Towards Managing Knowledge Collection from Volunteer Contributors

Timothy Chklovski & Yolanda Gil

University of Southern California
Information Sciences Institute (USC/ISI)
Marina del Rey, California, 90292
{timc, gil}@isi.edu

Abstract

A new generation of intelligent applications can be enabled by broad-coverage, up-to-date repositories of knowledge. One emerging approach to constructing such repositories is proactive knowledge collection from volunteer contributors. In this paper, we study the quality of the knowledge repository resulting from collecting spontaneous, little guided contributions of volunteers. In a representative collection of part-of information contributed by volunteers, we study the coverage and quality of the resulting collection. As a possible way to address the deficiencies, we outline a more managed, three-stage approach to the collection process, consisting of collection, evaluation & revision, and publication.

Introduction

Broad-coverage knowledge repositories stand to enable a new generation of intelligent applications and natural language understanding systems (Chklovski, 2003; Lenat, 1995). The variety of tasks and applications which can benefit from broad-coverage semantic resources are exemplified by uses of WordNet (Miller, 1990), a broad-coverage semantic resource which emphasizes lexical semantics. WordNet bibliography (Mihalcea, 2004) illustrates hundreds of uses in research.

One approach to constructing broad-coverage semantic (and lexical) resources is by employing a relatively small team of highly trained knowledge engineers. This approach has been taken by WordNet, CYC (Lenat, 1995), and DOLCE (Gangemi et al., 2003). This approach faces issues stemming from shortage of person-hours available, which can limit the coverage of facts and even limit which semantic relations are included (Lenat, 1995; Miller, 1990). This shortage can also lead to encoding viewpoints or statements that may require later reengineering or refinement (Gangemi et al., 2003; Friedland et al, 2004).

Another approach to constructing broad-coverage resources is text mining from large corpora and Web sources (Hearst, 1992; Berland and Charniak, 1999; Riloff & Jones 1999; Schubert, 2002; Girju, Badulescu, & Moldovan 2002; Etzioni et al., 2004). Through sophisticated statistical analysis and training algorithms, these approaches extract entities and discover useful lexical and semantic relations. While the level of precision and recall varies, the extraction of semantic relations remains a challenging topic.

An emerging approach that we are exploring is to collect knowledge from a multitude of minimally instructed volunteers. The approach can be traced back to at least 1857, when many volunteers aided the construction of the Oxford English Dictionary by mailing in knowledge about earliest known word usages. The recent advent of the Web has greatly simplified distributed contribution of knowledge, attracting a growing amount of research, including *Open Mind Common Sense* (*OMCS*), (Singh et al. 2002), LEARNER (Chklovski 2003a, 2003b), LEARNER2 (Chklovski, 2005), the *Fact Entry Tool* (*FET*) by the CYC team (Belasco et al, 2002), *Open Mind Word Expert* (*OMWE*) (Mihalcea & Chklovski 2004), and *Open Mind Indoor Common Sense* (*OMICS*) which adapts OMCS to collect knowledge about indoor objects (Gupta and Kochenderfer, 2004). The issues of reasoning over evidence of varying quality collected in mass collaboration settings are also being looked at (Lam and Stork 2003) and (Richardson and Domingos 2003). A key benefit of the mass collaboration approach is its inherent ability to bring orders of magnitude more effort to the construction process, since the approach can tap volunteers with minimal or no training. These volunteers also can be prompted with extensively conditioned questions, answers to which may be challenging to automatically extract from bare text. Also, because different contributors may have different backgrounds and contexts, the collection gathered from them is likely to include statements which are rare but true. Practical uses of broad-coverage knowledge collections collected from volunteers are also being developed and include detecting sentiment in emails, identification of potentially relevant images by reasoning about their annotations and indoor vision and robot navigation (Lieberman et al, 2004; Gupta and Kochenderfer, 2004).

This paper presents an analysis of the statements collected with one such system in terms of its coverage and acceptability. This analysis was done on a representative corpus, specifically statements about parts of everyday objects collected by the LEARNER2 system (Chklovski, 2005). Our analysis shows that if statements are spontaneously contributed, achieving broad coverage is unlikely since coverage grows ineffectively over time and over number of contributors. Our analysis also shows that some of the collected statements should be discarded, and that such statements can be detected when several humans agree on discarding a statement. We also show that there are a variety of sources of disagreement regarding the acceptability of specific statements, and that in those cases further qualification would be useful. This suggests that there is a role for volunteer contributors in evaluating and qualifying knowledge contributed by others.

