Amplifying Community Content Creation with Mixed-Initiative Information Extraction

Raphael Hoffmann, Saleema Amershi, Kayur Patel, Fei Wu, James Fogarty, Daniel S. Weld
Computer Science & Engineering
DUB Group, University of Washington
Seattle, WA 98195
{raphaelh, samershi, kayur, wufei, jfogarty, weld}@cs.washington.edu

ABSTRACT

Although existing work has explored both information extraction and community content creation, most research has focused on them in *isolation*. In contrast, we see the greatest leverage in the *synergistic pairing* of these methods as two interlocking feedback cycles. This paper explores the potential synergy promised if these cycles can be made to *accelerate each other* by exploiting the *same edits* to advance both community content creation and learning-based information extraction. We examine our proposed synergy in the context of Wikipedia infoboxes and the Kylin information extraction system. After developing and refining a set of interfaces to present the verification of Kylin extractions as a non-primary task in the context of Wikipedia articles, we develop an innovative use of Web search advertising services to study people engaged in some other primary task. We demonstrate our proposed synergy by analyzing our deployment from two complementary perspectives: (1) we show we *accelerate community content creation* by using Kylin's information extraction to significantly increase the likelihood that a person visiting a Wikipedia article as a part of some other primary task will spontaneously choose to help improve the article's infobox, and (2) we show we *accelerate information extraction* by using contributions collected from people interacting with our designs to significantly improve Kylin's extraction performance.

ACM Classification:
H5.2. Information Interfaces and Presentation: User Interfaces; H1.2. Models and Principles: User/Machine Systems.

Keywords: Community content creation, information extraction, mixed-initiative interfaces.

INTRODUCTION AND MOTIVATION

The explosion of information available on the Web presents important human-computer interaction challenges. Many techniques developed to address these challenges leverage the *structure* of Web content. For example, faceted browsing exploits a set of attribute/value pairs for objects in a collection [26]. Browser enhancements like Sifter parse structured content, such as product search results, to enable interactive sorting and querying [9]. Bibliographic sites like Citeseer locate and parse citations, enabling reference counting and navigation among related documents [5]. Web search interfaces like Assieme identify and leverage relationships among and within Web pages to better support common search tasks [7]. Despite the differing goals of this variety of systems, a fundamental challenge underlies all such systems: *How can systems scalably obtain the necessary structured information*?

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

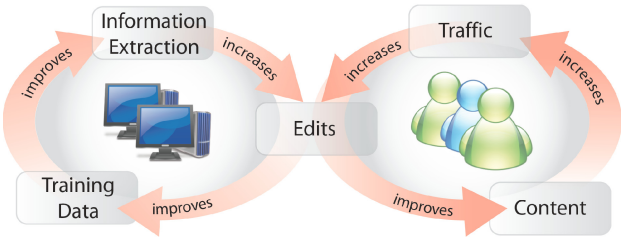

Figure 1: We envision the synergistic pairing of *information extraction* with *community content creation*, using the same edits to accelerate both feedback cycles.

One popular approach is *community content creation*. People visiting the photo-sharing site Flickr or the social bookmarking site Delicious, for example, can browse photos and bookmarks using tags applied by other people. Amazon and Netflix provide recommendations based on community-contributed ratings. Finally, Wikipedia is well known for its community-created articles. Despite such examples of extremely successful community approaches, many other sites have been unable to bootstrap themselves to critical mass or to overcome work/benefit disparities [6]. Significant research has therefore explored how and why people contribute to sites like Wikipedia [1, 10, 15, 19]. This research has shown that the vast majority of work is usually done by a relatively small set of people. We are therefore interested in new methods for lowering barriers to editing and for incenting broader contribution.

A second popular approach to structured information is *information extraction*. Sifter, for example, uses a set of heuristics to identify typical patterns, such as page numbers in search results [9]. Systems like Citeseer and Assieme use a combination of heuristics and statistical machine learning algorithms [5, 7]. Although learning-based approaches can be powerful and robust, they have at least two important limitations. First, learning algorithms typically require

Figure 2: An example page containing several opportunities for mixed-initiative contribution to Wikipedia. The person viewing this page has moused over an icon in the page that indicates that the system has analyzed the text of the article and found a potential value for Ray Bradbury's birthplace. The person's response to this question will be used to improve both the information extraction system and the content of this page.

numerous labeled training examples, whose collection is typically expensive and time consuming. Second, learning methods can be error prone, and state-of-the-art systems with precisions of 80 to 90% are considered successes. Although this performance can enable many applications, it is unacceptable for Wikipedia and many other sites.

Existing work has explored both information extraction and community content creation, but has focused on these approaches in *isolation*. In contrast, we see the greatest leverage in the *synergistic pairing* of these two approaches. We envision a pair of interlocking feedback cycles, illustrated in Figure 1. The left cycle corresponds to the traditional training of an information extraction system, wherein a person manually annotates a corpus with labels. After learning and testing an extractor, a person can examine the results and provide additional data to improve performance. Similarly, the right cycle corresponds to the familiar bootstrapping problem in community content creation, wherein quality content is required in order to attract people so that they might further contribute.

This paper explores the great potential synergy promised if these cycles can be made to *accelerate each other* by exploiting the *same edits* to advance both learning-based information extraction and community content creation. This synergy might enable many benefits, such as the semi-automated maintenance of portions of community content sites, the bootstrapping of new sites with knowledge extracted from the larger Web, and even the eventual semantification of much of the existing Web. Realizing this synergy requires new designs that both (1) leverage information extraction to increase visitor contribution rate, and (2) leverage visitor contributions to improve the reliability of information extraction.

We explore these challenges in the context of Wikipedia and the *Kylin* information extraction system [24, 25]. More specifically, we focus on Wikipedia infoboxes, tabular summaries present in many Wikipedia articles (Figure 2). Kylin analyzes Wikipedia articles containing infoboxes and learns to extract values from untagged articles (e.g., analyzing articles with infoboxes containing a birthdate field and thus learning to extract birthdates from other articles). Although the details of our current work are tuned for Wikipedia, we argue that our synergistic approach is potentially relevant to many different types of websites.

