Common Consensus: a web-based game for collecting commonsense goals

Henry Lieberman
MIT Media Lab
20 Ames St. E15-384A
Cambridge, MA 02139, USA
+1-617-253-0315
lieber@media.mit.edu

Dustin A Smith
MIT Media Lab
20 Ames St. E15-320F
Cambridge, MA 02139, USA
+1-617-253-0315
dustin@media.mit.edu

Alea Teeters
MIT Media Lab
20 Ames St. E15-X
Cambridge, MA 02139, USA
+1-617-253-0315
alea@media.mit.edu

ABSTRACT

In our research on Commonsense reasoning, we have found that an especially important kind of knowledge is knowledge about human *goals*. Especially when applying Commonsense reasoning to interface agents, we need to recognize goals from user actions (plan recognition), and generate sequences of actions that implement goals (planning). We also often need to answer more general questions about the situations in which goals occur, such as when and where a particular goal might be likely, or how long it is likely to take to achieve.

In past work on Commonsense knowledge acquisition, users have been directly asked for such information. Recently, however, another approach has emerged—to entice users into playing games where supplying the knowledge is the means to scoring well in the game, thus motivating the players. This approach has been pioneered by Luis von Ahn and his colleagues, who refer to it as *Human Computation*.

Common Consensus is a fun, self-sustaining web-based game, that both collects and validates Commonsense knowledge about everyday goals. It is based on the structure of the TV game show *Family Feud*[1]. A small user study showed that users find the game fun, knowledge quality is very good, and the rate of knowledge collection is rapid.

ACM Classification: H.3.3 [INFORMATION STORAGE AND RETRIEVAL]: Information Search and Retrieval; I.2.6 [ARTIFICIAL INTELLIGENCE]: Learning

Keywords: knowledge acquisition, commonsense, games, goals, contexts, commonsense knowledge retrieval

INTRODUCTION

Common Consensus is an on-line game, designed to motivate users to contribute Commonsense knowledge about people's everyday goals to a large knowledge base. A knowledge base of goals and associated information will serve as a resource for intelligent interfaces to model the motivations and actions of their users. We seek to make the game entertaining enough to motivate players to contribute, while making sure to get knowledge that answers the questions that interest us and that represents consensus knowledge according to our users.

Collecting Commonsense Knowledge from Volunteers

A major disparity between computers and humans is that computers do not have the vast resource of everyday knowledge that we humans rely on to solve-problems and communicate. Information such as "people sleep at night" and "doors can be opened" is trivial and implicit to people, but is absent from computer software. The problem of acquiring this enormous body of knowledge is known to the artificial intelligence community as the *knowledge acquisition bottleneck*. Offering an innovative solution to this problem, the Open Mind [19] and CYC [9] projects demonstrated that the Internet can be used for distributed knowledge collection, particularly commonsense knowledge, which, by definition, is non-expert and possessed by everyone. Subsequently, many similar projects have been developed [3] [8] that collect knowledge from volunteers and store them in various formats. The OpenMind project, for example, maintains the knowledge in basic English statements.

Motivating Volunteers to Contribute

If we expect to continue collecting knowledge from volunteers, we must focus on ways to motivate them to contribute high-quality knowledge. Although we have collected a lot of knowledge from projects like OpenMind, we are far from the hundred-of-millions to billions of "pieces of knowledge" that are estimated to be involved with human intelligence [15]. This challenge is exacerbated by the fact that the number of volunteer contributors drops over the life of the project.

In 2004, von Ahn and colleagues started building web-based games which serve the dual purposes of acquiring knowledge and providing entertainment to their users. Notable such efforts include the ESP Game for annotating images; Peek-a-boom, a game designed for segmenting objects in images; and Verbosity, a game for collecting commonsense knowledge [20], [22] and [21].

[1]Family Feud is a trademark of FremantleMediaOperations BV.

IUI'07, January 28–31, 2007, Hawaii, USA..

Goal-Oriented Commonsense Knowledge

Having a large amount of knowledge is just one part of the commonsense reasoning problem: we also need good ways to retrieve, represent and reason with this knowledge [14]. Particularly important is knowledge about human *goals*. Starting from Maslow's Hierarchy of Human Needs [12], people have desires that motivate their behavior. These goals are broken down into subgoals, and the goal tree terminates in concrete actions. Goals often answer the *why* questions about human behavior, and provide good clues as to the *when*, *how*, and other considerations. They are therefore fundamental to explanation.