To address these issues, this paper proposes extensions to current approaches to managing volunteer contributors, including: 1) a *tight feedback loop* to increase *coverage* during collection; 2) an *evaluation & revision step* in which contributors state agreement or disagreement with statements collected from others and add qualifications to statements which elicited disagreement; 3) a separation between an iterative cycle of evaluation and revision of the collected knowledge and a subsequent *publication* of knowledge that selects relevant subsets of the collection based on application or usage requirements.

The next section introduces LEARNER2 and the knowledge collection studied. After that, we motivate our analysis with examples of deficiencies in coverage and quality from the collection analyzed. Next, we analyze and characterize the nature of the challenges in broadening the coverage and improving the quality of the collected knowledge. We close with recommendations about how the collection process can be modified and extended to address these challenges.

Knowledge Collected by LEARNER2

LEARNER2 (Chklovski, 2005) has been deployed for six months as an interactive kiosk at a science museum as part of a traveling exhibit called "Robots and Us[1]," which will continue for 3 more years. LEARNER2 has collected more than 100,000 raw entries from museum visitors of all ages, collecting *meronymy* (*part-of*), *typical purpose*, *similarity*, and other semantic relations about everyday objects. LEARNER2 uses a template-based, fill-in-the-blank approach. For example, to learn about parts of a "*car*," LEARNER2 partially instantiates a template to form a fill-in-the-blank knowledge acquisition question:

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

To exclude malformed entries, the collected knowledge has been automatically postprocessed, removing all entries not found in a large lexicon (which removed approximately 25% of the 100,000 raw entries). Spelling mistakes were also discarded to avoid introducing errors by automatically correcting them. The postprocessed knowledge is available as the Learner2-v1.1 dataset[2]. To simplify evaluation, we focus on the meronymy statements (there were a total of 24,747 such statements). LEARNER2 used a seed set of 326 objects (selected from WordNet's tree of "instrumentation or device"). Users were allowed to introduce other objects as well. The seed objects were semi-automatically selected to exclude very rare objects; the resulting set contains objects such as *axe*, *briefcase*, and *compass*. Since the collection focused on the seed set of 326 objects, we restrict our analysis to them. The resulting analyzed set contains total of 6,658 entries, specifying 2,088 distinct statements.

Phenomena Identified in the Collected Statements

In this section, we introduce the issues present in the data: the ineffective coverage, the presence in the collected knowledge of statements which would need to be identified and discarded, and the presence of statements which are neither clearly acceptable nor clearly discardable but may be one or the other upon further qualification.

Coverage

Systems that collect knowledge from volunteers typically collect what can be called "spontaneous" contributions, that is statements about whatever topic or object comes to mind. As a result, there can be high redundancy in typical items and also spotty coverage in more unusual ones. This was the case in the collection we analyze here. Some statements are entered dozens of times at the expense of other acceptable statements, which are never entered. To illustrate, the 5 most frequently contributed (0.24% of all distinct) statements attracted a total of 533 (8.0% of all collected) entries:

part-of(handle, hammer)	136	*part-of*(blade, knife)	99
part-of(wheel, car)	121	*part-of*(wing, airplane)	75
part-of(engine, car)	102		

At the same time, some useful statements such as *part-of*(radiator, car), *part-of*(crankshaft, car), and *part-of*(aileron, airplane) were never entered.

These observations raised the issue of whether to stop collecting redundant contributions and if so how many times should a statement be collected before the utility of additional identical contributions becomes negligible. Another important issue is whether and how to steer contributors to contribute new statements when the collection contains a sizeable amount of what could be considered the most common or typical statements. Below, we will show an analysis based on data from the LEARNER2 corpus regarding these issues.

Collected knowledge: the good, the bad, and the needing qualification

Another important set of phenomena that we observed in the LEARNER2 data is a wide variety of quality or acceptability of the knowledge. There are statements arising from contributors occasionally disregarding the collection instructions, such as *part-of*(chicken, knife) and

[1] http://www.smm.org/robots/

[2] The live system and the collected data are available at http://learner.isi.edu

part-of(truck, pot), that should clearly be discarded. We noticed that these are a very small portion of the collection.

Judging the quality of the collection in terms of its correctness or accuracy is a non-trivial task. This is case with many kinds of knowledge and is not specific to part-of relations. Whether a given statement is indeed a *part-of* statement involves a number of subtleties. For example, (Winston et al, 1987) have discussed the types of the part-of relation, such as component/integral object, member/group, place/area, and others while Miller (1990) highlights instances of non-transitivity of the relation. The issues we observed had more to do with how the notion of the part-of relation and the terms in the relation need to be qualified to determine whether a given statement is acceptable.

