Design Patterns for Crowdsourced Document Annotation

Steve Berry
University of Texas School of Information
4702 Fieldstone Drive
Austin, TX 78735 USA
+1 512 914 9994
sberry@unideal.net

ABSTRACT

Document annotation plays an important role in semantic metadata assignment for articles and other published content but often falls outside of a publisher's standard editorial workflow process. In this paper we explore how crowds might be leveraged to perform this work as well as methods to incentivize those workers. We build a prototype annotation system that targets specific crowds, applies design patterns borrowed from video games, a targeted incentive system, and a combined set of refined aesthetic and interaction design principles to create and engaging user experience for annotation workers. We then propose methods for evaluating this system and discuss future work.

Keywords

Document Annotation, Interaction Design, Game Principles

1. INTRODUCTION

Production of descriptive and categorical metadata is an important method of adding value to digital content. This extra layer of semantic information can be used to improve accessibility, aid in the accuracy of information retrieval systems, disambiguate entities contained in documents and also provide links between those documents. Much progress has been made in Natural language processing both by search engine providers and entity extraction software packages but the accuracy of these systems is limited by the training of their algorithms and lack of expert training for categorizing particular domains of documents. Humans are still by far the most reliable judges of nuanced semantic meaning buried in documents but the cost of human annotation by experts often makes it infeasible to implement.

In the past decade crowdsourcing has emerged as an effective method for breaking down complex tasks and distributing them to pools of non-expert workers in the form of microtasks on platforms like Mechanical Turk.

Crowdsourced systems have been proven to be an effective approach to producing annotations at a lower cost then traditional expert annotation. Leveraging these crowdsourcing methods does, however, introduce a complex set of issues to be addressed in interface and incentive design. In the following section we will review the specific problems we attempt to address in our annotation prototype.

2. ISSUES IN CROWDSOURCING APPLICATIONS

2a. Low Quality Work

Microtask platforms like Amazon's Mechanical Turk and Crowdflower are powerful resources for task requesters to access a broad global population of crowd workers. Unfortunately, the inherent anonymity and systems of monetary reward incentives in these platforms makes them susceptible to gaming and contributes to the production of biased or inaccurate results of low quality [6]. This annotation quality issue may be mitigated through costly redundant verification processes [6], inclusion of "Gold Standard" questions [9], as well as simple manual review of work by task requesters but these verification methods may be difficult to implement and the increased cost may offset the gains of using crowdsourcing in this type of annotation work

2b. Worker Expertise

It's been shown that amateur crowd workers can effectively be employed to perform such complex domain-specific tasks as building a concept hierarchy for the discipline of Philosophy [5] or even mapping protein structures [3]. Successful employment of a non-expert worker pool is made possible through microtask platforms which provide some capability for limiting worker participation to qualified workers suited to requester tasks through the use of screening "Qualification Tests". These tests have utility but they are only partially effective in eliminating low quality work [6].

Some researchers have created innovative ad-hoc methods of promoting and retaining a pool of trusted crowd workers based on demonstrated performance through the use of a tiered incentive scale which rewards certain manually selected workers based on the quality of their work [1].

Successful employment of a non-expert worker pool may also be accomplished through the development of skillfully customized interfaces which break down the task to units of work that diminish the need for worker expertise to solve computation problems. This task-specific approach is used effectively in applications like Foldit [3] and OntoGame [13] but can create a need for intense specialized design and development to properly enable that work.

Finally, requesters are finding it increasingly more attractive to bypass the expertise problem altogether through use of pools of pre-verified or expert workers to improve result quality. New businesses are emerging to provide on-demand expert workforces for vertical problem spaces in the form of 'Vertical Crowds'. [7]

2c. Ethical Issues and Work Transparency

A frequently cited ethical issue inherent in anonymous crowd work platforms is the disintermediation of the requester/worker relationship. Workers may unwittingly be exposed to spam-oriented tasks that exploit the crowd to perform illegal or unethical work [8].

A further complication of this disintermediation and lack of transparency may also make the eventual end result of the workers' labor opaque. This opacity introduces the potential for workers to unknowingly perform work for organizations or goals they might normally be philosophically opposed to.

