Using Verbosity: Common Sense Data from Games with a Purpose

Robert Speer and **Catherine Havasi**
MIT Media Lab
Cambridge, MA

Harshit Surana
Carnegie Mellon University
Pittsburgh, PA

Abstract

Verbosity, a "game with a purpose", uses the collective activity of people playing an Internet word game as a body of common sense knowledge. One purpose of Verbosity has always been to provide large quantities of input for a common sense knowledge base, and we have now achieved this purpose by connecting it to the Open Mind Common Sense (OMCS) project. Verbosity now serves as a way to contribute to OMCS in addition to being an entertaining game in its own right. Here, we explain the process of filtering and adapting Verbosity's data for use in OMCS, showing that the results are of a quality comparable to OMCS's existing data, and discuss how this informs the future development of games for common sense.

Crowdsourcing for Common Sense

Common sense is the body of intuitive facts about the world that is shared by most people. When people communicate or interact, they assume others have this knowledge and tend to leave it implied. This makes the collection of common sense knowledge an ideal candidate for human computation; this knowledge is shared across most of the population, yet it is often difficult for a computer to acquire on its own, because so much of it is left unsaid.

Verbosity is a "game with a purpose" or GWAP (von Ahn and Dabbish 2008) for collecting common sense knowledge by Luis von Ahn, Mihir Kedia, and Manuel Blum (Luis von Ahn and Blum 2004). GWAPs are a form of crowdsourcing designed to take a difficult problem, such as protein folding or image labeling, and use online volunteers' game playing time to collect data which will help a computer become better at performing the task. The purpose of Verbosity has always been to contribute to a database of general knowledge, and now we have done so by incorporating it into Open Mind Common Sense.

Open Mind Common Sense, one of the largest common sense projects, has been collecting statements of common sense from volunteers on the Internet since early 2000 (Singh et al. 2002) (Havasi 2009), making it an early adopter of the idea of "crowdsourcing". OMCS's main focus has been to collect common sense about objects in the world and their properties, people's goals and emotions, and information about events. In the original OMCS site, users entered knowledge using one of several "activities" which asked users to enter data in certain forms, respond to prompts, or explain existing data. The current OMCS site allows users to enter new pieces of common sense knowledge, to rate existing knowledge, and to answer questions that verify knowledge inferred by AnalogySpace (Speer, Havasi, and Lieberman 2008). The interactive site has difficulty retaining users, however, because it is lacking in entertainment value.

Verbosity

Verbosity is a game about guessing a secret word, with gameplay comparable to other word games such as "20 Questions" or "Taboo". It is also designed in such a way that it collects common sense knowledge from its players.

Players in Verbosity switch between two roles: the describer and the guesser. As a describer, they have to convey the given secret using the fixed templates for common sense facts that Verbosity provides. When playing as a guesser, the player has to guess the secret word based on these clues from the describer. This process collects common sense facts from the describer, and verifies them when the guesser correctly guesses the word. The interface enforces some rules to minimize cheating and bad information, such as requiring that all words used by the describer must appear in a dictionary, and that they cannot include any form of the target word itself.

All user actions are stored in a database, which can be examined to extract common sense facts. A fact is stored in the database only if the guesser was able to answer the secret word correctly. More specifically, the database stores: the target word, the clue (including the template or "relation" that it filled), the frequency with which that clue was used for that target word, the average position within the clue order that the clue appeared, the conditional probability of that clue appearing, and other similar measures. For example, the target word *squirrel* is associated in the database with information like "it is a type of *tree rodent*" and "it looks like *chipmunk*", because describers used those descriptions in successfully cluing the word "squirrel".

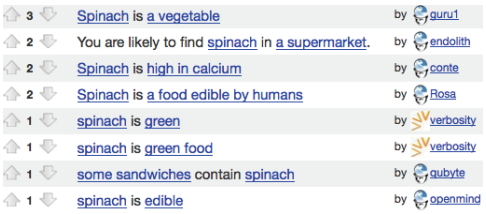

Figure 1: Knowledge about *spinach* is incorporated into the OMCS database.