Given the lack of a formal or intensional definition of correct part-of relations, we decided not to treat correctness as an all-or-nothing matter but rather as something that can be increased by additional context to the statement. For example, *part-of*(film, camera) was entered by several contributors and is not clearly wrong. Yet, the statement does not hold for digital cameras, or newly purchased, not yet loaded cameras, and so on. What should be counted as an object and therefore as its parts is also not always clear cut. For example, acceptability of *part-of*(elevator shaft, elevator), and *part-of*(sail, mast) depends on whether the elevator refers to just the elevator cab or to the whole elevator structure, and whether the mast refers to the structure with the sail and the rigging or just the bare structure. Other statements drew disagreement because the part was not tangible, as in *part-of*(hole, tube), *part-of*(flame, torch). Word senses can also play a role. For example, *part-of*(row, table); *part-of*(mouse, computer) drew disagreement in evaluation scores. Although collecting explicit information on senses in which words are used would be useful, such collection involves an entire set of research issues (e.g., Mihalcea and Chklovski, 2004) which have not been engaged by LEARNER2.

Given that our ultimate goal is to collect common knowledge about everyday objects, we would prefer to keep all of these statements in some form within the collection. This is a very challenging issue, and one that we discuss below in more detail. It is worth noting that such statements are often not included in manually engineered and highly curated resources such as WordNet. In construction of knowledge repositories by knowledge engineers, the knowledge encoded is typically prescriptive. That is, if a statement is often, but not necessarily true, it would likely not be included. For example, WordNet specifies that a dog is a mammal, but does not provide any indication that dogs are (often) pets. By contrast, the statements we collect tend to include statements which are only sometimes true, such as *part-of*(remote control, stereo) and *part-of*(rope, pulley). Harnessing the ability to collect such statements and perhaps qualifications of the context in which they hold may be a potential strength of the approach of collecting from volunteers.

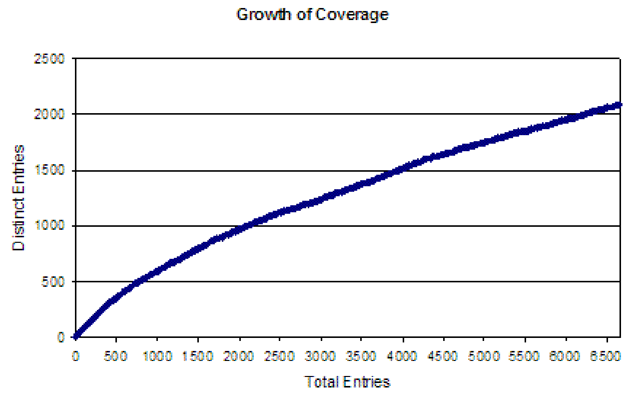

Figure 1. Contributions of distinct statements over time.

Detailed Analysis of the Collected Statements

In this section, we analyze in detail how contributor statements are distributed and the impact of this distribution on coverage. We also suggest possible indicators of acceptability of knowledge and analyze their merits based on the data collected.

Coverage

Out of a total of 6,658 entries collected, only 2,088 are distinct; 68.6% of entries were spent on getting redundant knowledge, adding nothing to coverage. Furthermore, examining all entries contributed three or more times reveals that 4,416 entries (66.3% of all entries) yielded only 350 distinct entries (16.8% of all distinct entries). This suggests that contributor effort was inefficiently exploited and could be redirected from these areas to other areas that have poorer coverage.

Furthermore, as shown in Figure 1, as the collection grows, the ratio of distinct to all statements contributed so far keeps decreasing. The diminishing returns seem to come from two sources. The first source is simple saturation of distinct answers. As the more frequent answers are collected the new ones become increasingly rare. The second source stems from the variability in the number of acceptable answers to a question. For example, even though in the collection studied all parts of a *hammer* and an *axe* have probably been collected, many parts of a *watch* have not yet been. Yet, the system currently keeps querying about objects without any preference for those about which knowledge is less complete. Hence, coverage suffers from contributor effort not being directed both at the question and at the answer level.