This paper makes the following contributions:

- We identify the potential for synergistically pairing community content creation with learning-based information extraction, using the same edits so that both feedback cycles accelerate each other.
- Using the Wikipedia community as a case study, we examine the challenge of simultaneously addressing the needs and norms of both learning-based information extraction systems and social communities.
- We develop and refine a set of interfaces to present the verification of Kylin extractions as a non-primary task in the context of Wikipedia articles. Our designs leverage Kylin extractions to make it possible to contribute to improving a Wikipedia article with just a few mouse

clicks, and we develop several designs to explore a tradeoff between contribution rate and unobtrusiveness.

- We develop an innovative use of Web search advertising services to study people interacting with our interfaces while engaged in some other primary task (the task that prompted them to perform the Web search that eventually brought them to our page).
- We demonstrate our desired synergy through a pair of complementary analyses: (1) we show we *accelerate community content creation* by using Kylin's information extraction to significantly increase the likelihood that a person visiting a Wikipedia article as a part of some other primary task will spontaneously choose to help improve the article's infobox, and (2) we show we *accelerate information extraction* by using contributions collected from people interacting with our designs to significantly improve Kylin's extraction performance.

BACKGROUND: KYLIN EXTRACTION

In order to provide appropriate context for the remainder of this paper, we briefly review the Kylin information extraction system; space precludes a comprehensive explanation, but more information is available in [24, 25].

Obtaining Data. Kylin obtains training data by analyzing existing infoboxes in Wikipedia articles. Each infobox has a class (e.g., the Ray Bradbury infobox in Figure 2 is of class *writer*). Kylin collects examples of articles containing infoboxes of a given class, then analyzes fields appearing in those infoboxes (e.g., birthdate and nationality). Kylin next heuristically chooses the best sentence in the article which contains the same value as the infobox field (e.g., a sentence containing the same date that the infobox provides for birthdate). These sentences provide positive training examples, and other sentences provide negative examples.

Document and Sentence Classifiers. Kylin learns two types of classifiers. For each infobox class, a *document classifier* is used to recognize articles of that class. Each sentence in an article is then examined by a *sentence classifier* trained to predict whether a sentence is likely to contain the value of a field (and thus whether to apply the extractor).

Extractors. Extracting a value from a sentence is treated as a sequential data-labeling problem. Kylin trains conditional random fields with a variety of features (e.g., presence of digits, part-of-speech tags, capitalization). Although Kylin learns accurate extractors for popular infobox classes, the majority of classes do not have enough existing examples to effectively train Kylin. A recent Wikipedia snapshot shows that 72% of classes have 100 or fewer instances and 40% have 10 or fewer instances. Therefore, this paper's approach to synergistically obtaining more training data is an important step for improving Kylin's accuracy.

METHOD

We first interviewed three senior members of the Wikipedia community (two administrators and a veteran contributor, all of whom had been contributing for at least four years), meeting face-to-face with each for approximately two hours. Given our interest in using the same edits to drive both feedback cycles, our interviews focused on why people do or do not contribute to Wikipedia, what aspects of the Wikipedia community are difficult for newcomers, and the role a system like Kylin could play in Wikipedia.

Informed by prior work on interruptions and ambiguity resolution [3, 11, 13, 16], we next designed three interfaces examining different ways to leverage Kylin's information extraction to accelerate community content creation. Our designs share a focus on promoting ambiguity resolution as a *non-primary* task, but they explicitly probe the tradeoff between contribution rate and unobtrusiveness.

We refined and informally evaluated our designs in informal talk-aloud sessions. We presented them to nine participants (randomizing order to address potential carryover effects) and asked them to comment on aspects of the interaction they found difficult, discuss aspects of the interface they found obtrusive, and to provide overall indications of their preference. Because our goal was to refine the designs, we made improvements throughout these sessions.

Because it is difficult to envision a laboratory study which measures how often people spontaneously contribute to Wikipedia, we evaluated our synergistic approach through a novel use of Web search advertising services. By placing ads for 2000 Wikipedia articles, we attracted visitors who were engaged in some other primary task. We assigned these visitors to different study conditions, logged their interaction with our designs, and examined their contribution rate.

DESIGNING FOR THE WIKIPEDIA COMMUNITY

Our interviews with veteran Wikipedia contributors, together with prior work examining the Wikipedia community [1, 10, 15], helped us identify two critical constraints governing the integration of information extraction into Wikipedia: (1) a need to balance Wikipedia policy regarding *bots* with policy that contributors *be bold,* and (2) the opportunity to encourage greater participation in Wikipedia. Taken together, these have led us to pursue a mixed-initiative approach using interfaces designed to solicit contribution as a non-primary task.

Being Bold, Bots, and a Mixed-Initiative Approach

Wikipedia policy states that people should *be bold* when updating pages [21]. This policy recognizes that some edits are contentious and must wait for discussion to yield consensus, but that Wikipedia develops faster when more people contribute. It is therefore important that people be bold enough to make edits. The policy emphasizes, for example, that people should feel comfortable correcting copy-editing mistakes and factual errors, rather than flagging content for discussion or for others to correct.

Wikipedia also has an explicit policy regarding automated *bots* [22], including the requirement that they be both "harmless and useful". At the time of this writing, Wikipedia lists 392 approved bots. These perform such tasks as updating links between different language versions of Wikipedia, maintaining lists, and archiving old discussion pages. Bots are appropriate for this work

because simple programs provide error-free performance and because the automation frees members of the community to do other work. In contrast, it is clear Kylin should not autonomously add infobox values, because its precision currently ranges between 75 to 98% [24, 25] and the errors associated with this state-of-the-art performance would likely be considered harmful.