One of our long-term goals is to index Commonsense knowledge via goals and knowledge associated with goals. To our knowledge, this has not yet been done in a comprehensive way for large-scale, everday Commonsense knowledge. Many applications of Commonsense knowledge rely on relatively simple matching techniques. The simplest of these fall back on Information Extraction technologies such as keyword matching and statistics of word co-occurrence such as Latent Semantic Analysis to match up goals with statements of their methods and results. More complex structural techniques perform limited kinds of reasoning over semantic networks, such as the spreading activation reasoning of ConceptNet [11], Case-Based Reasoning, and structure-mapping analogy. Plan recognizers also have a limited capability to recognize when a sequence of actions is consistent with the desire to accomplish a certain goal.

But in order to do plan recognition, generation, monitoring, and debugging over a wide spectrum of everyday situations, it is desirable to collect enough explicit knowledge about goals to reduce the burden of inferring every detail about goal-oriented behavior from first principles. People are often (but not always!) quite articulate about why they are doing something, when asked, even though they leave out this knowledge as already understood in normal discourse. Thus an application that explicitly asks users for goal-oriented explanations can quickly amass a large collection of goal knowledge.

Design Objectives for the Game

We developed a game, Common Consensus, that has the following characteristics: 1) provides entertainment to the users and thus motivation to contribute; 2) defines the quality of an answer by the number of consenting answers; and 3) avoids convergence by replenishing seed questions with common answers. In this section these properties are explained in more detail:

Motivating volunteers to contribute knowledge Our research group's prior knowledge acquisition projects [19] have the problem of attrition: contributions from volunteers tend to diminish quickly over time. With Common Consensus, we sought ways to motivate more users to contribute and continue to contribute commonsense knowledge. Although commonsense knowledge providers may be in short supply, Internet gaming is tremendously popular and people spend many hours playing web-based games. Recently, researchers have attempted to "tap" this human resource by developing knowledge collection games that are attractive to Internet gamers. To motivate users to contribute, we developed a competitive multi-player game where users race against each other to contribute commonsense knowledge. Users have found the scoring and user interaction elements of the game enjoyable, and a small user study suggested that most subjects wanted to continue playing.

Automatic knowledge validation by consensus Commonsense knowledge bases developed by volunteers must anticipate noisy data, which comes in the form of misspellings, incorrect information, and, most commonly, knowledge at varying levels of detail. This granularity problem leads to knowledge that is difficult to reason with reliably. The game structure of Common Consensus inherently provides a data-validation mechanism: the scores for players' answers are computed by counting the number of other people who contributed the same answer. The more people agree on a specific answer, the more confidence we have that this answer is valid.

Four subjects evaluated the data obtained during the user testing and all of the answers that one third of all users had entered were consistently marked as excellent answers for the question. The consensus mechanism can serve as a way to screen data. For example, when users were presented with the question: *What are some things you would use to: cook dinner?* their aggregate answers gravitated toward the superordinate and basic categories [13]. The most common answers (by the number of unique users) were: food (7), pots (3), pans (3), meat (3), knife (2), oven (2), microwave (2)... We also collected specific and atypical answers, like *garlic press* and obscure answers, like *cans* but they had a low count (in this case, 1). It should be noted that there is a trade-off involved with only using the popular answers: many good uncommon answers are neglected.

Goals as questions and answers: a continuous supply of questions We are collecting first-person goals, which we define as a verb and at least one object (*e.g.*, "to write an email"). Goals can be represented in a hierarchy, where each goal may have parent and children goals. A plan is a specific sequence of sub-goals, and each parent goal can have many particular sequences of sub-goals (plans). In other words, there may be many ways to satisfy a goal (with the email example: "open Gmail", "use your Blackberry" are both valid sub-goals), and each of those ways can be expressed as a series of sub-goals.

The hierarchical nature of goals allows us to ask questions that retrieve parent or children goals. When many users suggest a particular goal as an answer (and it has a specific syntactic signature), that answer is recycled into a new seed question.

If the question is too high level, low level (*i.e.*, actions[2]) or is malformed, we have provided a way for users to 'flag' bad questions by clicking on a button during the questioning round. The goal is then moved to a table where we can review and remove it if necessary.