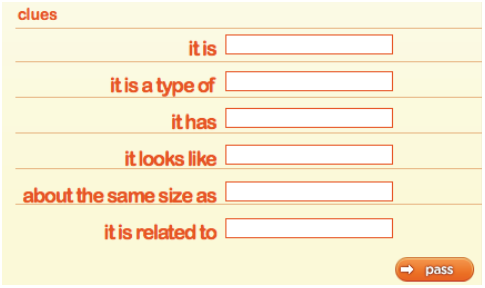

Figure 2: Verbosity's clue structure resembles OMCS's frame system.

ConceptNet

The statements collected by Verbosity take a form that makes them very useful to the Open Mind Common Sense project. In fact, they almost exactly match the data representation of ConceptNet, the semantic network representation of OMCS.

ConceptNet (Havasi, Speer, and Alonso 2007) is largely created from the free-text sentences that users have contributed to the OMCS database using syntactic patterns and shallow parsing. Much of the recent information in the OMCS database has been collected using a system of elicitation frames to ensure that their results will be easily integrated into ConceptNet (Speer 2006) without the need for parsing.

Data entered in these frames takes the form of (*concept*, *relation*, *concept*) triples. This conveniently matches Verbosity's clue structure, as we can see in Figure 2. To make the the Verbosity statement "It is a type of plant." fit into the ConceptNet framework, all one needs to do is replace the word "it" with the target concept.

Because of this, the OMCS project has begun to integrate knowledge from Verbosity into ConceptNet, filling a need for more entertaining ways to add knowledge to ConceptNet. An example of how information on *spinach* has been integrated can be seen in Figure 1. Using a filtering process that we will describe shortly, the OMCS project has imported over 200,000 statements of English-language common sense from Verbosity.

Other games for common sense

In an early questionnaire (Speer et al. 2009), top OMCS users expressed that the more interactive a common sense collection interface was, the more they would contribute to the project. Relatively early on, OMCS built games in an effort to get users to work on tasks that may be perceived otherwise as less fun or to populate data in sparse areas. The first of these games was the Game for Interactive OpenMind Improvement or GIOMI(Garcia, Havasi, and Todd 2002). GIOMI was created to draw upon the OMCS community in a new effort to rate assertions in the OpenMind database, an unpopular activity.

The game Common Consensus (Lieberman, Smith, and Teeters 2007) was created to quickly acquire more common sense information on a topic or topics. Common Consensus is based on the game *Family Feud*, a television game show where contestants compete to give the most common answers in response to prompts. Ideally, Common Consensus is played with a large group of "contestants." Additionally, OMCS has created its own "twenty questions" game (Speer et al. 2009). This game's purpose is not only to guess what concept the user is thinking of, but also to acquire specific pieces of knowledge which will be most useful in inferring additional information about the concept.

Kuo *et al.* (Kuo et al. 2009) created two community-focused games to collect knowledge in Chinese. These games collected over 500,000 verified statements which have become the OMCS Chinese database. These games, hosted on Facebook and PTT, take advantage of the community nature of the platforms to create goal-oriented games.

Another major common sense project, Cyc, acquires common sense information from knowledge engineers (Lenat 1995). Cyc uses their game, FACTory (Cycorp 2009), to have members of the public check facts which have been collected from knowledge entry, machine reading, and inference.

Verbosity in Practice

Verbosity has been played by over thirty thousand individual players. On an average, there are a few thousand games every day, and it remains the most popular game on gwap.com based on gameplay statistics.

In their paper on Verbosity(Luis von Ahn and Blum 2004), von Ahn, Kedia, and Blum say they intend on "creating a system that is appealing to a large audience of people, regardless of whether or not they are interested in contributing to Artificial Intelligence." With so many users providing so much knowledge, they succeeded at their goal. However, recruiting those additional users comes with some unanticipated challenges.

Existing word games that are similar to Verbosity, like Taboo and Charades, often encourage "outside-the-box" ways of giving clues, such as through rhyming and word play. Charades, in particular, comes with conventions such as indicating the number of words and syllables, which are not just allowed but are often expected. Tricks such as these do not fit the purpose of Verbosity, because they tend to produce statements that are not actually true, but players may be bringing their assumptions from other games to Verbosity. Because the player's primary goal in the game is to successfully communicate a word, and generating correct common

sense knowledge is only a vaguely related activity, many players may not even realize this is undesirable.