One approach would be to automatically post Kylin extractions on article *talk* pages, hoping people will manually make the necessary edits (each Wikipedia article has an associated *talk* page where changes to the article can be discussed). Although this would ensure extraction errors are not automatically introduced into Wikipedia infoboxes, the compatibility of this approach with the spirit of *be bold* is less clear. Updating an infobox with a birthdate that already appears in the body of an article is not likely to be contentious or require consensus. Instead, it is simply important that the extraction be confirmed as correct.

In addition to being a poor match to Wikipedia's *be bold* policy, posting extractions to article talk pages also fails to enable our desired synergistic pairing. In order to advance the information extraction feedback cycle, a system needs to collect additional labels by learning whether extractions are correct. Even if a system monitored changes to an article in order to observe whether a talk page suggestion was eventually enacted, it is not clear how to interpret edits. For example, a page might change significantly between the time an extraction was posted and the time the infobox is edited. Furthermore, a person might update an infobox in a manner similar, but not identical to, the suggested edit. In these and many more situations, it is unclear what relationship an edit might have to a posted extraction.

We therefore focus on a mixed-initiative approach [8], wherein potential infobox contributions are automatically extracted but then manually examined and explicitly confirmed before being published. This addresses all of the challenges discussed above. We enable the information extraction feedback cycle with additional training data collected through explicit indications of whether an extraction is correct. We address the requirement that *bots* be harmless with the manual confirmation of Kylin extractions, and we address the spirit of the *be bold* policy by designing interfaces to present the confirmation of Kylin extractions as a non-primary task in the context of Wikipedia articles, as discussed next.

Contribution as a Non-Primary Task

Any community content creation system must provide an incentive for people to contribute, and there are many ways one might incent people to examine and confirm Kylin extractions. For example, Doan *et al.* propose requiring people provide a small amount of work before gaining full access to a service [12]. We believe, however, that coercive approaches are unacceptable to the Wikipedia community, whose culture is based in altruism and indirect author recognition [1, 10, 15]. Existing systems, such as the AutoWikiBrowser [20] and Cosley *et al.*'s SuggestBot [2], focus on experienced Wikipedia contributors who are already motivated to contribute, helping them to find work.

Instead of targeting experienced Wikipedia contributors (perhaps by posting links in article talk pages that bring experienced contributors to a page where they could explicitly confirm mixed-initiative extractions), we believe our desired synergistic pairing is better served by focusing on people who are not already Wikipedia contributors. This is because Wikipedia contributions currently follow a power law, with a relatively small number of prolific editors making most contributions [15, 19]. Prior work (e.g., [1, 10, 15]) and our interviews with veteran contributors suggest this is because people do not know they can contribute, are time-constrained, are unfamiliar with Wikimarkup, feel unqualified, or feel their contributions are not important.

Overcoming these challenges and soliciting contributions from new people offers the potential to advance the community content creation feedback cycle in two ways. First, shifting the work of validating extractions onto newcomers frees experienced contributors to focus on other more demanding work. Second, it provides a quick and easy way for newcomers to make meaningful contributions. Bryant *et al.* report that newcomers become members of the Wikipedia community by participating in peripheral yet productive tasks that contribute to the overall goal of the community [1]. Making it easy for newcomers to examine and confirm Kylin extractions might therefore encourage more people to become active Wikipedia members.

This paper therefore focuses on soliciting contributions from people who have come to Wikipedia for some other reason, perhaps because they are seeking a specific piece of information or simply browsing out of curiosity, but did not already intend to work on Wikipedia. Contribution is therefore not a person's primary task. The challenge is then to design interfaces that make the ability to contribute by verifying Kylin extractions sufficiently visible that people choose to contribute, but not so obtrusive that people feel contribution is coerced (which would be seen as a violation of the Wikipedia community's goal of supporting free access to knowledge for everyone).

INTERFACE DESIGN AND REFINEMENT

In considering how to integrate the verification of Kylin extractions into Wikipedia articles, we note that Wikipedia already uses *cleanup tags* within articles [23]. Figure 3 shows an example of one cleanup tag, drawing attention to the need to add references to an article. Although these tags illustrate the fact the Wikipedia community considers it appropriate to embed small indications of the need for work

Figure 3: The Wikipedia community already uses *cleanup tags* to indicate opportunities for contribution, but these provide little assistance to potential contributors who are time-constrained or unfamiliar with Wikimarkup.

within articles, the current tags provide little or no assistance to potential contributors who are time-constrained or unfamiliar with Wikimarkup. In contrast, we aim to not only present the need for work within an article but also to leverage Kylin extractions so a person can contribute very quickly and easily (with just a few clicks). This section discusses general strategies we apply in all of our designs, as well as the details of our *popup*, *highlight*, and *icon* designs.

Ambiguity in a Non-Primary Task

Each of our designs explores a different approach to drawing attention to verifying Kylin extractions, but an important aspect of contribution as a non-primary task is the fact many people will never notice the potential to contribute. All of our designs therefore never present unverified information in a way that might be mistakenly interpreted as a part of the article. Figure 2, for example, shows an infobox populated with the placeholder "Check our guess." Although prior work where ambiguity resolution is a part of the primary task might suggest other approaches (such as presenting an ordered list of potential values or the most likely value together with some indication of confidence) [3, 11, 16], it would be inappropriate to introduce the potential for a Wikipedia visitor to see a value in an infobox without realizing the value is unverified. All of our designs present unverified information within dialogs clearly separate from and floating above article content. The "Check our guess" placeholder is used throughout our designs whenever space should be allocated to content that is currently unverified.

Inviting Contributions from Visitors

All of our designs include a callout above the infobox explaining the opportunity to help improve the infobox (see Figure 2 and Figure 4). We originally used a banner across the top of the page, but early talk-aloud participants were unsure how their actions were improving the page. We switched to the current callout to better draw attention to the infobox, consistent with the need to ensure people feel their contributions are important [1, 10, 15].