Towards classifying knowledge by acceptability

Given the considerations discussed above on how to judge quality and acceptability and lack of a working definition, we turned to evaluation by majority vote of human judges, a methodology previously selected by Berland and Charniak (1999) and Girju (2003) to evaluate automatic text extraction techniques. While an imperfect indicator of acceptability, as has been pointed out by Berland and Charniak (1999), majority vote provides a practical way to assess it. In our analysis, we asked 3 subjects (judges) to rate collected statements on a scale ranging from 0 to 3 ("is not," "probably is not," "probably is," and "is" a part-of relation). Statements were presented in a randomized order.

We consider two potential indicators of the acceptability of statements: redundancy and generation frequency.

We first examine whether the number of times a statement has been entered (its *redundancy*) is indicative of the opinion of the judges. To that end, we sampled 250 items from several redundancy categories: 1, 2, 3, and "4 or more".[1] The results from the "4 or more" category have additionally been broken out into 4 and "5 or more" when the data was analyzed. These categories were not shown to the judges.

Table 2. Redundant contributions and majority vote

# times statement was contributed (redundancy)	# distinct statements in this category	% for which majority voted "is" or "is probably" part-of
1 or more (all statements)	**2,088**	**70.5%**
2 or more	735	89.8%
3 or more	469	93.8%
4 or more	350	95.9%
5 or more	271	97.5%
exactly 1	1,353	60.0%
exactly 2	266	82.8%
exactly 3	119	87.4%
exactly 4	79	90.6%

Table 2 presents the results. The number of times statements were contributed is shown as well as the proportion of statements rated as "is" or "is probably" part-of by the majority of the three judges. The bottom of the table shows the number of statements contributed a given number of times. When our sampling is weighted by true number of statements in each sampled subset of statements, 70.5% of all statements receive the majority vote of judges. Of statements contributed more than once, majority vote is received by 89.8%. The majority vote increases monotonically with the number of times a statement has been contributed, with 97.5% of statements with contributed five or more times receiving the majority vote. In our evaluation sample, all 52 statements with contribution frequency of 15 or more were accepted (although in 3 cases one judge dissented). All three judges accepted all 35 evaluated statements that were entered 23 times or more (the maximum times a statement was entered is 136).

[1] There were only 119 items with redundancy of 3, so only 119 samples were used from that category.

The entries that we suggested earlier as ones that should clearly be discarded, such as *part-of*(chicken, knife) and *part-of*(truck, pot), primarily have redundancy of 1. Manual review indicates that there were approximately 25 such statements in the set evaluated by judges, no such statement was contributed more than three times. Also, all such statements received the lowest evaluation score from at least two of three judges, giving promise to future work on their identification.

A second potential indicator that we used in our analysis is *generation frequency*. It is based on the notion that more common or typical statements that are spontaneously brought up by many users are more likely to be acceptable. We define the generation frequency (*gf*) of a statement about a given part and an object as the frequency with which this part has been contributed out of a total number of times a statement has been made about any part of this object. For example, *part-of*(handle, hammer) was contributed 136 times out of a total of 203 statements about parts of a hammer. This yields the generation frequency of *part-of*(handle, hammer) to be 136/203=0.67. We expected answers with higher generation frequencies to be more accurate.

Table 3 shows the results. We show separately the results for statements contributed once, twice and so on into two sets: those with *gf* below 0.1 and those with *gf* of at least 0.1 (splitting the evaluation data into two sets of roughly equal size). Surprisingly, for redundancy greater than 1, items with lower generation frequency tend to be more accurate than items with the higher generation frequency. This finding suggests that collecting low-frequency items may not negatively impact the quality of the collection.

Table 3. Generation frequency and majority vote

# times statement has been contributed	**Gen freq < 0.1**		**Gen freq >=0.1**	
	% receiving majority vote "is" or "is probably"	Num in sample	% receiving majority vote "is" or "is probably"	Num in sample
exactly 1	55.7%	158	67.4%	92
exactly 2	89.8%	127	75.6%	123
exactly 3	92.3%	52	83.6%	67
exactly 4	92.0%	25	89.3%	28
5 or more	100.0%	44	96.7%	153

To sum up, the number of times that a statement has been contributed is a strong indicator of majority vote and therefore acceptability to judges. However, high generation frequency, for statements contributed more than once, is not. The mixed assessment of statements contributed once suggests that more information is needed about the acceptability of these statements. The positive assessment of the statements contributed many times suggests that they require relatively little further assessment effort.