Sometimes players go to even greater lengths to clue the target word quickly, such as by spelling it out over several clues, or by saying words that are extremely similar to the target. This defeats the point of the game somewhat, but can allow the player to earn points very quickly. Simply put, people cheat at Verbosity.

In an ideal world, the clues that players entered into Verbosity would be directly usable as ConceptNet assertions. In reality, the process of importing the Verbosity data into ConceptNet requires a considerable amount of filtering, for the reasons we have seen above. At this point, we need to detect the patterns that indicate cheating, frustration, or "bending the rules", and remove or alter those assertions so that what remains is reasonable common sense knowledge.

Working with Verbosity data

Use and mention

Many erroneous statements in the Verbosity data come from the fact that players often disregard the use-mention distinction. Verbosity expects players to *use* words to convey meaning, but in some cases, players instead *mention* the words to refer to the words themselves. An example of the use-mention distinction is illustrated by the following sentences:

- A cow has four stomachs. *(Use)*
- "A cow" has four letters. *(Mention)*

Statements that mention words as words are not particularly useful in building a common sense knowledge base. Computers are, after all, already good at determining the properties of strings without regard to their meaning. Because people did not actually use quotation marks to distinguish use from mention, we need to detect these "mention" statements and discard them.

Flag words

When certain words appear in a clue, they are "red flags" that the describer is mentioning the target word instead of using it. By inspecting the database, we decided to filter out statements that contained any of the following words or their derived forms: *letter, rhyme, blank, word, syllable, spell, tense, prefix, suffix, start, end, plural, noun, verb, homonym, synonym, antonym*.

Other words indicate that the describer is referring to previous clues or the guesser's previous guesses, which also does not help in collecting common sense. For example, one player gave the clues "it looks like accord", "it has dance", and "it is add them" to clue the word "accordance". To deal with these cases, we also filtered out the words *guess, opposite, close, only, just, different, this, that, these,* and *those*, and clues that started with *add*, *delete*, or *remove*.

Along similar lines, we filtered out clues that ended with a single letter word or that contained fewer than three letters overall (because these were often involved in spelling the target word), and we filtered out the word *mince* because for some reason it was included in over 1300 nonsensical clues (most likely all from the same player). In all, these checks discarded 48354 out of 448339 statements.

Look-alike and sound-alike clues

One of the relations that Verbosity prompts for is "it looks like". This was intended for statements about physical objects, such as "a clock looks like a circle", but it may have accidentally encouraged players to violate the use-mention distinction by describing what the target word itself looks like. This results in statements that are not true at all, but which lead the guesser toward the correct answer anyway.

One common way that players describe what a word looks like is to mention another word, or pair of words, that look like the target word, such as "*farmer* looks like *farm err*". This would sometimes require some creativity, because Verbosity requires that all the words players use in their clues are in its dictionary. In some cases, players would give this kind of clue even when "it looks like" was not the relation they were being prompted for, producing statements that don't even make sense as statements about words, such as "*carpet* is a kind of *car pet*".

Related to look-alike clues are sound-alike clues, where players would give words that are pronounced like the target word, or – very frequently – that rhyme with the target word. In many cases, these overlap with look-alike clues, and players frequently (but not always) used the relation "it looks like" for sound-alike clues as well.

Some more examples of look-alike clues and sound-alike clues include:

attack is *a tack*
belief is a kind of *be leaf*
chord is typically in *rhymes sword*
heat looks like *feat meat*
machine looks like *mush sheen*
passion looks like *fashion*
wander is a type of *wonder*

To deal with the many instances of look-alike and sound-alike clues, whether they used the "it looks like" relation or not, we ran a script to detect them. As false negatives (letting erroneous statements through) are worse than false positives (keeping correct statements out of the database), this script aggressively matched and discarded anything that could be a look-alike or sound-alike clue.

To determine if two strings looked alike, we calculated the average of (a) the length of their longest common prefix, (b) the length of their longest common suffix, (c) the length of their longest common substring, and (d) the edit distance between the two strings, subtracted from the length of the shorter string. The *look-alike factor* can then be defined as this average divided by the length of the shorter string. By experimenting on test data, we decided that we should reject all statements where the look-alike factor between the target word and the clue was over 0.35.