Popup Interface. Our first design is intended to solicit a greater number of contributions at the risk of being more obtrusive. It uses an *immediate* interruption coordination strategy [13], presenting a popup dialog as soon as a page is loaded. Dialogs are positioned adjacent to relevant content (as opposed to in the center of the browser). An immediate popup for each extraction in an article yielded an interface that was obviously too obtrusive. The first version tested in our informal talk-aloud sessions therefore randomly chose four non-overlapping popups and presented those when the article was loaded. Talk-aloud participants still considered this overly obtrusive, and so our current design displays only one of the four popups at a time. If a person contributes via the popup, the next of the four is presented, but no additional popups are presented if a person closes a popup. The popups are non-modal, repositionable, do not scroll the browser or request focus, and otherwise do not interfere with article content except for the area obscured

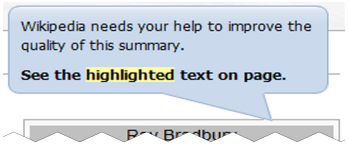

Figure 4: Each of our designs includes a callout drawing attention to the opportunity to help improve the article's infobox (see Figure 2 for the *icon* callout).

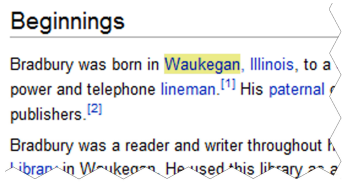

Figure 5: Our *highlight* design places a yellow highlight behind article text corresponding to potential extractions.

by the popup. Nevertheless, talk-aloud participants unanimously ranked this as the least acceptable interface.

Highlight Interface. Our second design is intended to better balance contribution rate against obtrusiveness. It uses a *negotiated* interruption coordination strategy [13], placing a yellow highlight behind text corresponding to potential extractions. Figure 5 illustrates this within the body of an article, and we also highlighted any "Check our guess" infobox placeholders. Mousing over either type of highlight presents a dialog allowing an indication of whether an extraction is correct. Responding to this dialog updates the infobox and removes any other obviated highlights.

Icon Interface. Our third design is intended to be the least obtrusive. It also uses *negotiated* interruption coordination [13], showing an icon for each potential extraction. These icons are placed on the left side of the infobox and along the left side of the article (as in Figure 2). Upon mousing over an icon, the appropriate article text is highlighted and a dialog allows an indication of whether an extraction is correct. As in the highlight design, responding updates the infobox and removes any icons obviated by the response.

The biggest difference between *highlight* and *icon* pertains to intrusiveness. *Highlight* displays its cues in the article's body (highlighting words within the article) while *icon* does not disturb the contents of the article (displaying icons on the periphery). Three of our nine talk-aloud participants ranked *highlight* as their favorite, while six chose *icon*.

Presenting Ambiguity Resolution in Context

Our designs take two approaches to providing context for verifying Kylin extractions. The first, illustrated in Figure 6, is presenting a dialog in the context of the article sentence from which Kylin obtained an extraction. We display the name of the infobox field in bold text and highlight the correspondence between the extracted value and the location of that value in the article. This dialog would look the same regardless of whether it was presented immediately as a part of the *popup* design, in response to mousing over the highlighted word "American" in the *highlight* design, or in response to mousing over an *icon* positioned off the left edge of Figure 6. There is also an

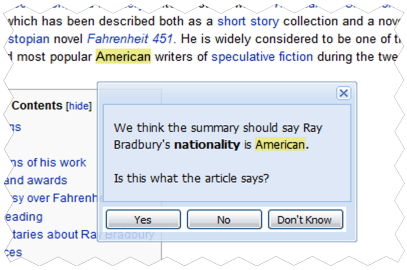

Figure 6: If interaction is initiated via the article, we annotate the extraction in the article and position our dialog to take advantage of the article context.

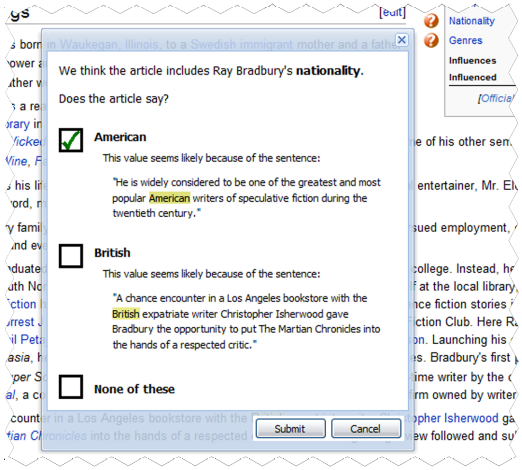

Figure 7: If interaction is initiated via the infobox, there may be different potential extractions at different locations. We therefore replicate the appropriate article context.

important subtlety in the wording of this dialog we revisit in discussing our Web search advertising deployments.

The second location our designs present extractions is in the infobox. It is important to take advantage of both locations to enhance the salience of opportunities to contribute, but the presentation of context in the infobox is more difficult. For example, Kylin may have identified multiple sentences in an article that suggest potential values for a field, these sentences may not be located near the infobox, and they may also not be located near each other. Figure 7 shows our approach, duplicating a small amount of context within the ambiguity resolution dialog. The dialog highlights the extracted value in each sentence, and a confidence metric is used to indicate whether the match is "likely" or just "possible". In this case, the dialog is visible because a person moved their mouse over the *icon* in the upper-right corner of the figure. *Highlight* works similarly, but this dialog is not immediately presented in the *popup* interface. As a part of reducing the obtrusiveness of the *popup* design, we instead present a smaller dialog indicating that potential extractions are available. Clicking in that dialog then presents the larger verification dialog.

WEB SEARCH ADVERTISING DEPLOYMENT STUDY

Although our talk-aloud sessions were critical to refining our designs, it is difficult to imagine a laboratory study convincingly demonstrating whether our interfaces increase

Figure 8: We used Web search advertising services to attract visitors to our pages. All of our visitors therefore had some other primary task, and we wanted to see whether they would spontaneously choose to contribute to Wikipedia.

spontaneous contributions. We therefore developed a novel method using Web search advertising services to deploy our interfaces as an actual non-primary task.

