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

Modern web search and browsing interfaces often leverage the structure of Web content. In our work we explore the synergistic pairing of information extraction and community content creation as two interlocking feedback cycles for generating structured information. 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 and social communities. We then develop and explore several approaches to inviting contribution to a community, each presenting ambiguity resolution as a non-primary task.

ACM Classification:

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

MIXED-INITIATIVE INFORMATION EXTRACTION

Two methods have proven successful at generating structured information from the internet: machine-learning based information extraction and communal content creation. *Information extraction* (IE) has demonstrated incredible success (e.g., Google Scholar), but it has important limitations. First, machine learning algorithms typically require numerous labeled training examples, and these can be expensive to collect. Second, statistical approaches can be error prone; systems with precisions of only 80-90% are considered successful. The second approach for creating structured information is *community content creation* (CCC), as used in creating Wikipedia. Despite examples of successful community approaches, bootstrapping to critical mass and overcoming difficulties related to work/benefit disparities can be challenging (e.g., the vast majority of Wikipedia work is done by a relatively small set of people).

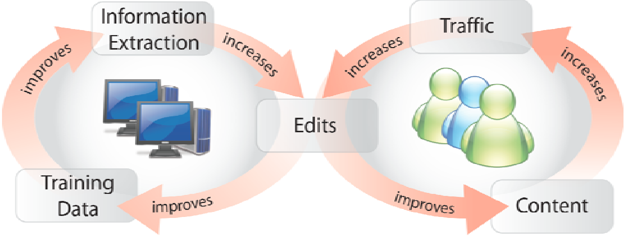
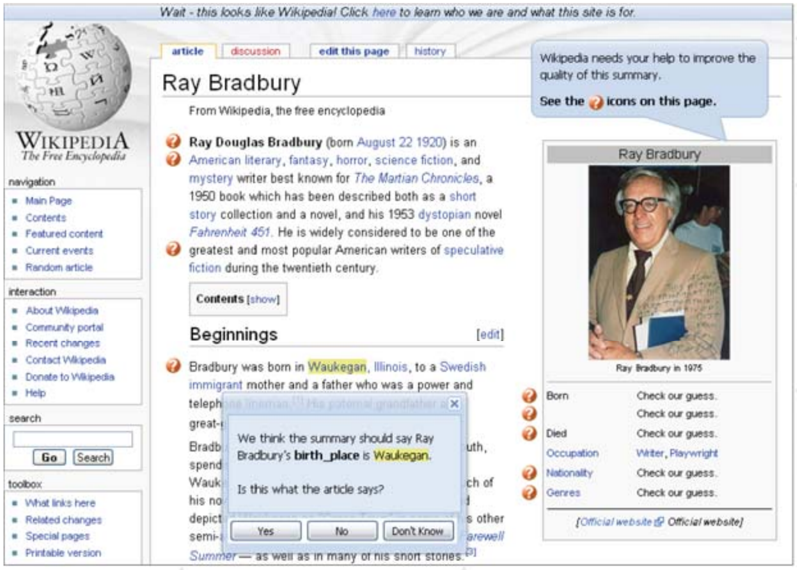


Figure 1: We envision the synergistic pairing of automated information extraction with community content creation, using the same edits to accelerate both feedback cycles.

We believe that these techniques are complementary [1]. For example, IE can be used to bootstrap content on a site to attract traffic and CCC can be used to correct errors, improve training data, and enable a virtuous cycle as shown in Figure 1. But surprisingly there has been almost no work

Figure 2: An example page containing several opportunities for a person to easily contribute 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.


Wait - this looks like Wikipedia! Click here to learn who we are and what this site is for.

Wikipedia needs your help to improve the quality of this summary.
See the ? icons on this page.

Ray Bradbury

From Wikipedia, the free encyclopedia

Ray Douglas Bradbury (born August 22 1920) is an American literary, fantasy, horror, science fiction, and mystery writer best known for *The Martian Chronicles*, a 1950 book which has been described both as a short story collection and a novel, and his 1953 dystopian novel *Fahrenheit 451*. He is widely considered to be one of the greatest and most popular American writers of speculative fiction during the twentieth century.

Contents [show]

Beginnings [edit]

Bradbury was born in Waukegan, Illinois, to a Swedish immigrant mother and a father who was a power and teleph... great-... Bradb... uth, spend... Wauk... ch of his no... depict... s other semi-... arewell *Summer* — as well as in many of his short stories.[3]

We think the summary should say Ray Bradbury's **birth_place** is Waukegan.

Is this what the article says?

Yes | No | Don't Know

Ray Bradbury

Ray Bradbury in 1975

Born	Check our guess.
	Check our guess.
Died	Check our guess.
Occupation	Writer, Playwright
Nationality	Check our guess.
Genres	Check our guess.

[Official website Official website]

A callout draws attention to the editing icons.

A Wikipedia infobox provides a summary of the key attributes of an article.

Icons by infobox attributes provide an overview of extracted attribute values. Mousing over an icon opens a dialog for choosing.

Icons allow the user to quickly find extraction sites, without reducing article readability.

When mousing over an icon in the article, a popup prompts for feedback. The extraction is then highlighted in the article.

UIST'08, October 19–22, 2008, Monterey, California, USA
ACM 978-1-59593-975-3/08/10.