Some phonetic similarities, of course, are not detected by the look-alike factor. The *sound-alike factor* of two strings could be calculated by replacing all words with their first entry in the CMU Phonetic Dictionary (`cmudict.0.7a`), and calculating the look-alike factor of the result.

Finally, because some players used multiple look-alike or sound-alike words in their clue, as in the example "*heat* looks like *feat meat*", we also calculated the average look-alike factor and sound-alike factor between the target word and each individual word in the clue.

So at the top level, we rejected a statement if any of these quantities were over 0.35:

- The look-alike factor between the target and the clue
- The sound-alike factor between the target and the clue
- The average look-alike factor between the target and all words of the clue
- The average sound-alike factor between the target and all words of the clue

We call this maximum the *text similarity factor*. The distribution of text similarity factors over all statements that passed the "flag words" step is shown in Figure 3. We discarded 47115 clues that had a text similarity factor over 0.35.

Filtering based on score

Another way of filtering the statements that resulted from Verbosity is to examine the statistics that were collected for each Verbosity statement, especially the number of times it was asserted and where it appeared in the sequence of clues.

By experimentation, we decided on the following rules for which statements to accept based on these statistics:

- Start with a score equal to the number of times the statement was asserted.
- If the statement always appears first in the clue order, increase its score by 1.
- If the relation is "it has" (which appeared frequently and was frequently misused), decrease its score by 1.
- If the resulting score is 2 or more, accept the statement.

The effect of this process is to mostly accept statements that are confirmed by more than one person entering them, but to also accept many one-shot statements that were entered first in the clue order. (These statements were of generally higher quality than other one-shot statements, because they reflect what comes to mind first for that concept, and because players seem to enter less appropriate clues later in the sequence when they are more frustrated.)

This is the final filter that determined whether a statement would be accepted or rejected. Figure 4 shows the breakdown of statements that were filtered out for various reasons and statements that were accepted.

Rephrasing

In some cases, we rephrased Verbosity clues so that they would be more useful and reliable as ConceptNet assertions.

A common type of clue was to express the target word in terms of what it is not. Sometimes players used the relation "it is the opposite of" for this, but sometimes they simply expressed this using whatever relation was handy, as in "*reserve* looks like *not active*", "*inch* is not *foot*", and "*plant* is related to *not animal*".

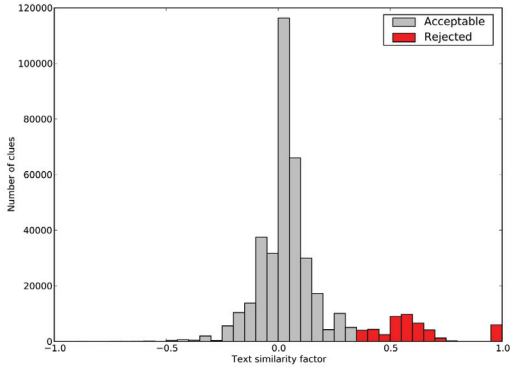

Figure 3: The distribution of text similarity factors among statements that passed previous checks.

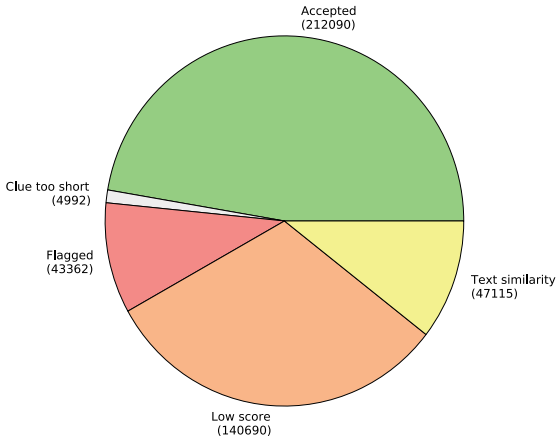

Figure 4: The overall classification of statements collected from Verbosity. "Accepted" represents statements that passed all the checks and were accepted into ConceptNet.

