Improving the Design of Intelligent Acquisition Interfaces for Collecting World Knowledge from Web Contributors

Timothy Chklovski
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292 USA
timc@isi.edu

Yolanda Gil
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292 USA
gil@isi.edu

ABSTRACT
An emerging approach to knowledge acquisition is to collect statements from volunteer contributors over the Web. In this approach, the design of the acquisition interface is key to focusing on statements of interest, avoiding spurious entries, retaining the contributors, etc. Several such volunteer-contribution-based systems have been deployed to date, each with its own idiosyncratic interface. This paper discusses some key challenges faced by volunteer collection interfaces, and outlines the design features that we have found effective in addressing some aspects of those challenges. The paper discusses how these features have been implemented in deployed collection systems, and reflects on the data collected to extract lessons for future work in this research area.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning – *knowledge acquisition*; I.2.4 Knowledge Representation Formalisms and Methods: *frames and scripts, semantic networks*

General Terms: Algorithms, Design

Keywords: Knowledge acquisition, intelligent user interfaces, interfaces for knowledge elicitation, broad-coverage knowledge repositories, collecting knowledge from volunteers

INTRODUCTION
Knowledge collection from volunteer contributors ([10], [16]) has recently emerged as an alternative to traditional knowledge engineering (e.g., [18]) and to text extraction from large corpora (e.g. [25], [11], [26]). Although some applications that leverage the knowledge collected have already been developed (e.g., [20], [8]), many challenges must be addressed to make this approach of practical use. Developing this source of broad-coverage knowledge would help address brittleness in knowledge systems and enable a new generation of AI applications.

Learner [5], [6] is a system that collects knowledge about everyday objects and events from volunteers. We continue to extend and improve Learner based on empirical analyses of the data collected [7]. Continued evolution of its design in our work suggest that intelligent acquisition interfaces for Web volunteers present their own distinct challenges and that some kind of guidance about what is effective in such collection efforts would be extremely helpful.

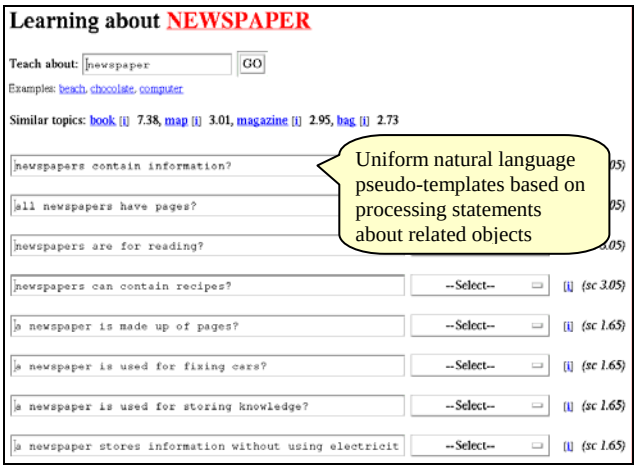
This paper outlines some *design features* that we have found to be effective in assessing interfaces by deploying them and analyzing the collected data. We present five key design features:

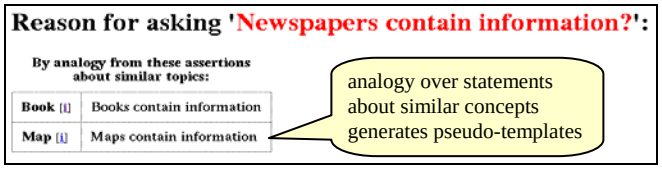
1. Create and fine tune *templates* to acquire specific types of semantic relations
2. Provide *guidance and feedback* on the form and type of the answer sought
3. Acquire knowledge *incrementally*, breaking up collection of complex statements into several acquisition steps
4. Automatically *postprocess* the knowledge to repair or discard entries.
5. Direct multiple contributors to *validate and evaluate previously entered statements*.

The paper describes in detail how these features were embodied in our implementations, and through either data or examples points out the resulting improvements in the collected knowledge.