2d. Incentive Design

Incentive design can have an important self-selection effect on crowd composition. Tailoring the reward system to appeal to a specific population of workers can greatly offset issues that occur on platforms that focus solely on a basic monetary compensation in exchange for workers' effort. Small changes to incentives or their delivery method can have dramatic effects on participation levels and crowd composition. The txtEagle project recognized that cell phone minutes were an effective reward system for the crowd they were attempting to engage, increasing participation in their system [4]. Amazon's decision to distribute rewards in Rupees may be a contributing factor in their significant worker demographic shift towards Indian workers [10][12].

In addition to typical reward systems, researchers have had great success in adopting game mechanics used for years in video games to motivate workers with an engaging and entertaining work experience. In some cases this experience alone can displace the need for any type of monetary reward system entirely. Crowdsourcing projects have also shown that workers will donate their time to projects which they believe align with their own interests or help to achieve some greater altruistic goal like classifying galaxies [11] or furthering research in biology [3].

2e. Interface Design

Usability and clarity of interface design can have a significant impact on the level of engagement and quality of work performed by crowd workers. Poorly designed and untested interfaces can lead to misunderstanding of task goals and lack of proper validation methods can introduce low quality data in spite of the best intentions on the part of workers. Researchers employed by Amazon Mechanical Turk cite a lack of tested design patterns as a critical research area for improvements in quality and participation from workers [2]. Attention to interaction quality and aesthetics may influence perceived usability of interfaces and these factors should receive strong consideration when creating worker interfaces [14]

3. PROTOTYPE CONSTRUCTION

3a. Goals and participating publishers

The goal of our prototype annotator system is to demonstrate an application that leverages self-selecting semi-expert crowds to annotate documents for domains that require specific tacit knowledge of a subject. We apply multiple design principles to counteract the inherent issues of crowd tasks to arrive at a system that could be deployed by content publishers. Two publications were targeted as the test cases for the prototype:

Hemmings Motor News' Sports & Exotics publication. This publication's articles consist of lengthy editorial discussions of collectable import vehicles and the process of maintaining, buying and selling those vehicles. Annotators of these documents will need to have knowledge of specific terms, tools, and classifications found in the collector car hobby that would not be considered common knowledge.

The Austin Business Journal – This weekly publication produces articles that discuss business news and events specific to the Austin Metropolitan area. Annotators of these documents would ideally have a specific geographical and business domain understanding to accurately annotate these documents.

These two publications were selected primarily because of their domain-specific content and their large, active and enthusiastic reader audiences who routinely engage with these publications on and offline as candidates for recruitment of a publisher-specific "vertical crowd". We address this through the use of a tailored incentive system designed for the individual publications.

The primary goal of the annotation interface is to allow workers to explicitly and accurately specify semantic entities and concepts in these document sets that may be later leveraged by publishers for purposes of document linking, information retrieval accuracy or training natural language extraction algorithms. The core interface of the system presents the worker with the text of an article and

tools to positively identify what entities and concepts exist within that document.

In order to reduce the cognitive load of generating these annotations we created a base set of possible entities by submitting the document text to Reuters' Open Calais extraction API. This web service processes document text and returns a set of entities found within the text along with a score reflecting the confidence that the entity is relevant to the document. The accuracy of this automated process is exceptional but also contains many false-positives and ambiguous results. We present these results to the user as a basis for annotating the document, allowing them to easily confirm or discard the annotations. Additionally, we allow the users to add entirely new annotation entries that may not have been detected by the automated extraction process [Figure 1].