Procedure

We deployed a local Wikipedia mirror using a recent database dump and then randomly selected 2000 articles containing a *writer* infobox. We next used Kylin to extract values for the infobox fields. To ensure there would be an opportunity for contribution, we randomly removed up to ten existing infobox fields from the set which Kylin had extracted. This is appropriate for evaluating our designs, as we are not yet making actual edits in Wikipedia.

We then used two Web search advertising services (Google AdWords and Yahoo Search Marketing) to place ads for each of the 2000 writers. Figure 8, for example, shows an advertisement that appeared in response to a Google query for "ray bradbury" while our ads were active. Clicking on our ads directed people to our local Wikipedia mirror, where we could add our interfaces. Note that our ads intentionally do not mention *contributing* to Wikipedia. We therefore believe that all of the people who visited our pages had some other primary task motivating their visit. We deployed four interfaces: our *popup*, *highlight*, and *icon* designs as well as a *baseline*. The *baseline* included a callout (analogous to Figure 4) which prompted people to "Please edit this summary." Like Wikipedia's existing cleanup tags (see Figure 3), *baseline* did not highlight text or otherwise ease contribution. Visitors were assigned to interface conditions in a round-robin manner.

Our proxy injected JavaScript for the appropriate interface into each page. We used AJAX calls to log a unique session identifier and time-stamped events (including the firing of page load and unload events, clicks on components of our interfaces, and interaction with the normal Wikipedia presentation of edit functionality in the *baseline* condition). We also injected a short questionnaire into each page. This appeared as a popup 60 seconds after the page load event. It asked participants whether they saw they could help improve the quality of the article, how disruptive they considered the prompts in the article, whether they would be willing to use the interface as an addition to Wikipedia, and then provided a field for freeform comments. We used referral information to remove from our analyses any visits that did not originate from our ads (including visits by our team and by automated crawlers).

Deployments

We initially deployed our study using Google AdWords, receiving 1131 visitors. Examining the freeform feedback

from our survey revealed a potential misinterpretation of the wording used in our designs. Specifically, our initial dialogs said "*We think Ray Bradbury's nationality is American. Is this correct?*" Although we presented this in the context of the article and used highlighting to indicate the relationship to the article, we received comments like "*If I knew would I really need to look*" and "*Please check with the Britannica*" that underscored visitor feelings they were unable to contribute. In retrospect it is clear our initial wording can be interpreted as asking for factual validation, and so we clarified the wording to "*We think the summary should say Ray Bradbury's nationality is American. Is this what the article says?*" We then conducted a small Google AdWords test of our revised wording with the *icon* design, acquiring another 285 visitors. Satisfied with the results of our change, we redeployed our study using Yahoo Search Marketing, receiving another 1057 visitors.

The next section quantitatively analyzes the results of our deployments to demonstrate the synergy of our pairing of community content creation with information extraction, but we include this discussion to illustrate the challenges of designing for ambiguity resolution as a non-primary task. Further iteration could likely improve our designs, but we also believe significant future research is motivated by a need to better understand designing for non-primary tasks.

DEMONSTRATING SYNERGISTIC FEEDBACK

In order to show the synergy between community content creation and information extraction, we analyze our results from two complementary perspectives: (1) we show we *accelerate community content creation* by using Kylin's information extraction to significantly increase the likelihood that a person visiting a Wikipedia article as a part of some other primary task will spontaneously choose to help improve the article's infobox, and (2) we show we *accelerate information extraction* by using contributions collected from people interacting with our designs to significantly improve Kylin's extraction performance.

Accelerating Community Content Creation

Figure 9 summarizes the impact of our designs on the contribution rates of people who visited our pages as part of some other primary task. Recall *baseline* presented a callout asking visitors to help improve the quality of the infobox, analogous to Wikipedia's current cleanup tags, but did not leverage Kylin's information extraction to ease contribution. We analyzed contribution likelihood using chi-squared tests in a sequential Bonferroni procedure, finding that all of our designs result in a significantly greater likelihood of contribution than *baseline* (*icon*: $\chi^2_{(1,N=1345)} = 23.0$, $p < .001$, *highlight*: $\chi^2_{(1,N=1039)} = 53.0$, $p < .001$, *popup*: $\chi^2_{(1,N=1041)} = 55.4$, $p < .001$) and that *highlight* and *popup* yield a significantly greater likelihood of contribution than *icon* (*highlight*: $\chi^2_{(1,N=1432)} = 14.6$, $p < .001$, *popup*: $\chi^2_{(1,N=1434)} = 16.5$, $p < .001$). Analyzing the contributions per visit using Mann-Whitney tests in a sequential Bonferroni procedure finds the same differences (note that a large majority of people make no contributions, so finding the same differences is somewhat unsurprising).

	Baseline	Icon	Highlight	Popup
Visitors	476	869	563	565
Distinct Contributors	0	26	42	44
Contribution Likelihood	0%	3.0%	7.5%	7.8%
Number of Contributions	0	58	88	78
Contributions Per Visit	0	.07	.16	.14
Survey Responses	12	24	25	18
Saw I Could Help Improve	11/33 (33%)	30/73 (41%)	23/58 (40%)	24/52 (46%)
Intrusiveness (1:not – 5:very)	3.0	3.3	3.5	3.5
Willing to Use	11/33 (33%)	49/72 (68%)	34/57 (60%)	33/50 (66%)

Figure 9: Summarizing the results of a total of 2473 visits to Wikipedia articles during our deployments. All of our designs significantly improve the likelihood of contribution.