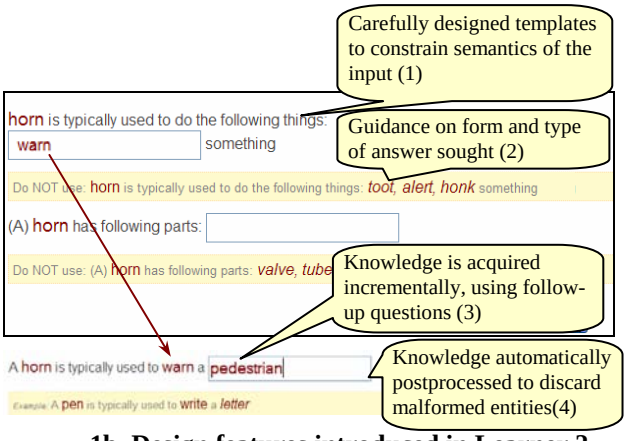
The paper also motivates these design features with three challenges presented by the design of intelligent acquisition interfaces from volunteers: 1) collecting interpretable knowledge despite ambiguity in natural language contributions, 2) collecting piecemeal contributions to describe a highly complex and interrelated world, and 3) detecting and handling spurious and non-consensus contributions. A vast body of prior research in cognitive

K-CAP'05, October 2–5, 2005, Banff, Alberta, Canada.

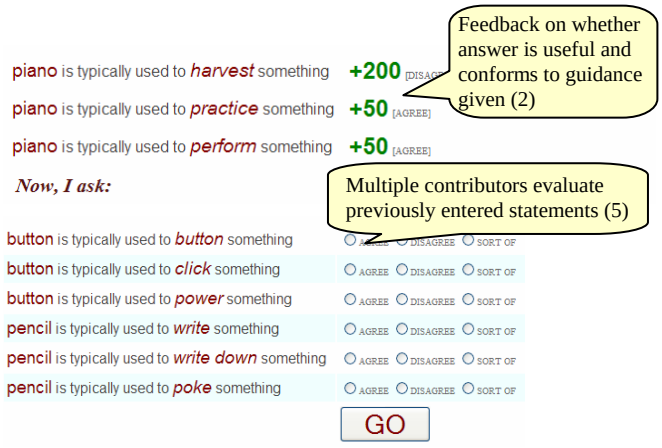




1a. Pseudo-template approach used in Learner



1b. Design features introduced in Learner 2



1c.) Design features introduced in Learner 2.5

Figure 1. The evolution of Learner's interaction with the user, illustrating the five design features discussed throughout this paper

science, user interface design, natural language, human factors, and knowledge capture is relevant and should be brought to bear as the community continues to deploy these collection efforts. This paper takes a first step in addressing some important aspects of these and other challenges.

DEPLOYED SYSTEMS & IDENTIFIED CHALLENGES

Over the past four years, we have been collecting knowledge from volunteers. We have fielded two major systems for collecting world knowledge (Learner [5],[6], and Learner 2 [7]). We have also deployed two systems for collecting lexical knowledge: word senses in contexts in Open Mind Word Expert, (OMWE) [22],[9], and focused paraphrase knowledge in *1001 Paraphrases* [8]. To date, these efforts have collected more than 600,000 entries from thousands of contributors[1]. These experiences have provided a reasonable amount of empirical data on the challenges that arise in collecting world knowledge from volunteers. The data has been contributed both via the Web and via a computer kiosk at a science museum exhibit. The focus of this paper is on the lessons that we learned from this experience.

In collecting knowledge from volunteers, two major factors need to be taken into account:

- *Homogeneity gap*: The world about which knowledge is being acquired is nuanced and heterogeneous, while the knowledge which we aim to acquire is more easily usable if it is homogeneous and semantically interpretable.
- *Comprehension gap*: Our system, when compared to a human, has extremely limited world knowledge, extremely limited ability to process natural language, and extremely limited reasoning and learning abilities. Thus, the way knowledge can be imparted to such a system needs take into account these severe limitations on the side of the learner. In addition, volunteer contributors are unlikely to have had experience or training in teaching a system that is not even remotely as good at learning as a child.

We have been evolving Learner in several stages: Learner, Learner2, which introduced significant changes, and Learner2.5 which extended Learner2 with new features. Snapshots of the interface used in these versions of Learner are shown in Figure 1 with highlights of the features that are discussed throughout the paper, and summarized at the end. We have also analyzed knowledge collected by an earlier project, Open Mind Common Sense (OMCS), and used that knowledge to seed the acquisition in Learner.

[1] The live collection system and collected statements are available from http://learner.isi.edu/

Through these experiences, we have had to deal with practical challenges which arise in such knowledge acquisition. Specifically, the *homogeneity gap* and the *comprehension gap* combine to give rise to the three challenges that we discuss in the remainder of the section. After describing the challenges, we introduce the approaches we have been investigating and our observations on where they succeed and fall short.