[2] In this framework, an action is just a sensorimotor goal.

GAME ARCHITECTURE

The game model is similar to the television game show *FamilyFeud* where contestants are asked to answer a trivial, open-ended question and are rewarded based on the commonality of their answers. On the game show, the scores would be determined by pre-screening the audience; however, our game computes the score dynamically based on the answers for the given round or, if there are too few players, the answers from prior games.

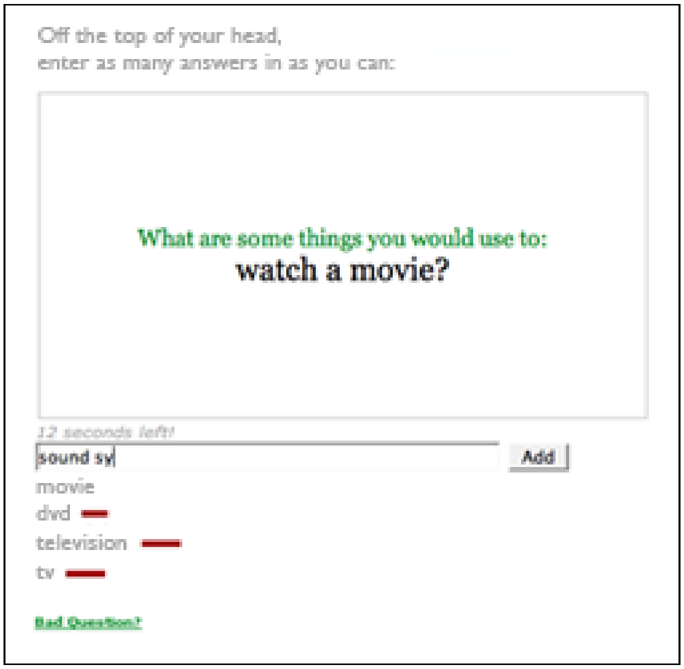

Figure 1: This is the answer box, where the user submits answers. The red bars grow as more users enter the same answer.

Here is an overview of a typical round of Common Consensus. First, a single goal is selected from a database of goals. The goal is embedded into one of six question templates:

1. *Why would you want to* X*?*
2. *What is something you can do if you wanted to* X*?*
3. *What is another goal similar to* X*?*
4. *About how long would it take to* X*?*[3]
5. *What are some things you would use to* X*?*
6. *Where are some places you would* X*?*

The first two questions are designed to accumulate parent and children goals for a given goal. Answers to the third question provide orthogonal connections between goals. Certain types of answers to these questions are recognized as candidates for use in future questions.

[3] This question is a special case: only one answer can be selected and the most common answer "wins."

In figure 1, users were asked to describe objects related to the goal "watch a movie." Users are encouraged to supply as many answers as possible. Only unique strings can be entered and users cannot enter the same answer twice. After each answer is entered, it is compared with the other answers to see if another user has entered the same answer. The player receives immediate feedback in the form of a expanding red bar that indicates how many other users have entered the same answer.

Scoring Answers: Identifying the Most Common Knowledge

After the timer has elapsed, each user receives a score (a product of the number of unique users who provided the same answer) and all of the answers and scores are displayed (see figure 2). When three players or fewer are playing, answers from prior game-rounds are used as stock answers.

Figure 2: Answers are grouped, scored and displayed after each round.

Determining Semantic Similarity in Answers We took a pre-emptive measure against user frustration by designing the game so that virtually identical answers would be scored as the same. To this end, we built a program that quickly determines whether two sentences are "the same" or not.[4] Similarity is a vague concept that is difficult for both man and machine to assess [4]. With the primary considerations being speed and performance predictability, we regarded two strings as the same if, after removing stop words and expanding contractions, all words are identical or share a common lineage in WordNet's inheritance hierarchies [7].

[4] http://web.media.mit.edu/∼dustin/Pair-A-Phrase-1.0.tar.gz

COLLECTING AND REPRESENTING GOALS

Having a lot of knowledge is only one part of the problem; we also need good ways to represent and reason with this knowledge. In this section we discuss the reasons we have chosen to collect goals.