We believe the lack of contribution in *baseline* is typical of people who come to Wikipedia for some reason other than a pre-existing intent to contribute, as Wikipedia's current cleanup tags provide little or no assistance to potential contributors who are time-constrained or unfamiliar with Wikimarkup. To further validate our analyses, we examined typical Wikipedia contribution rates. We analyzed three months of recent Wikipedia log data and found that only 1.6% of Wikipedia visits involve editing. Although it is impossible to know how many of these begin as a non-primary task, this contribution rate includes, for example, the work of people who dedicate extended periods of time to contribution as a primary task as well as the work of people using tools designed to help experienced and motivated contributors quickly make large numbers of edits [20]. Also relevant is the fact 32% of edits were anonymous, meaning 0.5% of Wikipedia visits involve anonymous editing. There are many potential reasons for anonymous editing, including the possibility a person is sufficiently new to the community that they do not have an account and the possibility a person had not intended to edit and so had not logged in to their account.

Our designs clearly succeed in *accelerating community content creation* by leveraging Kylin's information extraction to obtain statistically significant improvements in contribution rates that herald practical implications. Every participant came to our pages with some other primary task, yet our *highlight* and *popup* designs, for example, yield an average of one contribution for every seven visits. Importantly, we obtained these results by emphasizing ease of contribution, not through coercion. Our results provide compelling initial evidence of the promise of using information extraction to identify opportunities for people to quickly and easily contribute to community content creation systems.

Each of our designs promotes contribution in a different manner. Although our focus is that all of our designs were successful in leveraging Kylin's information extraction to increase contribution, it is also interesting to consider

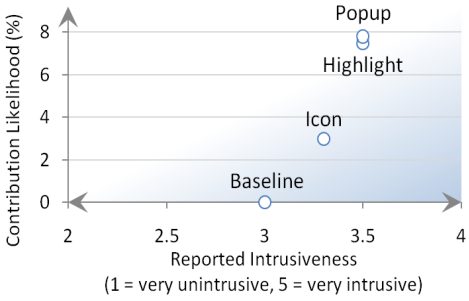

Figure 10: Contrasting contribution likelihood with reported obtrusiveness motivates future work exploring new designs that leverage information extraction.

differences in reactions to our designs. Figure 10 plots average intrusiveness (as reported by survey respondents) against the contribution likelihood for those designs. We see the apparent correlation here as a motivation for future work further exploring this tradeoff, as we believe significant opportunities remain to explore designs that leverage information extraction to solicit greater contribution rates without being perceived as intrusive. We also believe there are a number of questions to explore regarding how well different approaches will work with different types of data and in different communities.

Accelerating Information Extraction

Having shown that information extraction can amplify community content creation, we now demonstrate these contributions similarly improve information extraction. We first examined the reliability of the 224 community-created labels collected in our deployments. We removed 13 ambiguous labels, where our system had presented visitors with the entire sentence containing a correct value rather than the value itself. Of the remaining 135 extractions that visitors marked as correct, we found that 122 (90.4%) were indeed valid. Such high precision shows that making it easy for people to contribute does not necessarily mitigate quality. Of the 76 extractions that visitors indicated were incorrect, we found that only 44 (57.9%) were actually errors. This high false negative rate likely indicates people were conservative during validation, but may also be due in part to confusion over factual verification versus extraction validation (as discussed in the previous section).

We next examined whether these noisy community–created labels actually improve Kylin's information extraction performance. Because Kylin learns by analyzing existing infoboxes, we expected the impact of community-created labels to diminish if there were numerous existing infobox examples available. Thus, we test the effect of the community labels with models trained on 5, 10, 25, 50, and 100 existing infobox examples. We chose these numbers because 72% of infobox classes have 100 or fewer articles and 40% have 10 or fewer articles. Furthermore, Kylin frequently cannot obtain a training example for every field from an article containing an infobox. For example, an infobox may not contain a value for a field or there may not be a sentence in the article that matches the field. When

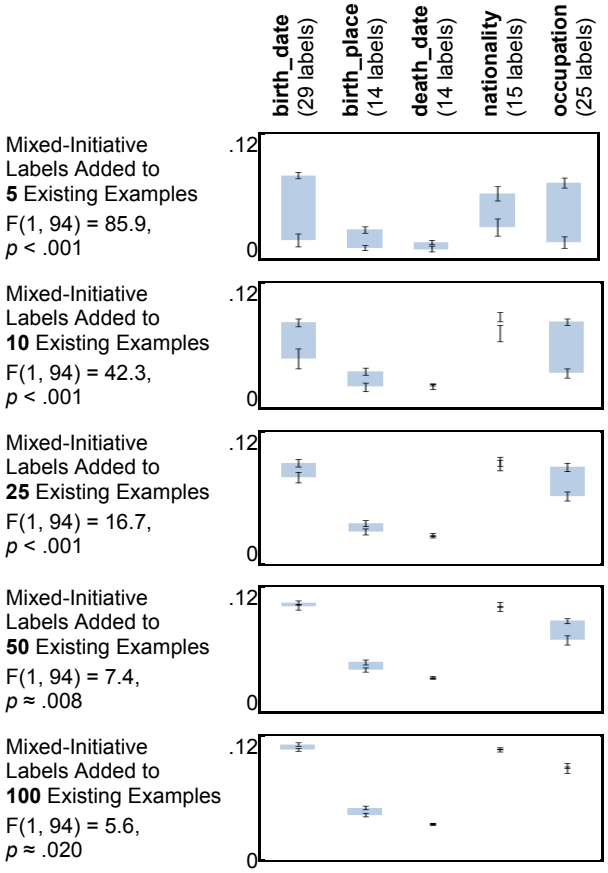

Figure 11: Adding our community-created labels to examples found from existing article infoboxes significantly improves Kylin's extraction performance in all five groups of trials (as measured by the area under the precision-recall curve). This impact is most dramatic when relatively few existing infoboxes are available. This is the typical case in Wikipedia, where 72% of inbox classes appear in less than 100 articles.

given an article with a *writer* infobox, for example, Kylin was able to generate a positive training example for only an average of 14.5% of the attributes.