Challenge A: Collect *semantically interpretable* knowledge while interacting in natural language, which can be highly ambiguous

To capture a wide variety of world knowledge from contributors who are not versed in advanced knowledge representation formalisms, it is appealing to base the interaction on natural language. Natural language is both flexible and ubiquitous. However, natural language expressions are also notoriously ambiguous in a number of ways, including: i) underspecification of the semantic relations between statement elements (for example, English noun-noun phrases such as "brick house", "paper tray", or "coffee cup" omit the semantic relation altogether), ii) structural ambiguity of the statement (as in the attachment of "flying" to "I" or to "mountain" in "I saw the mountain flying over New York City"), iii) referential ambiguity (e.g. in "punching a wall causes pain" the experiencer of the pain is the agent doing the punching), and iv) word sense ambiguity (e.g. in "a hospital can have a part called a wing," *wing* is not the type used for flying). Interpreting such ambiguities with human-level precision without large amounts of lexical and world knowledge presents a significant challenge in the state-of-the-art natural language research. Some useful techniques to manage ambiguity include controlled grammars [29] and limiting range of input [2].

An additional feature of natural language is the non-uniqueness of ways to express (paraphrase) the same statement. This can hinder structuring of and generalization across the collected knowledge, as well as recognizing what is already known and what needs to be learned.

Given these features of natural language, the challenge for collecting knowledge basing on natural language becomes how to draw on its expressivity and ubiquity while sidestepping the challenges of ambiguity and non-uniqueness.

Challenge B. Knowledge about the world may be *difficult to fully and correctly specify with a single interaction*

The everyday world descriptions of which we seek is highly nuanced and heterogeneous. Statements about it need both careful delineation and qualification. At the same time, when contributors are asked to explain something, they tend to underestimate the richness of the knowledge they are imparting. Consider, for example, specifying typical usage of a *car horn*. The typical usage can be specified as an action (to *warn*), which can range over a variety of objects (e.g., *pedestrians, drivers, animals, children*, and so on). Furthermore, the objects being warned need to be in proximity of the vehicle, are typically in potential danger from the vehicle that the horn is a part of, need to be able to hear the horn for the horn to have an effect, and so on.

Challenge C: Contributors occasionally provide input which is *spurious, non-consensus, or malformed*

This challenge has to do with some contributors providing input which should not be used as is and needs to be either repaired or discarded altogether. The main types of such input are: *spurious input* (nonsensical statements such as "chicken is part of a knife"), *non-consensus* input such as "a dial is part of a television" or "arrow is part of a bow", and *malformed input* which includes typos and usage of plural when filling in a template which calls for a singular "a table has a piece or a part called a(n) *legs*" (sic) as well as spelling errors.

We briefly present some data on the extent of these problems.[2] The system was fielded at a kiosk in the Science Museum of Minnesota[3] for three months and has collected a set of 42,446 statements. Manual evaluation of 1000 statements suggests that around 5%-9% of the input is spurious. An automatic analysis showed that 5.2% had entries in plural where a singular was expected (e.g. *robots, dolls*). 0.85% entries contained arguments with a leading article *a*, *an*, or *the*, which had to be discarded to align with WordNet entires. Approximately 5% of the entries contained misspellings (e.g., *footbal, missle, tounge, antena*). Contributors occasionally entered either a person's first name (presumably that of a friend) or a "taboo" word (a curse or a slur). Preliminary comparison with rates in data collected over the web (rather than in a kiosk at a Science museum) suggests that the rates for these types of malformed input when collecting on the web has so far been slightly lower though still significant.

DESIGNING INTELLIGENT ACQUISITION INTERFACES FOR COLLECTING WORLD KNOWLEDGE FROM VOLUNTEER CONTRIBUTORS

This section presents five approaches which we have been investigating to cope with the above challenges. While not the final word on how to address these challenges, we believe our experience with these approaches will help future work in this research area.

---

[2] Additional discussion of acceptability of the knowledge collected and the coverage achieved can be found in [7].

[3] The collection occurred during the first stop (in St Paul, MN) of a 3.5 year traveling exhibit titled "Robots And Us". The exhibit will be featured at Ft. Worth, TX, Portland, OR, Boston, MA, Chicago, IL, and other locations.

Design feature 1: Create and fine tune *templates* to acquire specific types of semantic relations

In LEARNER, our first effort at designing an interface for collecting knowledge from volunteers, we aimed to rely heavily on natural language and to carry out disambiguation dialogues where necessary. To simplify processing the collected knowledge, LEARNER used "pseudo-templates" – new statements (hypotheses) were generated by replacing terms in previously collected statements to make plausible new statements. For example, "maps contain information" would be used to generate the hypothesis "newspapers contain information." While allowing the collection process to be quite expressive (new statements in parsable natural language could be added at any time), the collected knowledge turned out to be highly ambiguous, and designing all the needed clarification interactions would be no small task. Furthermore, this approach was suited for collecting a broad range of statements, but made it more difficult to focus the collection on specific statement types.