Commonsense knowledge is context-specific

Commonsense knowledge is *defeasable*. Similar to the intended sense of an ambiguous word, the truth of a given assertion is dependent on the context in which it is used [16]. When solving everyday problems in a changing environment, it can be useful to represent objects in an environment in terms of their "affordances:" what they can offer toward solving the current problems, and what problems are they most useful for solving.

Why collect goals?

We are representing our goals in a hierarchical structure, where each goal can have parent goals, children goals and sideways "analogous goals". Other information, such as related objects, locations, and duration can be attached as properties of the goals.

Given the hierarchical structure of goals, goals are a useful way to organize knowledge at varying levels of detail. Additionally, they are related to solving problems in general, which is of course an important characteristic in Commonsense reasoning systems. Here are some examples of ways goals could be used in commonsense reasoning applications.

Identifying goals from context, and context from goals Plan recognition and plan generation are important components of Commonsense reasoning. Systems that plan or infer a user's goal would benefit from having knowledge about how goals are related and the characteristics of these goals. Identifying the plan *watch a movie* may vary greatly depending depending on surrounding context, for example "date" versus "airplane."

Roadie [10] is an intelligent user interface for consumer electronics based on interacting about the user's goals rather than the specific functions of the consumer electronics devices it controls. Roadie contains both a planner for mapping from goals to procedures and a plan recognizer for mapping from procedural steps to possible goals. The user can express a goal in natural language, for example, "I want to record a song". The system constructs a set of steps that can accomplish the goal, displays the steps to the user, and insofar as is possible, software controlling the devices can perform steps automatically. It generates context-dependent help for steps that need to be performed manually by the user.

It can also watch user actions such as pressing buttons on the devices, and infer a set of possible goals. Turning on the power to a DVD player indicates a possible goal of watching a movie or listening to a CD. Further actions by the user constrain the possible interpretations of the actions.

Application context provides a natural filter for relevance. For example, in the Common Consensus example above, the user is asked what you need to watch a movie. Answers like "a DVD player" or "a videocassette recorder" would generate possibilities for Roadie to interpret the "watch a movie" goal, but "popcorn" does not, because Roadie intersects its goal knowledge with its knowledge of consumer electronics.

Creo/Miro/Adeo [6] is a set of browser tools that allow users to record goal-oriented Web procedures using Programming by Example. They use Commonsense reasoning for the generalization step necessary to abstract away a procedure from particular examples. Miro recognizes the occurrence of possible goals on a Web page that can be satisfied by procedures recorded with Creo.

Constructing a model of common human goals We believe an extensive resource of everyday human goals would be a great asset to the artificial intelligence and cognitive science communities. There are a lot of interesting questions on the nature of goals that could be investigated with such a resource: *How different are people's goals?* We could look at how many users agree on the sub-goals they suggest. *How do people categorize and classify goals; in what aspects are two goals "similar"?* Looking at the ways users suggest analogous goals would help shed light on this question.

Goal knowledge is also critical for applications which model the attitudes and behaviors of a user or community. We can use these representations of many users' goals to build a caricature of a community's goals and problem solving strategies. Having a rich goal structure may prove a good way to model similarities and differences among cultures, and eventually, among individuals (for example, by comparing their goals against prototypical commonsense models). Having a rich user model would be useful in customizing plans (*i.e.*, planning around other goals like "to save time," "to save money") [1] and predicting how a user may react to a situation.

Reasoning with meta-knowledge about goals Knowledge about a goal can be useful even when the system has no understanding of a way to achieve that particular goal. For example, if a planning system needs to construct a new plan to solve a goal, it would be useful to know analogous goals so that plans to achieve existing goals can be adapted to suit the current situation.

EVALUATION

User Interactions: Was it fun?

An 11-person user study was conducted to survey whether users enjoyed playing the game and were asked to comment on their overall interaction in the context of an entertaining experience. The users all played each other for 20 to 30 minutes, which is 30 to 45 rounds of questions, and then were directed to complete a user interaction survey with optional comments and feedback.

From the feedback we collected, 10 of the 11 subjects enjoyed playing the game and 8 claimed they would play again. Most users identified the multi-user interaction was the most entertaining aspect of the game. One user suggested that the anxiety produced by the timer would contribute to more spelling mistakes. Overall, we were satisfied with the users' reactions and in the near future we plan to open the game to a larger audience.