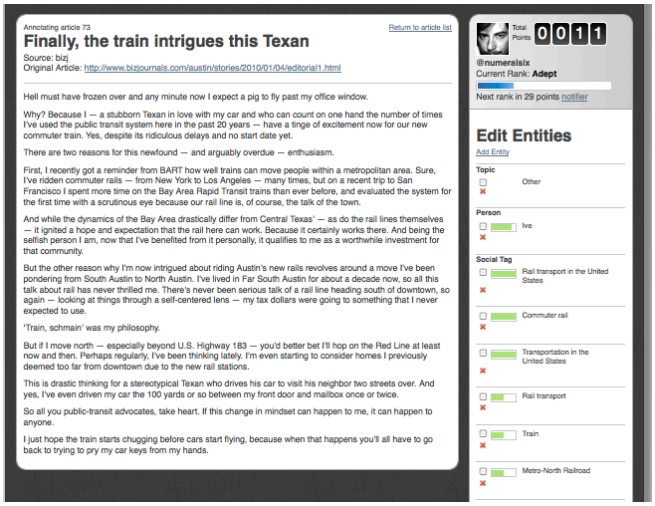

Figure 1. Annotation Screen

4a. Establishing Worker Identity

The system intentionally eliminates anonymity by tying worker annotations to an online identity. For the purposes of the prototype we use Twitter accounts as the authority for this identity [Figure 2] but in a production system this identity could instead refer to the publisher's internal user account records or any external social profile. This authentication scheme should be configurable to suit the needs of the publisher and their audience. Furthermore, some additional level of identity verification or skills test could accompany the establishment of an annotator user before they are permitted to participate. The prototype forgoes this level of complexity for the purposes of demonstration.

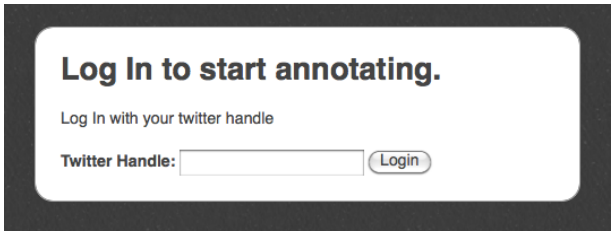

Figure 2. Simple Twitter Authentication

Once an identity is established in the system it is used as the basic user record for allotting points, rank, badges and rewards from publishers. The identity and associated attributes persist across annotation sessions.

4b. Points and Incentives

The core incentive system in the annotator interface centers on the accumulation of 'points ' for work done that mimics the points systems found in video games [Figure 3]. As users confirm, dismiss or add to the base extracted entity annotations for an article they are rewarded a 'point' for each discrete annotation action. Longer articles covering more concepts will naturally require more review and offer the possibility of more points. Once all suggested entities and user-provided entities are handled the article is considered completely annotated the user is rewarded an additional 3 points and encouraged to continue annotating other articles. The workers' progress is saved so they may pause and return to the article mid-annotation.

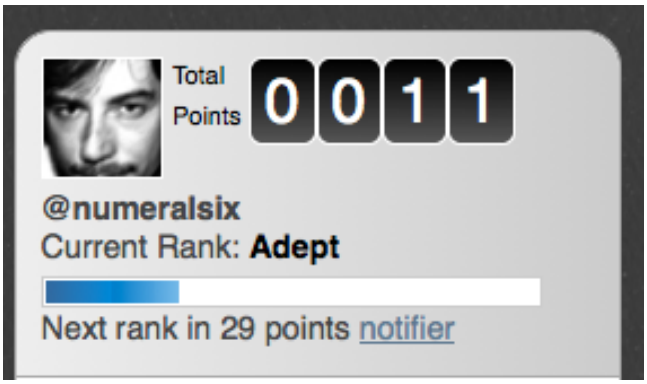

Figure 3. Persistent Points Scoreboard

Any available article may be annotated by any approved user in the system exactly once. Across users, each is shown the default state of the document without awareness of annotations performed by previous users that may introduce annotation bias or support spam agreement behavior. The publisher may configure the maximum number of times an article may be reviewed by users until it is automatically removed from the pool of documents available for annotation, tuning it until they find an ideal balance of verification vs. cost. Once removed from the pool, the document state is marked as "Annotation Complete, Needs Admin Review".

The administrator of the system is presented with an interface showing the combined results of the annotation work for review and approval. This interface presents the annotations in order of agreement allowing the administrator to quickly confirm entities with high agreement and focus on manually verifying those with low agreement. The administrator approved set of annotations is then finalized and stored with the original document. During this review process the administrator has an

opportunity to judge worker behavior and eject workers who appear to be gaming the system or providing low quality annotations to prevent them from future participation. We anticipate that the complexity of these decisions will differ depending on the level of controversy inherent in a particular article or domain.