In Learner2 we used *templates* to acquire knowledge, with contributors filling in template blanks rather than entering full statements. This has been the approach we adopted for our ongoing work. Template-based collection is also used in OMCS and OMICS.

In this section, we discuss advantages and limitations of using templates to acquire knowledge, motivate guidelines we have formulated and adhered to in designing templates, and finally discuss our observations on the quality of the knowledge collected using templates.

Using templates to acquire knowledge

While templates are still phrased in natural language, they can be more precise than language in its common usage. This reason alone goes a long way to justify collection using templates.

Using templates also allows us to focus acquisition on acquiring specific types of knowledge. In Learner2, we use templates to specify a type of question, and then instantiate these templates on specific slot fillers. For example, a template for collecting part-of relations can read:

"a <*object1*> has a piece or a part called a(n) <object2>".

Instantiating the template to acquire parts of a *car* would produce the following knowledge acquisition question:

"a *car* has a piece or a part called a(n) _____".

This approach also allows us to not only focus on the type of knowledge we are acquiring, but also on specific objects about which additional learning is necessary. Collecting uniform knowledge and being able to guide collection may also support more extensive automated analysis of the knowledge, for example generalizing over the collected statements it to make further acquisition more intelligent.

The rigidity of a template-based interaction also has the disadvantage of preventing a contributor from providing knowledge which the contributor may think relevant and important but which does not fit the template. In OMCS, such freedom was recaptured at the cost of interpretability by allowing free-form input in addition to template-based interaction. Because we gave been targeting the collection of specific types of semantic relations, we have not found a strong need for collecting knowledge in this less interpretable but more flexible form.

Designing templates

In designing our systems, we have studied the knowledge collected by OMCS, an earlier system. OMCS used loosely phrased templates such as "*a <action> is for* _____". The collected statements may be difficult to interpret semantically, because remarkably many interpretations for conventional expressions can and do crop up. For example, the above template has collected, without distinction, assertions in which *is for* stood for several semantic relations, including: *results in a (emotional) state*, as in "riding a horse *is for* pleasure," *is done by*, as in "eating breakfast in bed *is for* sick people," and *has the aim of*, as in "getting a job *is for* using your skills."

These early observations and our later experience with designing more precise templates suggest that it is desirable to provide a lot of guidance about the kind of answer desired. If the collection task allows it, it is desirable for template blanks to solicit non-compositional concepts. This ensures that structural and referential ambiguities of natural language will not creep in into the answers supplied by contributors when they fill in the template. Finally, the template itself also needs to be carefully designed to avoid structural, referential, and word sense ambiguities.

In our experience, satisfying all of the above desiderata wile anticipating the types of entries contributors may make is often an iterative process of trial and error. In some cases, a helpful methodology may be to inspect, in a large text corpus, the surface manifestations of the desired semantic relation, as well as identify other semantic relations which may have the same surface forms.

Given the design and testing effort involved in deploying the templates, it is intriguing to contemplate whether very broad acquisition can eventually be made more autonomous by (reliably) delegating to volunteer contributors the tasks of proposing, critiquing and refining knowledge acquisition templates.

Experiences with using carefully designed templates

We have deployed templates to collect different types of knowledge, including semantic relations such as *part-of* and *typical-use-of*, as well as more contextualized knowledge such as what an administrative assistant may need to do to prepare a piece of equipment for use in a videoconference or a meeting. We have also used them to collect knowledge about problems which can arise when taking a certain action, and what can be done to address these problems. Finally, we have collected arguments

which are used in analysis of an issue, and the key aspects of such arguments.

In our experience, templates go a long way in addressing issues of ambiguity and non-uniform surface expression of the same relation in the collected knowledge, especially if the templates can be designed to collect only one, non-compositional answer at a time.

The only ambiguity that templates can offer little to help with is word sense ambiguity of the contributor input. Addressing that problem involves its own set of challenges. We have studied some of these in separate work on a volunteer contributor based system for collecting information about word senses, called Open Mind Word Expert (OMWE) [9],[22]. In future work, we aim to integrate into Learner both automatic and volunteer-based methods for decreasing word sense ambiguities.

A lingering issue with even the carefully crafted templates is that contributors take a loose view of what is admissible. For example, they specify that a "seam" is part of a "baseball," an "end" is part of a "beam," and even "notebook" is part of a "waiter". We speculate that additional guidance to contributors on what are acceptable or desired answers can help focus the contributions, as treated in the next design feature.