Assessing the quality of the data

After the 11-person study was completed, we had collected 549 unique answers out of a total 800 answers. On average, each round contributed at least 18 answers, or 27 per minute.

To assess the quality of the data, we took a random sample of 40 answers and asked four other subjects to judge the answer in relation to the question. In our experience, agreement among the judges is typically high, so a large number of subjects is not needed for this phase of the evaluation. All of the answers that were entered individually by at least $\frac{1}{3}$ of the users were ranked as excellent answers by all four judges. Unfortunately, user answers only overlapped in a small percentage of the answers, and one third (11) of the answers that were contributed by only one user (33) were also considered excellent by all four judges. Here, a larger-scale user test would significantly reduce this problem, as lower-priority answers would garner more votes.

The random sample was not representative of the game (17.5% had multiple answers in the sample, versus 31.4% in the game), so we had subjects review the answers that selected to confirm or reject this one-third selection criteria. Consistent with the results from the random sampling, all 39 out of 549 answers where at least a third of user's agreed upon were ranked as excellent by the human reviewers. In the public release of the game, we plan to extend the length of the questioning round so that more overlapping answers are accumulated.

RELATED WORK

The most closely related work is Von Ahn et al.'s game for collecting Commonsense knowledge, Verbosity [21]. As mentioned above, we were inspired by their more general research program, Human Computation, which envisions using multiplayer games for knowledge acquisition. Verbosity uses the structure of the parlor game *Taboo* to encourage a pair of players to input general Commonsense facts. Common Consensus is more focused on the specific area of acquiring knowledge about goals, rather than facts in general. Indeed, one of the templates in Verbosity, "X is used for Y", does often elicit goals, and some of our templates could also be imported into Verbosity. One problem with Verbosity is that the prohibition of the taboo words, the central constraint of the game, also has the effect of discouraging the most straightforward way of expressing the knowledge, thereby introducing artifacts into the resulting knowledge. Like the ESP game [20], but unlike Verbosity, Common Consensus encourages the user to explicitly choose answers that match what they expect an anonymous person to say, rather than just anything that counts as a true answer to the question. It thus ensures that users are motivated to respect the "commonality" of knowledge.

The website 43Things.com also collects goals from users, and in turn it provides a way for users to find other users who have the same goals, even if they are uncommon. We have found most of 43Things.com's goals to be too abstract and high-level for the type of everyday knowledge we wish to collect. For example, this website has goals of the nature "spend more time with family," and "make a difference in the world" which contain rich information about human values, social and self knowledge, but are far from the foundational level of everyday activities we are currently investigating.

The Human Computation paradigm for collection of Commonsense knowledge was recently applied in a game called The FACTory, a Java applet on Cycorp's Web site[5]. It presents a randomly chosen fact from Cyc's database, e.g. "Earache is a symptom of conjunctivitis", and asks the user whether it is true, false, don't know, or doesn't make sense. Interestingly, the game can actually disagree with you, in the case that your answer differs significantly from the majority of other users or a high-confidence answer. The FACTory is intended as a single-(human)player game, with the other players only implicitly and asynchronously represented as the CYC consensus.

More generally, Common Consensus touches on the area of Knowledge Acquisition in AI [2]. Much of knowledge acquisition, as it has been traditionally performed in Expert Systems and Knowledge-Based Systems, is concerned with eliciting from informants problem-solving rationale as well as specific context-dependent actions. Understanding the goal of the informant is central for the knowledge engineer to figure out how to represent the informant's expertise in rule-based form, and how to generalize the informant's experience from a specific situation. However, knowledge acquisition work has traditionally been performed in highly specialized domains such as medicine and engineering, and little work has been done in acquiring general goal-based models of everyday activities. Recent work on Web mining of Commonsense knowledge [5] and [18] also holds out the promise that some knowledge of goals could be inferred indirectly from Web material intended for human readers.

There have been also several attempts to theoretically model the relationship between goals and actions, for example, Belief-Desire-Intention models [17] in multi-agent systems, and dialogue models in natural language processing. Again, our emphasis here is on collecting knowledge that could then be reformulated to serve as input to such models.