While the relations associated with these negative statements were not always reliable, they shared an informative core idea – "there is an important difference between these two things". We rephrased negative statements that used the word "not" at the start of the clue or the "opposite" relation to be in the form "*x* is not *y*", so that they would form the negation of an IsA statement in ConceptNet. (We simply discarded clues that used the word "opposite" in the *text*, as mentioned before, because too many of these appeared to be referring to previous guesses.)

Similarly, although statements with the relation "'it looks like" were not often literally true, they could usually be replaced by weaker, more correct statements with the relation "it is related to".

Evaluating and Comparing Verbosity Data

To evaluate the result of our filtering process, we ran an evaluation similar to previous evaluations that we have run

on ConceptNet (Speer, Havasi, and Lieberman 2008). People participating in the evaluation went to a Web page that presented an assortment of ConceptNet-like statements, and asked them to evaluate the quality of each statement. These statements came from shuffling 20 statements from the filtered Verbosity data, 20 statements from ConceptNet that did not come from Verbosity, and 20 randomized statements. The randomized statements were created by taking real Verbosity statements and randomly swapping their keywords or clues with each other, to produce generally nonsensical statements that looked like they could come from Verbosity.

Our participants were 15 people who were familiar with the Open Mind Common Sense project, who were not told as they were taking the study what the different sources of statements were. These participants as a whole evaluated 631 statements. Participants had seven options for how to assess each statement: "Generally true", "Generally true (but grammar is bad)", "Sometimes true", "Sometimes true (but grammar is bad)", "Don't know / Opinion", "Not true", and "Doesn't make sense".

The options containing "(but grammar is bad)" were used to take into account that a statement, especially given the constraints of Verbosity, could have awkward grammar and still be valid. ConceptNet, too, contains statements with bad grammar, often resulting from an input frame being too constraining or from answering a question that was automatically generated by the system. Out of 179 ConceptNet statements that were rated positively, 23 of them (12.8%) were assessed as having bad grammar. Of 169 such Verbosity statements, 49 (29.0%) were assessed as having bad grammar.

Disregarding assessments of the grammar, the breakdown of responses is shown in Figure 5. For the purposes of comparison, the above responses were also converted to a 7-point quality scale, using consecutive integers from -2 ("doesn't make sense") to 4 ("generally true").

Source	# samples	Mean score	Std. error
ConceptNet	211	2.787	0.123
Verbosity	209	2.378	0.148
Random	211	-1.014	0.0776

Table 1: The scores resulting from the evaluation, on a scale from -2.0 to 4.0.

We confirmed the difference in quality across the different sources, as shown in Table 1, using a one-way ANOVA. The average score varied significantly among the three different data sources ($F(2, 628) = 305.6$, $p < .0001$). A Tukey HSD test showed that the difference in score between randomized statements and the two other data sources was significant at the $p < .01$ level, and the difference between Verbosity and the rest of ConceptNet was significant at the $p < .05$ level.

Using Verbosity scores

Like ConceptNet, Verbosity assigns scores to its statements according to how many people have asserted them. We have used the Verbosity score so far as a rough filter, by throwing

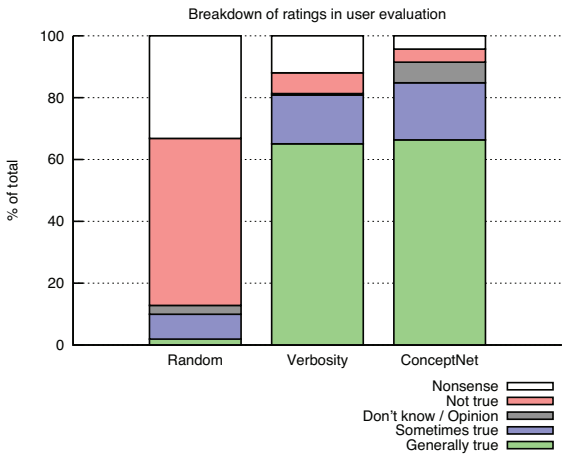

Figure 5: The results of the evaluation, broken down by the three sources of statements and the evaluations that people gave them.

out some of the statements with a score of 1, but aside from that, we have not taken the Verbosity score into account. Is it possible that we could be getting more information about the validity of statements by taking into account higher scores?