Design feature 2: Provide guidance and feedback on the form and type of the answer sought

In designing Learner 2, we noticed that while the knowledge acquisition templates can be phrased to provide some guidance about the type of answer sought, what can go into the blanks is left to the contributors' interpretation.

In exploring how much additional guidance to provide, we chose to provide examples of how a template may be filled in whenever we present the template. For instance, when collecting statements about why a certain action may be difficult to accomplish we used a template such as "something can be difficult to *move* because it is _____". We aimed to collect single adjectives or simple adjectival phrases for a variety of actions such as *move*. Below the template, we presented a sample way it may be filled in: "something can be difficult to *burn* because it is *wet*"; this allowed us to communicate the spirit of the type of answer we wanted without resorting to technical jargon such as "adjectival phrase." In Learner 2.5, we extended the example mechanism to show the previously provided answers to this instantiation of the template, when those are available.

Our experience suggests that the initial simple approach of providing examples of clearly acceptable answers may not constitute sufficient guidance. The experience of deploying the system brought to our attention that unless explicitly guided, contributors to our system are unlikely to know how narrowly to interpret the knowledge acquisition questions they encounter. That is, is it appropriate to provide answers which are marginally acceptable or only sometimes true? For example, to a question "a *car* has a piece or a part called a(n) _____", it is not clear whether it is appropriate to give the answer "*airbag*" (which only some cars have) or the answer "*piston*" (which cars with internal combustion engines have, but indirectly, as part of an engine).

In future work, we plan to extend the mechanism of guidance to provide not only prototypical, but also negative and extreme examples of what is and is not considered appropriate input. Such examples, as well as a brief description of the type of answer sought, would be useful to associate with the template.

Another technique to guide contributors is by providing feedback about whether what they just contributed is in line with what we sought to collect. In Learner 2.5, we added a mechanism for providing such feedback. For every answer provided, feedback is given in the form of a score added to or subtracted from the cumulative score maintained for the contribution session. Each score also comes with a brief explanation for the reason for it.

When generating feedback on the contributed answers, it can be difficult to distinguish a previously unseen good entry from a spurious one. While we were testing various schemes, one approach we tried was to award a significant number of points for a previously unseen answer. However, many contributors discovered that this scheme can be exploited with spurious answers. One contributor has captured the sentiment well when in a template, in place of providing an answer as instructed, the contributor typed "I can enter anything I want and you will keep giving me points". The new scoring scheme we use gives the most points for entries which match those which we already collected previously, but have not collected from sufficiently many contributors to be highly confident about the validity of the statement. This guides contributors to provide answers that others may also provide, while discouraging focusing solely only the most salient or obvious answers. Such a mechanism is motivated by the misallocation of contributor effort when spontaneous contributions are not managed [7].

In future work, we plan to explore automatic assessment of plausibility of an answer (e.g., using clustering techniques to see if an answer is an outlier). We also plant to award or subtract points for previously unseen answers if they when they are later validated or rejected by multiple other contributors, or successfully used in reasoning steps. In addition to creating an incentive to enter the most useful knowledge, this may also encourage contributors to check back on how their contributions are faring.

Design feature 3: Acquire knowledge *incrementally*, breaking up collection of complex statements into several acquisition steps

One of the challenges discussed earlier is that a piece of knowledge about the world may be difficult to fully and correctly delineate with a single entry. The approach we have been investigating is to acquire knowledge incrementally, using a cascade of one or more follow-up questions to acquire additional detail on any given entry.

The cascaded acquisition, introduced in Learner 2, takes advantage of the template-based approach by basing follow-up questions on slots of the statement being followed up. For example after collecting statements such as "A *microphone* may be useful while *setting up a videoconference,*" our system posed the following follow-up question: "As an admin assistant, if helping with *setting up a videoconference*, if you need to deal with a(n) *microphone*, important activities may be: _____ it". This follow-up question collected such answers as: *turn on*, *test*, and *adjust*. Our follow-up mechanism can also pose questions by using knowledge from several previously collected statements at once, for example to pose comparison questions.

Applicability of this design feature depends on presence of identifiable "intermediate mileposts" in the statements being collected to provide stages of acquisition. In the above example, the intermediate milepost is the piece of equipment which needs to be identified before the action needed to prepare it can be elicited.

An additional benefit of such "cascaded acquisition" is that it allows validation of the knowledge being collected, allowing us to address one entry at a time. Validation is discussed further below.