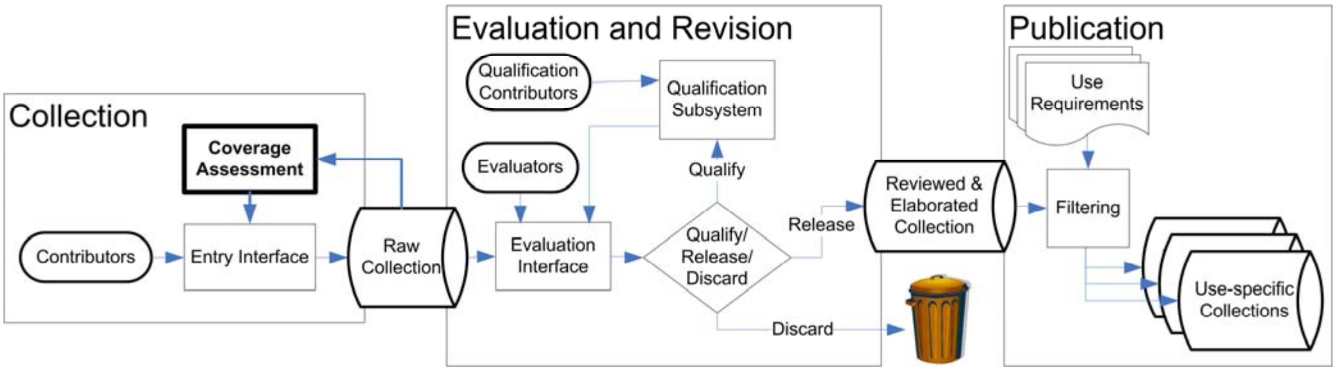

Figure 2. Collection Process

Analysis of Comparable Resources

An area that requires further work is the detailed comparisons of the content of our collections versus resources created through other approaches such as ontology engineering and text extraction. There are many subtleties involved in such comparisons. Here, we present some initial results which indicate that the approaches may be complementary and amenable to combination.

Extracting the part-of relation from text has been attempted by Berland and Charniak (1999), reporting 55% accuracy and citing issues such as lack of unequivocal syntactic indicators for the part-of relation in text. Girju et al., (2003), resorted to ontological knowledge and a large amount of manually annotated training data to improve extraction precision, reporting precision of 83% on an extracted set of 119 statements. For statements contributed 2 or more times, our accuracy is 89.8%, which surpasses the results from text extraction. Still, automatic extraction from very large corpora (e.g., Hearst, 1992; Etzioni et al, 2004; Riloff & Jones 1999; Schubert, 2002) may uncover valuable statements to augment or complement volunteer collections.

WordNet also contains part-of relations, although an appropriate comparison is difficult to formulate because, while WordNet's coverage is not complete, some statements in it are extremely general and some are extremely specific. For example one of the senses of a "pen" in WordNet is a "female swan," which, as a "whole object" has a part "part" and, as a "bird" has parts such as "oyster" (a small muscle of a bird). Comparing with direct parts of only primary senses of the concepts studied, we find overlap of statements LEARNER2 collected with WordNet to be 10-15%.

Towards Managing Volunteer Collection Efforts

We have presented an analysis on a system which collects knowledge from volunteer contributors in a simple fashion. Contributors interact with the entry interface to provide what we have referred to as "spontaneous" contributions. The analysis of the knowledge so collected suggests that the coverage and quality of the resultant collection may benefit from extending the collection process in several significant ways to manage volunteer efforts. Figure 2 diagrams the proposed conceptual stages of a more managed collection process. The three major stages we identify are *Collection*, *Evaluation & Revision*, and *Publication*. The *Collection* stage, the only one clearly present in current systems (OMCS, OMICS, LEARNER, LEARNER2), collects raw statements. Examining coverage achieved by spontaneous collection indicates ineffective allocation of effort in this stage. As a remedy, we propose a feedback loop which assesses achieved coverage and guides contributors to extend it. Before proceeding with the description of the other stages, we briefly discuss four specific methods to increase coverage. *(a) Guide contributors away from known answers*, showing a "taboo list" made up of the top most frequent answers. *(b) Collect knowledge about insufficiently covered objects*, using some saturation criterion to guide contributors towards objects about which new answers continue to be contributed (for example, a hammer has fewer parts than a car, and the collection process should reflect this. Together with these feedback-based methods, it may be useful to allow each contributor to *(c) enter several answers per question*. The final method we mention is to *(d) prompt contributors with possible answers*. This method uses collected knowledge to suggest other possible answers or to generate similar answers. LEARNER used a form of this method by analogizing possible answers from statements about similar objects. A synergy with text extraction approaches is also possible: text fragments extracted from text corpora (e.g. the Web or an encyclopedia) using already collected knowledge may also be used to prompt contributors or bootstrap extraction.

