalism and other sophisticated types of work, and the latter often gets reverted more; (2) Different types of contributions are needed by articles in different quality stages, with elaboration and verification are needed more for articles in the starting stages, and simplification and process become more important as article quality increases.

Our proposed edit intention taxonomy and the constructed corpus can facilitate a set of downstream NLP applications. First, classifiers based on this intention taxonomy can help retrieve large scale and high quality revisions around simplification, neutral point of view or copy editing, which provides amazing corpora for studying lexical simplification, language bias detection and paraphrases. Second, as we showed in Section 5.2, determining how different edit types influence changes in articles is of great use to better the causes of quality variance in collaborative writing, such as detecting quality flaws (Anderka et al., 2012) and providing insights on which specific aspects of an article needs improvement and what type of work should be performed. The ability to identify the need for editing, and specifically the types of editing work required, can greatly assist not only collaborative writing but also individual improvement of text. Moreover, even though our edit taxonomy is for English Wikipedia, it can be applied to other language versions of Wikipedia. We are now deploying the same edit intention taxonomy for Italian Wikipedia, and plan to apply it to other low resourced languages in Wikipedia. Finally, beyond the context of Wikipedia, similar taxonomies can be designed for analyzing the collaboration and interaction happened in other online contexts such as academic writing (e.g., Google Docs or ShareLatex, etc).

Acknowledgement

This research was supported in part by a grant from Google to Robert Kraut. The first author was supported by Carnegie Mellon Presidential Fellowship. The authors would like to thank our reviewers and Wikipedian for helpful feedback.

References

B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 277–288.

Ablimit Aji, Yu Wang, Eugene Agichtein, and Evgeniy Gabrilovich. 2010. Using the past to score the present: Extending term weighting models through revision history analysis. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 629–638. ACM.

Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting quality flaws in user-generated content: the case of wikipedia. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 981–990. ACM.

Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. 2015. Functional roles and career paths in wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1092–1105.

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'12, pages 356–366.

Jose M Cortina. 1993. What is coefficient alpha? an examination of theory and applications. *Journal of applied psychology*, 78(1):98.

Johannes Daxenberger. 2016. *The Writing Process in Online Mass Collaboration: NLP-Supported Approaches to Analyzing Collaborative Revision and User Interaction*. Ph.D. thesis, Technische Universität.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, Mumbai, India. The COLING 2012 Organizing Committee.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, Washington, USA. Association for Computational Linguistics.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication*, 32(4):400–414.

Oliver Ferschke, Johannes Daxenberger, and Iryna Gurevych. 2013. A survey of nlp methods and

resources for analyzing the collaborative writing process in wikipedia. In *The Peoples Web Meets NLP*, pages 121–160. Springer.

Jill Fitzgerald. 1987. Research on revision in writing. *Review of educational research*, 57(4):481–506.

Peter Kin-Fong Fong and Robert P. Biuk-Aghai. 2010. What did they do? deriving high-level edit histories in wikis. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, WikiSym '10, pages 2:1–2:10, New York, NY, USA. ACM.

Viola Ganter and Michael Strube. 2009. Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACL Short '09, pages 173–176.

Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2012. The rise and decline of an open collaboration system: How wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, page 0002764212469365.

Manoj Harpalani, Michael Hart, Sandesh Singh, Rob Johnson, and Yejin Choi. 2011. Language of vandalism: Improving wikipedia vandalism detection via stylometric analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 83–88, Portland, Oregon, USA. Association for Computational Linguistics.

John Jones. 2008. Patterns of revision in online writing a study of wikipedia's featured articles. *Written Communication*, 25(2):262–289.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Aniket Kittur and Robert E. Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 37–46.

Robert Kraut, Moira Burke, John Riedl, and P Resnick. 2010. Dealing with newcomers. *Evidencebased Social Design Mining the Social Sciences to Build Online Communities*, 1:42.

Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst.*, 2(2):11:1–11:23.

Aurlien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from wikipedias revision history. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Rani Nelken and Elif Yamangil. 2008. Mining wikipedia's article revision history for training computational linguistics algorithms. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 31–36.

Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Overview of the 5th international competition on plagiarism detection. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 301–331. CELCT.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proceedings of ACL (short paper)*.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. 2011. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12:2411–2414.

Elif Yamangil and Rani Nelken. 2008. Mining wikipedia revision histories for improving sentence compression. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 137–140. Association for Computational Linguistics.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who did what: Editor role identification in wikipedia. In *Tenth International AAAI Conference on Web and Social Media*.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368. Association for Computational Linguistics.

Fabio Massimo Zanzotto and Marco Pennacchiotti. 2010. Expanding textual entailment corpora from wikipedia using co-training. In *Proceedings of the COLING-Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, volume 128.

Torsten Zesch. 2012. Measuring contextual fitness using error contexts extracted from the wikipedia revision history. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 529–538, Avignon, France. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2016. Using context to predict the purpose of argumentative writing revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1424–1430, San Diego, California. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.