The performance of any individual extractor is difficult to interpret, and the performance of extractors for different infobox fields cannot be directly compared. The nature of sentences containing a birthdate, for example, may make that date more (or less) difficult to extract than a person's nationality. We conducted our experiment with the five *writer* fields for which we obtained the greatest number of positive examples during our Web advertising deployment. We trained Kylin using a set of randomly-selected existing infobox examples for each field and tested the resulting extractor against 200 articles, the fields in which we had manually labeled. We then added the community-created labels (both correct and incorrect) and repeated the test. To minimize errors caused by sampling, we repeated this process for ten trials with different initial infobox examples. As an outcome measure, we chose the *area under the precision-recall curve*, a common summary statistic for information extraction performance.

Figure 11 summarizes our results. We show standard error bars around the mean area under the precision-recall curves from the ten trials. The means are connected by a wide blue bar whenever a paired *t* test indicated a statistically significant improvement from the addition of the community-created labels. Because we have noted that different fields can vary in the difficulty of their extraction, our analysis focuses on the impact of adding community-created labels to Kylin extractors trained on differing numbers of existing infobox examples. We analyze each group of trials using a mixed-model analysis of variance, treating field as a random effect. As reported in Figure 11, our analyses show that the addition of community-created labels significantly improves the area under the precision-recall curve in all five groups of trials.

These results provide strong initial evidence of the second half of our synergy, that despite containing errors, community-created labels *accelerate information extraction* by significantly improving Kylin's extraction performance. Our community-created labels most dramatically improved extraction performance when relatively few existing infobox examples were available, and we have noted that, indeed, most Wikipedia infobox classes have relatively few existing examples available.

RELATED WORK

Prior work has explored why people contribute to Wikipedia and what the implications of those motivations are for the Wikipedia culture [1, 10, 15]. Throughout this paper we have discussed designing for a culture that is motivated by altruism, supporting free access to knowledge for everyone, reputation, and indirect author recognition. Work by Cosley *et al.* [2] has examined the problem of finding appropriate articles for experienced contributors to work on, based in the idea that a person's editing history provides insight into what other articles they might be interested in editing. Instead of targeting experienced Wikipedia contributors, we leverage information extraction to solicit contributions from people who otherwise would be unlikely to contribute. Our approach frees experienced contributors to focus on more challenging work and provides a path for newcomers to become active members, consistent with Bryant *et al.*'s finding that newcomers become members of the Wikipedia community by participating in peripheral yet productive tasks that contribute to the overall goal of the community [1].

DeRose *et al.* [4] and Doan *et al.* [12] each propose different approaches to building communities based on both human and automated contributions. DeRose *et al.* build their MadWiki system with structured slots, which are attribute/value pairs expressed in terms of paths over entity relationship representations of database schemas. Doan *et al.* focus on schema matching, examining several design dimensions and proposing that one approach to obtaining contributions might be to require some amount of work before providing a service. In addition to examining a different type of inference, we take a different perspective by placing the needs and norms of the existing Wikipedia community on equal footing with the needs of information extraction systems. We work with Wikipedia's existing content format and community norms because we believe the full benefits of pairing information extraction with community content creation can be realized only by reinforcing both feedback cycles.

Von Ahn and Dabbish's *games with a purpose* channel player entertainment into productive work, such as labeling images through a game in which players guess what words other players will guess in response to an image [17, 18]. Although the verification of information extraction does not easily fit into the game templates described by von Ahn and Dabbish [18], we believe the more interesting contrast is a difference in perspective regarding community contribution. Von Ahn and Dabbish's games are ultimately a deception, disguising work as a game. In contrast we highlight the opportunity to contribute meaningful work to a community, leveraging our synergy with information extraction to make contribution easy. The approaches are clearly complementary, but we believe it is notable that our approach provides a path for newcomers to become active members of a community, at which point they may choose to take on more challenging work (work that cannot be reduced to a game or to a handful of clicks).

Prior work has explored interfaces for ambiguity resolution, including Mankoff *et al.*'s OOPS [11], Shilman *et al.*'s CueTIP [16], and Culotta *et al.*'s examination of corrective feedback in information extraction [3]. Such work has focused on ambiguity resolution as a part of the primary task, at least in the sense that the primary task cannot continue until the ambiguity is resolved. Although we leverage many of the techniques developed in such work, we have also shown that designing for ambiguity resolution as a non-primary task introduces new challenges.

DISCUSSION AND CONCLUSION

This paper proposes a novel synergistic method for jointly amplifying community content creation and learning-based information extraction. By enabling both techniques to exploit the same edits, two interlocking feedback cycles accelerate each other. We have demonstrated this synergy with two complementary analyses: (1) we show we *accelerate community content creation* by using Kylin's information extraction to significantly increase the likelihood of contribution by people visiting Wikipedia while engaged in some other primary task, and (2) we show we *accelerate information extraction* by using community-created contributions as training examples to significantly improve Kylin's extraction performance. Taken together, these analyses provide initial but compelling evidence of our proposed synergy.

Our use of Web search advertising services was a powerful way to expose people to our interfaces as a non-primary task, but it is clear that future work needs to address the more complete integration of our approach into a variety of sites. For example, community question and answer sites (Yahoo Answers), product review sites (Amazon), and

other Wiki sites (Wiktionary, Wikitravel) could improve presentation and search with more structure, but relevant extraction techniques are currently far from perfect. Similarly, extraction-based bibliographic sites (Citeseer) and aggregation sites (InfoZoom) suffer from incomplete or inaccurate information and could benefit from increased community contributions. Our ultimate goal is to add appropriate hooks to platforms like MediaWiki [14] (upon which Wikipedia and many other sites are implemented) so that anybody visiting such a site can be presented with mixed-initiative contribution opportunities. In the shorter term, MediaWiki includes support for people to request inclusion of a JavaScript file in every page they visit, so it will be possible to build a community of early adopters.