	Baseline	Popup	Highlight	Icon	Revised Icon
Visits	234	274	276	293	209
Contributions Per Visit	0	.19	.12	.06	.13
Intrusiveness (1:not – 5:very)	2.55	3.46	3.76	3.53	3.20
Willing to Use	7/11 (64%)	15/24 (63%)	13/20 (65%)	12/16 (75%)	11/15 (73%)

Figure 3: Results of 1286 visits to pages with our interfaces.

aimed at combining the methods, looking for additional synergies, or discerning the principles of combination. We explore methods for combining these techniques in the context of the Wikipedia community and the Kylin information extraction system [2]. More specifically, we focus on the creation of infoboxes, tabular summaries present in many Wikipedia pages (Figure 2). Kylin analyzes relationships between infoboxes and the text of corresponding Wikipedia pages, learning to extract new attribute values from the many untagged pages.

To identify and examine the most important aspects of integrating IE with CCC, we interviewed members of the Wikipedia community, developed three interfaces exploring different points in the design space, conducted think-alouds, and deployed our interfaces via Adwords.

USER-INTERFACE DESIGN DIMENSIONS

Space constraints prohibit a detailed discussion of our design dimensions, but all interfaces use a mixed-initiative approach, presenting inferred potential contributions within the normal context of Wikipedia pages with the goal of soliciting human feedback regarding whether an automatically inferred contribution is correct and can therefore be published into Wikipedia. An important focus is on contribution as a non-primary task, as we believe that our system will be most effective if it encourages contributions by people who had not otherwise planned to contribute. An important aspect of treating contributing as a non-primary task is the fact that many people will never even notice the *potential* to contribute. A design principle that therefore emerged in our process is that unverified information should never be presented in such a way that it might be mistakenly interpreted as a part of the page. This also raises the challenge of how to appropriately incentivize contribution. We note that Wikipedia's community culture is based on altruism and supporting free access to knowledge for everyone, and is incompatible with some approaches to soliciting contributions (such as requiring people to provide a small amount of work before gaining full access to a service). Our goal is to make the ability to contribute sufficiently visible that people will choose to contribute, but not so visible that people feel an interface is obtrusive and attempting to coerce contribution. We designed three interfaces to explore this tradeoff.

Our *Popup Interface* is intended to solicit a greater number of contributions at the risk of being more obtrusive. It uses an *immediate* interruption coordination strategy, presenting a popup dialog as soon as a page is loaded. The popups are non-modal, do not scroll the browser or request focus, and otherwise do not interfere with any page content except for the area obscured by the popup. Our *Highlight Interface* is intended to better balance visibility against obtrusiveness. It uses a *negotiated* interruption coordination strategy, placing a yellow highlight behind text in the page corresponding to potential extractions. Mousing over either type of highlight presents a dialog. Our *Icon* and *Revised Icon Interfaces* (Figure 2) are intended to be minimally obtrusive. They also use a *negotiated* strategy, placing icons within the page at locations of each extraction. Upon mousing over an icon, the extraction is highlighted and a dialog is presented. They differ in the wording of messages.

GOOGLE ADWORDS DEPLOYMENT STUDY

To explore whether our interfaces will lead people to spontaneously contribute, we loaded 2000 articles to a local Wikipedia mirror and placed an ad for each article using the Google AdWords service. During the course of our studies, for example, a Google query for 'ray bradbury' showed an advertisement for the corresponding Wikipedia article. Clicking upon this ad directed people to our local mirror, where we added our interfaces by injecting Javascript into the Wikipedia pages. Our ad intentionally does not mention contributing to Wikipedia. We believe that all of the people who visited our pages therefore had some other primary task motivating their visit. We also injected a short questionnaire into each page, which appeared as a popup 60 seconds after the page loaded.

Results. Figure 3 summarizes 1286 visits. All interfaces prompted a significantly greater percentage of people contributing than the *baseline* condition's callout (*baseline* is analogous to the cleanup tags that Wikipedia currently uses) (popup: $\chi^2_{(1,508)}=22.4$, $p<.001$, highlight: $\chi^2_{(1,510)}=17.6$, $p<.001$, icon: $\chi^2_{(1,527)}=7.3$, $p<.01$, revised icon: $\chi^2_{(1,443)}=11.5$, $p<.001$). The *popup* interface also resulted in a significantly greater percentage of people contributing than *icon* ($\chi^2_{(1,N=567)}=9.2$, $p < .01$), but there is no detectable difference between *highlight* and either other interface. We consider these initial results to be quite promising. Our *revised icon* interface resulted in an average of one small piece of work for every eight visitors to a page. We were able to obtain these contributions from people who were focused on some other primary task and had not come to our page to contribute. Perhaps most importantly, we obtained these results by emphasizing the visibility and ease of contributing, not by resorting to coercion.

To assess the quality of the information provided by people, we manually examined the extractions that people indicated were correct and found that 90% were indeed correct. This high precision shows that making it easy for people to contribute does not necessarily mitigate quality.

REFERENCES

1. Weld, D.S., Wu, F., Adar, E., Amershi, S., Fogarty, J., Hoffmann, R., Patel, K., and Skinner, M. (2008). Intelligence in Wikipedia. AAAI 2008.
2. Wu, F., Hoffmann, R. and Weld, D.S. Information Extraction from Wikipedia: Moving Down the Long Tail. KDD 2008.