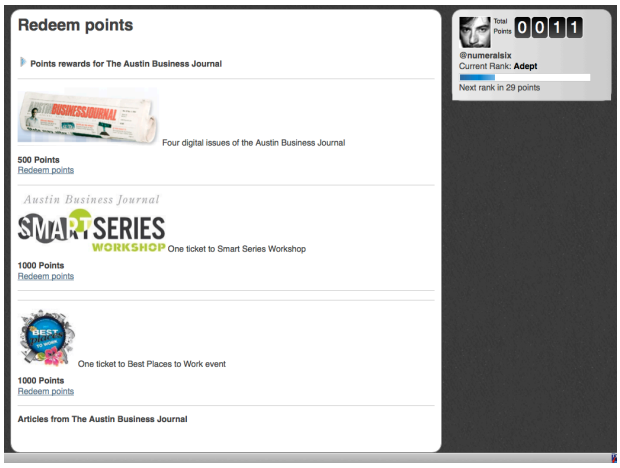

Figure 4. Points Redemption

The publication-specific rewards offered by the system [Figure 4] are designed to attract a group of semi-expert workers with some pre-existing knowledge of the domain. For the demonstration we created specific items which may be rewarded to workers in exchange for points accumulated during annotation. The publishers may offer access to content paywalls, publisher-specific merchandise or tickets to real world events. The ambiguous value of a 'point' allows publishers to independently decide what the value of a single annotation is to their organization and designate the redemption value of the prizes accordingly. As a side-effect of this reward system there should be a large degree of self-selection for workers already engaged with these publications who are more likely to enjoy reading the articles and directly benefit from information obtained while working

4c. Game Elements: Rank and Badges

As workers accumulate points they graduate through a publisher-configured 'rank' system. At some predefined number of points the worker can graduate to a new ranking level, progressing through a system of ranks (example multiple steps graduating from "Intern" at 10 points to "CEO" at 5000 points in a business publishing domain) . These ranks will be consistently displayed alongside the worker identity as a measure of reputation. A progress bar is shown that updates as each point is awarded and makes users aware of the number of points needed to reach the next rank. This sense of progress is designed to encourage the worker to continue annotating to achieve new ranks.

In addition to the pursuit of rank, the system is configured to recognize milestone events in the workers' progress. Examples include "First Article Annotated" or "200 Annotations Made". Upon reaching these milestones the system flashes a notification modal informing them of the badge and congratulating them. These surprise badge events are meant to periodically reward the user with positive reinforcement for their annotation activity.

Rank and badge rewards are intended reinforce worker progress and project reputation within the system. Ideally, a final published article would give attribution to the workers that assisted in the annotation of that article, displaying avatars, rank and accumulated badges of those users alongside the content. This is designed to give the workers a sense of ownership and reputation within the publisher domain as a reward for participating.

4c. Design: Aesthetics and Interaction Quality

The visual and interaction design quality of the interface consumed a large portion of the development effort for the prototype. The interface is clean, well-constructed and hopefully highly usable. The system is free from any non-essential distracting information, providing a modal experience for the workers. Controls are responsive ajax operations with appropriate feedback. It's difficult to quantify or evaluate the impact of time spent in this area beyond the research cited in the introduction, but a 'fun' experience was one of the primary objectives we pursued in construction

5. EVALUATION

Limited evaluation of this prototype has been conducted so far and the application should currently be considered the first iteration of a pilot interface. The prototype has been demonstrated and discussed with peers in the Information, Publishing and IT communities. Response and constructive feedback from these demonstrations have been very positive so far and done much to inform the current design of the incentive system and possible real-world deployment scenarios. The words "fun", "beautiful" and "well made" were used frequently in feedback.

Future evaluation will require more refinement to the interface and the underlying data model to fully support all of the interface goals and secure the system from manipulation. Badge and rank functionality need further development along with the administrative tools for publisher configuration and final verification of article entities.

Ideally, evaluation would be performed by the target worker group for a participating publication through the recruitment that publication's subscribers. Workers would be asked to interact with the interface to annotate a finite set of articles and respond to a qualitative survey to assess

their overall experience and willingness to use the interface again. Time-on-task would be measured by querying the data store for beginning and ending insertion time for the collection of worker annotations for a completed article.