Motivations for contribution may vary among different Web communities, but our methods suggest that our approach generalizes beyond Wikipedia. Our contributions were not solicited from people in the Wikipedia community (which may be a somewhat atypical Web community), but were instead solicited from people using the Web in their everyday primary tasks (due to our use of Web search advertising services). Prior research on contribution to different types of community content sites also reveals many similarities: contribution is often highly skewed, and so successful sites need to provide value to visitors, make the need for contribution visible, ensure it is easy to contribute (especially for newcomers), and ensure people perceive the contributions as meaningful, all principles that our synergistic approach is designed to address.

ACKNOWLEDGEMENTS

We thank all of our study participants and our Wikipedia interviewees. We also thank Eytan Adar, Ivan Beschastnikh, Oren Etzioni, Travis Kriplean, and the 2008 CSE 574 participants for helpful comments and suggestions. We appreciate the efforts of Evgeniy Gabrilovich and Ken Schmidt who greatly facilitated our second Web-advertising study. This work was supported by Office of Naval Research grant N00014-06-1-0147, CALO grant 03-000225, NSF grant IIS-0812590, the WRF / TJ Cable Professorship, a UW CSE Microsoft Endowed Fellowship, a NDSEG Fellowship, a Web-advertising donation by Yahoo, and an equipment donation from Intel's Higher Education Program.

REFERENCES

1. Bryant, S.L., Forte, A. and Bruckman, A. (2005). Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. *Proceedings of the ACM Conference on Supporting Group Work* (GROUP 2005), 1-10.
2. Cosley, D., Frankowski, D., Terveen, L. and Riedl, J. (2007). SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia. *Proceedings of the International Conference on Intelligent User Interfaces* (IUI 2007), 32-41.
3. Culotta, A., Kristjansson, T., McCallum, A. and Viola, P. (2006). Corrective Feedback and Persistent Learning for Information Extraction. *Artificial Intelligence* **170**(14). 1101-1122.
4. DeRose, P., Chai, X., Gao, B., Shen, W., Doan, A., Bohannon, P. and Zhu, J. (2008). Building Community Wikipedias: A Human-Machine Approach. *Proceedings of the IEEE International Conference on Data Engineering* (ICDE 2008), 646-655.
5. Giles, C.L., Bollacker, K. and Lawrence, S. (1998). CiteSeer: An Automatic Citation Indexing System. *Proceedings of the ACM Conference on Digital Libraries* (DL 1998), 89-98.
6. Grudin, J. (1994). Groupware and Social Dynamics: Eight Challenges for Developers. *Communications of the ACM* **37**(1). 92-105.
7. Hoffmann, R., Fogarty, J. and Weld, D.S. (2007). Assieme: Finding and Leveraging Implicit References in a Web Search Interface for Programmers. *Proceedings of the ACM Symposium on User Interface Software and Technology* (UIST 2007), 13-22.
8. Horvitz, E. (1999). Principles of Mixed-Initiative Interfaces. *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI 1999), 159-166.
9. Huynh, D.F., Miller, R.C. and Karger, D.R. (2006). Enabling Web Browsers to Augment Web Sites' Filtering and Sorting Functionalities. *Proceedings of the ACM Symposium on User Interface Software and Technology* (UIST 2006), 125-134.
10. Kuznetsov, S. (2006). Motivations of Contributors to Wikipedia. *ACM Computers and Society* **36**(2). 1-7.
11. Mankoff, J., Hudson, S.E. and Abowd, G.D. (2000). Interaction Techniques for Ambiguity Resolution in Recognition-Based Interfaces. *Proceedings of the ACM Symposium on User Interface Software and Technology* (UIST 2000), 11-20.
12. McCann, R., Shen, W. and Doan, A. (2008). Matching Schemas in Online Communities: A Web 2.0 Approach. *Proceedings of the IEEE International Conference on Data Engineering* (ICDE 2008), 110-119.
13. McFarlane, D.C. (2002). Comparison of Four Primary Methods for Coordinating the Interruption of People in Human-Computer Interaction. *Human-Computer Interaction* **17**(1). 63-139.
14. *MediaWiki*. http://www.mediawiki.org/.
15. Priedhorsky, R., Chen, J., Lam, S.T., Panciera, K., Terveen, L. and Riedl, J. (2007). Creating, Destroying, and Restoring Value in Wikipedia. *Proceedings of the ACM Conference on Supporting Group Work* (GROUP 2007), 259-268.
16. Shilman, M., Tan, D.S. and Simard, P. (2006). CueTIP: A Mixed-Initiative Interface for Correcting Handwriting Errors. *Proceedings of the ACM Symposium on User Interface Software and Technology* (UIST 2006), 323-332.
17. von Ahn, L. and Dabbish, L. (2004). Labeling Images with a Computer Game. *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI 2004), 319-326.
18. von Ahn, L. and Dabbish, L. (2008). Designing Games with a Purpose. *Communications of the ACM* **51**(8). 58-67.
19. Voss, J. (2005). Measuring Wikipedia. *International Conference of the International Society for Scientometrics and Informetrics* (ISSI 2005), 221-231.
20. *Wikipedia: AutoWikiBrowser*. http://en.wikipedia.org/wiki/Wikipedia:AutoWikiBrowser.
21. *Wikipedia: Be Bold*. http://en.wikipedia.org/wiki/Wikipedia:Be_Bold.
22. *Wikipedia: Bot Policy*. http://en.wikipedia.org/wiki/Wikipedia:Bots.
23. *Wikipedia: Cleanup Tags*. http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Cleanup.
24. Wu, F., Hoffman, R. and Weld, D.S. (2008). Information Extraction from Wikipedia: Moving Down the Long Tail. *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining* (KDD 2008), 731-739.
25. Wu, F. and Weld, D.S. (2007). Autonomously Semantifying Wikipedia. *Proceedings of the ACM Conference on Information and Knowledge Management* (CIKM 2007), 41-50.
26. Yee, K.-P., Swearingen, K., Li, K. and Hearst, M. (2003). Faceted Metadata for Image Search and Browsing. *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI 2003), 401-408.