Design feature 4: Automatically *postprocess* the knowledge to repair or discard entries

One approach to addressing spurious, non-consensus, malformed input includes automatic evaluation and repair of the input. This approach applies particularly well to identifying malformed input such as spelling and morphology errors. For example, some collection templates require a singular noun. Resources such as a spellchecker and a large lexical database such as WordNet, can be used to identify questionable entries. In Learner2, we have deployed automatic detection and repair of entries as a postprocessing step. Future work may integrate such validation into the acquisition loop. The kiosk installation of Learner2 also uses a stoplist of "taboo words" (swear words) in its interactive knowledge acquisition to automatically detect and suppress any such entries.

We analyzed a set of 42,446 statements collected via the kiosk at the science museum exhibit. Useful canonicalization included aligning with WordNet word pairs which were not in entries in WordNet, but could be mapped to a single word WordNet concept (e.g. *lap top*, *tea cup* were automatically mapped to *laptop* and *teacup*). In all, 464 (1.1%) entries were so mapped. Examining the 464 repairs indicated that all the repairs made were correct. Other repair techniques we found useful include discarding spurious articles and correcting wrong number. For example, given the fill-in-the-blank template-based question "a *human being* has a piece or a part called a(n) _____, contributed knowledge included entries such as "*the brain*" or "*hands*". The first one was automatically repaired by discarding the leading article (which may not appropriate if collecting, for example, names of music bands). The second was repaired automatically converted to "hand".

However, correcting misspelling and otherwise normalizing knowledge fully automatically can be difficult due to a number of factors. Some words admit multiple spellings (e.g. theater/theatre, color/colour, judgment/judgement), while some misspellings result in rare but legitimately spelled words. For example, in several cases, "handle" was misspelled as "handel," which is found (sans capitalization) in WordNet's term bank, and refers to Handel, the composer. Because of these subtleties, to determine what is an appropriate repair in a specific statement, it may be useful to supplement automatic methods with consulting contributors to instruct the system how a given problem should be addressed. We are also interested in investigating more semantically-based postprocessing on a more recent, larger collection of knowledge (approximately 160,000 statements after one year of collection).

Design feature 5: Direct multiple contributors to *validate and evaluate previously entered statements*

Learner 2.5 introduced additional functionality to validate the collected knowledge by volunteers. Validation allows us to detect spurious statements which should be discarded as well as other statements which need to be further qualified or repaired. The validation mechanism is motivated by the presence of spurious statements and of statements needing qualification [7]. We discuss the design of the validation interface and present some data on deploying the approach. Note that knowledge collected from other sources, such as knowledge obtained by text extraction can also be validated using this approach.

Designing the validation interface

The validation interface asks volunteers to rate previously collected statements using one of several choices. The choices we currently use are "agree", "disagree" and "sort of." For each statement, ratings are solicited from multiple volunteers until the statement can be classified as by the follow up action appropriate to it: should be kept, discarded, or other (should be qualified/repaired). We use an ad hoc formula to classify a statement based on its validation ratings. For example, a statement is currently classified as "should be discarded" if there are at least four evaluations and at least three-quarters of the evaluators

disagreed with the statement. To prevent one contributor from unduly influencing the validation of any given item, only one vote per item is considered per IP address.

One aspect that has proved surprisingly challenging is designing the set of choices that the validators can use to evaluate a given statement. The choices need to be clear about what they mean, capture a variety of possible evaluations, but without overwhelming the contributor with a large number of subtly different choices. Constructing such a set of choices is ongoing work.

A major feature of the validation interface is to assess quality of the validations – both for quality control reasons and to provide calibrating feedback to the human validator. In addition to statements which need additional evaluations, we occasionally plant statements which have already been classified by consensus of earlier validators. When validating such a planted item, contributors receive a large "bonus" score for assigning the same classification as expected, and are docked for a dissenting classification. Because it is not revealed which items are "planted," validators need to pay attention to all items to avoid the negative feedback.

Experiences with deploying validation

As expected, validation can proceed faster than entry of knowledge. In collecting validations of statements about parts of objects and bout their typical uses, we have found that in the current interface, it takes contributors approximately 10 seconds per validation. By comparison, when entering new answers, contributors proceed at the average rate of one in 20 seconds.