Subsequent evaluation steps would be to release the interface to production and use log analysis and analytics to model worker engagement in a real-world environment. Metrics such as points accrued, progress at time of abandonment and level of reward redemption activity would be calculated and analyzed to provide a picture of how workers respond to the annotation experience.

6. CONCLUSIONS AND FUTURE WORK

It's difficult to present any informed conclusions without a proper evaluation of the interface. Discussion of the demo has been promising in the sense that the value and intentions of the protoype seem to be clearly understood during demonstrations and the incentive model appears to be stable.

Further functionality could be developed to make the tool more comprehensive:

6a. Media Assets

Most of the test documents have accompanying media assets (photos, videos, etc.) which could benefit from annotation. It would be relatively easy to display those assets with tagging or paraphrasing fields that would provide missing semantic / accessibility metadata. Including the media assets would also improve the worker experience by making the article complete.

6b. Advanced Interactions

We've implemented very basic phrase highlighting in the prototype when users mouses over an entity. It would be beneficial to make this function more consistent and also to allow users to highlight text to initiate the addition of entities. User-generated entities would ideally have a function for disambiguation that would allow them to select from an authority list specifying what specific entity that addition represents.

6c. Smart Viewports

The existing interface has inherent usability limitations reconciling what is simultaneously viewable in the article and the entity tool panel. A most basic revision would allow these panes to be independently scrollable to allow the user to move between panels without losing their place in reading. It is also possible for these panels to intelligently scroll in response to each other but we would need to experiment with making this behavior intuitive and not overly-complex.

REFERENCES

1. Chen, D.L., Dolan, W.B. (2011). Collecting Highly Parallel Data for Paraphrase Extraction. *University of Texas Department of Computer Science* [Not Yet Published]
2. Chen, J.J., Menezes, N.J., Bradley, A.D (2011). Opportunities for Crowdsourcing Research on Amazon Mechanical Turk. *To appear at CHI 2011 Workshop on Crowdsourcing and Human Computation*.
3. Cooper, S., Khatib, F., Treuille, A., et al. (2010). Predicting protein structures with a multiplayer online game. *Nature* 466, 756-760
4. N. Eagle (2009). txteagle: Mobile Crowdsourcing. In *Internationalization, Design and Global Development*, 2009.
5. Eckert, K., Niepert, M., Niemann, C., Buckner, C., Allen, C., and Stuckenschmidt, H. (2010). Crowdsourcing the Assembly of Concept Hierarchies. In Proceedings of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL), Brisbane, Australia. ACM Press.
6. Ipeirotis, P.G., Provost, F. & Wang, J. (2010). Quality management on Amazon Mechanical Turk. HCOMP'10.
7. Ipeirotis, P.G. (2011). Crowdsourcing goes professional: The rise of the verticals. *http://behind-the-enemy-lines.blogspot.com/2011/03/crowdsourcing-goes-professional-rise-of.html*
8. Ipeirotis, P.G. (2010). Mechanical Turk: Now with 40.92% spam. *http://behind-the-enemy-lines.blogspot.com/2010/12/mechanical-turk-now-with-4092-spam.html*
9. P. Hsueh, P. Melville, V. Sindhwami (2009). Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. NAACL HLT Workshop on Active Learning and NLP, 2009.
10. Paolacci, G., Chandler, J., and Ipeirotis, P (2010). Running experiments on Amazon Mechanical Turk. Judgment and Decision Making, 5(5), 2010.
11. Raddick, M.J., Bracey, G., et al. (2010). Galaxy zoo: Exploring the motivations of citizen science volunteers. Astronomy Education Review, 9(1), 2010.
12. Ross, J., Irani, I., Silberman, M. Six, Zaldivar, A., and Tomlinson, B. (2010). Who are the Crowdworkers? Shifting Demographics in Amazon Mechanical Turk. CHI EA 2010. (2863-2872)
13. Siorpaes, K. and M. Hepp (2007), "OntoGame: Towards Overcoming the Incentive Bottleneck in Ontology Building", *International IFIP Workshop On Semantic Web & Web Semantics (SWWS '07)*, OTM conferences, Springer LNCS, Vilamoura, Portugal, 2007
14. Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. Interacting with Computers, 13, 127–145.