During the proposed *Evaluation & Revision* stage, acceptability of statements is evaluated. As a result of this evaluation, statements are to be released, discarded, or qualified (revised), as appropriate. The statements directed for qualification are to be re-evaluated in this stage after being qualified. The statements released after the evaluation form what we call the Reviewed and Elaborated Collection, which contains statements passing some minimum quality requirements. The statements may also be annotated with the evaluation information. To gauge prospects for evaluation of knowledge, we computed the inter-annotator agreement of the judges who rated the knowledge we used. Overall inter-annotator agreement of the judges who received little instruction was 76.6%, while agreement on answers with redundancy 4 or more was 85.1%. Providing explicit guidelines on how to treat the specific sources of subtlety is likely to result in further increase in agreement. These observations suggest reasonableness of further investigate integrating into the collection process of a stage of evaluation of acceptability by additional volunteers. If evaluation is to be carried out by volunteer contributors, note that quality of the evaluations by a given volunteer can be spot-checked by including known, gold standard items for validation, using these items to "evaluate the evaluators." The refinement can take a number of forms, rooted in deep knowledge representation issues. Some forms of refinement which the collected knowledge could benefit from include specifying tangibility of the part (e.g. "idea" in *part-of*(idea, textbook) could be reined to *intangible-part-of*(idea, textbook)), specifying senses of the terms used, and specifying whether the relation holds only for some instances of the object (as in *part-of*(remote control, radio), *part-of*(airbag, car)).

Finally, the Publication stage consists of filtering, which takes in *use requirements* and processes this collection to create a use-specific collection. We distinguish this stage because use or application requirements may vary, affecting what knowledge should be released in a given use-specific collection. For example, an intelligent user interface application may require the list of the most typical parts, most agreed upon parts, while an application for reference resolution is better served by a more inclusive, even if less reliable, list of parts.

Conclusions

This paper examined the issue quality of knowledge contributed by volunteers. We analyzed a representative collection of part-of information contributed by volunteers with little guidance, identifying certain systematic deficiencies in coverage and quality of the resulting collection. As a possible remedy to the deficiencies, we proposed a more managed three-stage approach to the collection process consisting of collection, evaluation & revision, and publication.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under Contract No.NBCHD030010.

References

Belasco, A., Curtis, J., Kahlert, R., Klein, C., Mayans, C., Reagan, P. 2002. Representing Knowledge Gaps Effectively. In *Practical Aspects of Knowledge Management*, (*PAKM*), Vienna, Austria, December 2-3.

Berland, M. and Charniak, E.. 1999. Finding Parts in Very Large Corpora. In Proceedings of the the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99).

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

Chklovski, T. 2003b. LEARNER: A System for Acquiring Commonsense Knowledge by Analogy. In *Proceedings of Second International Conference on Knowledge Capture (K-CAP 2003)*.

Chklovski, T. and Pantel, P. 2004. Path Analysis for Refining Verb Relations. In *Proceedings of KDD Workshop on Link Analysis and Group Detection (LinkKDD-04).* Seattle, WA.

Chklovski, T. 2005. Designing Interfaces for Guided Collection of Knowledge about Everyday Objects from Volunteers. In *Proceedings of Conference on Intelligent User Interfaces (IUI05)* San Diego, CA

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*.

Friedland, N., Allen, P., Matthews, G., Witbrock, M. et al. 2004. Project Halo: Towards a Digital Aristotle. *AI Magazine*, 25(4): Winter 2004, 29-48

Girju, R., Badulescu, A., and Moldovan, D. 2003. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. Proc. of the *Human Language Technology Conference (HLT)*, Edmonton, Canada.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. 2003. Sweetening WORDNET with DOLCE. *AI Magazine* 24(3): 13-24.

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

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*.

Lam, C. and Stork, D. 2003. Evaluating classifiers by means of test data with noisy labels. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico, pp. 513–518

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

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.

Mihalcea, R. and Chklovski, T. 2004. 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.

Mihalcea, R. 2004. WordNet bibliography, available online at http://engr.smu.edu/~rada/wn

Miller, G. 1990. Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography*, 3(4), 245-264

Richardson, M., Domingos, P. 2003. Building large knowledge bases by mass collaboration. In *International Conference on Knowledge Capture (K-CAP03)*

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

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

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

Winston, M. E., Chaffin, R. and Herrmann, D. 1987. A taxonomy of part-whole relations. *Cognitive Science*. 11:417–444