In few weeks of collecting validation information, we have collected 16,027 ratings from volunteers. Despite the need for further exploration of the set of choices to present to volunteers, and the need for a more principled way to combine individual ratings, the current validation ratings appear promising. A total of 340 statements were rejected. Manually examining the rejections suggests that they are appropriate in more than 95% of the cases, although the classification formula may need to be adjusted to reclassify the borderline cases. That is, rejection has few false positives. The most significant misclassifications seem to center around statements which are borderline (such as "a dial is a part of a television set") being classified as acceptable. In future work, we intend to evaluate performance of validation more extensively and investigate whether additional instructions and feedback to contributors about how to treat such cases can further improve correctness of the classifications.

SUMMARY

This paper provides rationale for desirable features in designing collection interfaces aimed at volunteer contributors, based on a set of challenges identified from deployed systems and data analysis:

Challenge A: Collect *semantically interpretable* knowledge while interacting in natural language despite it being highly ambiguous.

Design feature 1: create and fine tune templates to acquire specific types of semantic relations; we point out the need to *carefully design and pilot test* the templates

Design feature 2: provide guidance and feedback on the form and type of the answer sought

Challenge B: Knowledge about the world may be *difficult to fully and correctly specify with a single interaction*

Design feature 3: acquire knowledge incrementally, breaking up collection of complex statements into several acquisition steps

Challenge C: Contributors occasionally provide input which is spurious, non-consensus, or malformed

Design feature 4: automatically postprocess the knowledge to repair or discard entries, and

Design feature 5: direct multiple contributors to validate and evaluate previously entered statements.

Figure 1 highlights these features in the context of the Learner interface.

RELATED WORK

In prior work, we analyzed the knowledge collected in terms of acceptability, coverage, and complexity [7]. There is a direct dependency between the acceptability and coverage of the knowledge collected and the design of the collection interface. These dependencies need to be better understood. Related work on mining web sites with free form contributions such as ratings and opinions looks at credit assignment and effort allocation for volunteer contributions [24] and theoretical results on amount of validation required under different noise levels [17].

A wide variety of ontology editors [12], [15], [28] are being developed to acquire knowledge expressed in semantic markup languages with clear semantics such as the OWL W3C standard. The knowledge collected is in the form of ontology classes, relations, and constraints. Users need to have some training in ontology engineering, and must fit their contributions into what is possible to express in the target language. In contrast, our contributors can express a wider variety of knowledge and they do not need any training or background to be able to contribute from the start.

Other related work on knowledge acquisition aims to capture knowledge from subject matter experts [4], [3]. The techniques used in these systems for knowledge validation and acquisition dialogues cannot be directly applied to web collection from volunteers because they rely on inference and constraint reasoning exercised over the knowledge entered. However, if the knowledge we collect

were processed to support shallow inference and possibly logical reasoning then those techniques could be exploited.

CONCLUSIONS

We have described several important features that we have found effective in designing acquisition interfaces to collect knowledge from volunteer contributors. We framed these features in terms of some broad challenges to this type of collection efforts. Through deployment and analysis of the resultant collected data, we continue to refine and articulate the design features that can be important to other researchers engaged in similar efforts. Some of the techniques that we use may also be useful to knowledge capture tools that interact with more trained users or subject matter experts.

In future work we would like to draw more strongly from principles and lessons learned in related areas such as user interface design, human factors, engaging interaction [13], collaborative dialogue systems [19], [1], controlled languages and other natural language processing techniques [29], common sense knowledge formalisms [21], and knowledge capture techniques [3], [4].

Designing increasingly competent interfaces for collecting knowledge from volunteers may be a viable and practical approach to create broad repositories of increasingly semantically interpretable declarative knowledge, enabling a new generation of intelligent applications.

ACKNOWLEDGMENTS

We gratefully acknowledge funding for this work by DARPA under contract no. NBCHD030010. We thank Joshua Seaver of the Science Museum of Minnesota for his work on deploying the kiosk, and Push Singh for making OMCS data available.

REFERENCES

[1] Allen, J., Byron, D, Dzikovska, M, Ferguson, G, Galescu, L, and Stent, A. "Towards Conversational Human-Computer Interaction," *AI Magazine 22(4)*, pages 27-38, Winter, 2001

[2] Androutsopoulos, I., Ritchie, G. D., and Thanisch, P. "Natural Language Interfaces to Databases: An Introduction". Natural Language Engineering, 1(1), 1995.