Commonsense knowledge bases such as Open Mind, Cyc, and ThoughtTreasure already contain considerable knowledge about everyday goals and activities as as a result of their general collection of Commonsense knowledge. OpenMind's original "activities" (templates that served as prompts for knowledge enterers) included several that, like the template "X is used for Y", tend to (but not always) encourage the user to express goals. It was our desire to expand upon these and harness the power of gaming that led to the development of Common Consensus.

CONCLUSION

1. Why would you want to *collect knowledge about goals*?
    - To create better Intelligent User Interfaces(3)
    - Improve our ability to do Commonsense reasoning (3)
    - Understand human thinking better (3)
    - Have fun (3)

[5] http://www.cyc.com

2. What is something you can do if you wanted to *collect knowledge about goals*?
   - Create a game called Common Consensus (3)
   - Give Common Consensus to some users to play (3)
   - Use Common Consensus knowledge to build better interfaces (3)
   - Write a paper about Common Consensus (3)
3. What are some places you would *collect knowledge about goals*?
   - MIT Media Lab (3)
   - IUI 2007 (3)

ACKNOWLEDGMENTS

We would like to thank Amelia Servi for her work on the answer comparison module, Pair-a-Phrase; Burak Arikan for contributing his design talents; Bo Morgan, Hugo Liu, Sean Markan, and Benjamin Mako Hill for their helpful suggestions.

REFERENCES

1. Chris Baker, Josh Tenenbaum, and Rebecca Saxe. Bayesian models of human action understanding. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 99–106. MIT Press, Cambridge, MA, 2006.

2. Jim Blythe, Jihie Kim, Surya Ramachandran, and Yolanda Gil. An integrated environment for knowledge acquisition. In *IUI '01: Proceedings of the 6th international conference on Intelligent user interfaces*, pages 13–20, New York, NY, USA, 2001. ACM Press.

3. Timothy Chklovski and Yolanda Gil. Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pages 35–42, New York, NY, USA, 2005. ACM Press.

4. Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. *COLING 2004*, 2004.

5. Ian Eslick. Searching for commonsense. Master's thesis, MIT Media Lab, 2006.

6. Alexander Faaborg and Henry Lieberman. A goal-oriented web browser. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 751–760, New York, NY, USA, 2006. ACM Press.

7. Christiane Fellbaum and George Miller, editors. *WordNet*. The MIT Press, 1998.

8. Rakesh Gupta and Mykel J. Kochenderfer. Common sense data acquisition for indoor mobile robots. In Deborah L. McGuinness and George Ferguson, editors, *AAAI*, pages 605–610. AAAI Press / The MIT Press, 2004.

9. Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

10. Henry Lieberman and José Espinosa. A goal-oriented interface to consumer electronics using planning and commonsense reasoning. In *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*, pages 226–233, New York, NY, USA, 2006. ACM Press.

11. Henry Lieberman, Hugo Liu, Push Singh, and Barbara Barry. Beating common sense into interactive applications. *AI Magazine*, 25:63–76, 2004.

12. Abraham Maslow. A theory of human motivation. *Psychological Review*, 50:370–396, 1943.

13. Carolyn B. Mervis and Elanor Rosch. Categorization of natural objects. *Annual Review of Psychology*, 32(1):89–115, 1981.

14. Marvin Minsky. *The Emotion Machine*. Simon and Schuster, 2006.

15. Erik Mueller. Common sense in humans. Available at http://www.signiform.com/erik/pubs/cshumans.htm, 2001.

16. Mircea Gh. Negoita, Robert J. Howlett, and Lakhmi C. Jain, editors. *Commonsense Reasoning in and Over Natural Language.*, volume 3215 of *Lecture Notes in Computer Science*. Springer, 2004.

17. A. S. Rao and M. P. Georgeff. BDI agents: From theory to practice. In *Proc. of 1st International Conference on Multi-Agent Systems (ICMAS)*, pages 313–319. AAAI Press/MIT Press, 1995.

18. Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. Mindnet: acquiring and structuring semantic information from text. In *Proceedings of the 17th international conference on Computational linguistics*, pages 1098–1102, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

19. Push Singh. The public acquisition of commonsense knowledge, 2001.

20. Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM Press.

21. Luis von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: a game for collecting common-sense facts. In *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems*, volume 1 of *Games*, pages 75–78, 2006.

22. Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: a game for locating objects in images. In *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems*, volume 1 of *Games*, pages 55–64, 2006.