[3] Barker, K., Blythe, J., et al. "A knowledge acquisition tool for course of action analysis." *Proceedings of the Innovative Applications of Artificial Intelligence Conference (IAAI-2003)*. Acapulco, 43-50. 2003

[4] Blythe, J., Kim, J., Ramachandran, S., and Gil, Y. "An Integrated Environment for Knowledge Acquisition." Proceedings of the *2001 International Conference on Intelligent User Interfaces (IUI-2001)*, 2001

[5] Chklovski, T. "Using Analogy to Acquire Commonsense Knowledge from Human Contributors," PhD thesis. MIT Artificial Intelligence Laboratory technical report AITR-2003-002, 2003

[6] Chklovski, T. LEARNER: A System for Acquiring Commonsense Knowledge by Analogy. In *Proceedings of Second International Conference on Knowledge Capture (KCAP)*, 2003

[7] Chklovski, T. and Gil, Y. An Analysis of Knowledge Collected from Volunteer Contributors. To appear in *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*, 2005

[8] Chklovski, T. 2005. Collecting Paraphrase Corpora from Volunteer Contributors. In *Proceedings of International Conference on Knowledge Capture, K-CAP 2005*.

[9] Chklovski, T. and Mihalcea, R. Building a Sense Tagged Corpus with Open Mind Word Expert. In *Proceedings of the Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, ACL 2002

[10] Workshop on Distributed Collaborative Knowledge Capture (DC-KCAP 03). Held in conjunction with KCAP 03. http://www.isi.edu/~timc/dc-kcap/

[11] Etzioni, O., Cafarella, M., Downey, D., et al. 2004. Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison. In *Proc. of AAAI-2004*

[12] Gennari, J., Musen, M., Fergerson, R., Grosso, W., Crubezy, M., Eriksson, H., Noy, N., Tu, S. The Evolution of Protege: An Environment for Knowledge-Based Systems Development *International Journal of Human-Computer Studies*, 58(1), 2002

[13] Gottlieb, H. "The Jack Principles of the Interactive Conversation Interface". Jellivision Inc. 2002

[14] Gupta, R., and Kochenderfer, M. 2004. Common sense data acquisition for indoor mobile robots. In *Nineteenth National Conference on Artificial Intelligence (AAAI-04)*

[15] Handschuh, S., Staab, S. and Ciravegna, F., S-CREAM: Semi-automatic CREAtion of Metadata. *Proceedings of EKAW'02*. (2002)

[16] Symposium on Knowledge Collection from Volunteer Contributors (KCVC-05). AAAI 2005 Spring Symposium. http://teach-computers.org/kcvc05.html

[17] Lam, C. and Stork, D. Evaluating classifiers by means of test data with noisy labels, *IJCAI-2003*. pp. 513–518

[18] Lenat, D. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38 (11), 1995

[19] Lesh, N.; Marks, J.; Rich, C.; Sidner, C.L., "Man-Computer Symbiosis Revisited: Achieving Natural Communication and Collaboration with Computers", Transactions on Electronics, December 2004

[20] Lieberman, H., Liu, H., Singh, P., and Barry, B. 2004. Beating common sense into interactive applications. *AI Magazine*, Winter 2004, 25(4):63-76. AAAI Press.

[21] McIlraith, S., Peppas, P., and Thielscher, M. (Eds) Symposium Series on Logical Formalizations of Commonsense Reasoning, http://www.iccl.tu-dresden.de/announce/CommonSense-2005. 2005.

[22] Mihalcea, R. and Chklovski, T. Building Sense Tagged Corpora with Volunteer Contributions over the Web, book chapter in *Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing*, Nicolas Nicolov and Ruslan Mitkov (eds), John Benjamins Publishers, 2004

[23] Miller, G. WordNet: An On-line Lexical Database. In *International Journal of Lexicography*, Vol.3, No.4, 1990

[24] Richardson, M., Domingos, P. Building large knowledge bases by mass collaboration, in *Proceedings of Second International Conference on Knowledge Capture (K-CAP 2003)*.

[25] Riloff, E. and Jones, R. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proc. of AAAI-99*, pp. 474-479

[26] Schubert, L. 2002. Can we derive general world knowledge from texts? In *Proc. HLT 2002*, March 24-27, San Diego, CA, pp. 94-97

[27] Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., Zhu, W. Open Mind Common Sense: Knowledge acquisition from the general public. In Robert Meersman & Zahir Tari (Eds.), LNCS: Vol. 2519. *On the Move to Meaningful Internet Systems: DOA/CoopIS/ODBASE* (pp. 1223-1237). Springer-Verlag 2002

[28] Vargas-Vera, M., Motta, E., Domingue, J, Lanzoni, M., Stutt, A. and Ciravegna, F. MnM: Ontology Driven Semiautomatic and Automatic Support for Semantic Markup, *Proceedings of EKAW'02*. (2002).

[29] Wojcik, R. The Boeing Simplified English Checker, 2002, http://www.boeing.com/assocproducts/sechecker