In Figure 6, we plot the Verbosity scores of the statements used in our evaluation (on a log scale) against the score each statement received in our evaluation. As both scores are constrained to integers, we make the points distinctly visible by adding random "jitter" with a standard deviation of 0.1 to the X and Y coordinates of each point.

Perhaps surprisingly, a linear regression on these results shows no significant correlation between the Verbosity score and the evaluation score, with a correlation coefficient of only $r = 0.056$. There were a considerable number of well-rated statements that were asserted only once, for example, as well as some poorly-rated statements with higher scores.

Inaccurate relations

Most of the difference in ratings between Verbosity and ConceptNet came from the Verbosity statements rated as "nonsense". When we examine these statements, one common problem we see is that the describer would give a word that had some common sense relation with the target word, but it was not the relation being prompted for.

Some examples of statements rated "Doesn't make sense" were "*leg* has *lower limb*", and "*toy* is a kind of *little*". Statements rated "Not true" included "*sail* is *a boat*" and "*servant* has *paid help*", and statements rated positively but with bad grammar included "*pearl* is related to *shiny*" and "*produce* is a type of *fruits vegetables*". These statements could all have been improved by filling the same words into a different frame, such as "sail is an action you take using a boat", "a servant is paid help", and "produce includes fruits, vegetables".

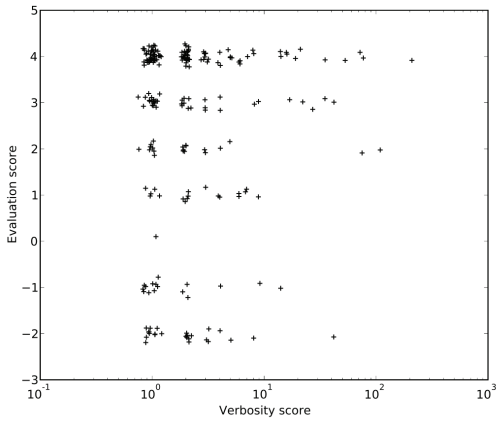

Figure 6: A comparison between the number of times each statement was asserted in Verbosity versus the score it received in the evaluation.

Further Thoughts

The speed of knowledge acquisition using a successful GWAP is much faster than the standard volunteer-based interface. But the data collected from GWAPs is also bound to be somewhat noisier, as they are designed to bring in those not sympathetic to or interested in the underlying greater purpose of collecting data for computer science applications. In this paper, we have identified many sources of noise and ways to correct or disregard the erroneous data they produce. Many of these, such as use/mention errors, would apply to any similar system as well, so such a system could benefit from the corrections we have described here.

Future games can be built these issues in mind, and one way to do so is to similarly crowdsource the verification step. This way, not only does the game collect new data, it also weeds out data that will not be useful for the target system. Such verification systems can be a separate game, or could be built into the game itself, such as by allowing the guesser to reward good clues from the describer. A built-in verification game has been used to good effect in the Chinese-language pet game (Kuo et al. 2009).

Overly-strict constraints in a game such as Verbosity are a common source of error, as players subvert the constraints to get their idea across. As we have seen, because Verbosity chooses the relations that the describer must fill in, the describer often ignores the relation and says what they mean. One lesson we can learn is that games such as this should remove constraints that are not important to the gameplay.

A future version of Verbosity could mitigate this problem by allowing a choice of relations, leaving the problem of how to fix existing data. One idea we propose is a part of the game in which players would choose, out of two relations, which one they thought connected the words most meaningfully (with "neither" available as an option). The players are given the incentive of a bonus score, if both agree on the same statement. If enough players agree against the status quo, the existing statement could then be changed.

Having gone through the process of augmenting OMCS using a GWAP, this leaves us with many ideas and opportunities for future work. We hope that the lessons we have learned and the processes we have developed are useful in future endeavors.

References

Cycorp. 2009. Cyc web games. `game.cyc.com`.

Garcia, D.; Havasi, C.; and Todd, K. 2002. Interactively improving the OpenMind database: A game for interactive OpenMind inprovement (GIOMI). Media lab working paper, Massachusetts Institute of Technology.

Havasi, C.; Speer, R.; and Alonso, J. 2007. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*.

Havasi, C. 2009. *Discovering Semantic Relations Using Singular Value Decomposition Based Techniques*. Ph.D. Dissertation, Brandeis University.

Kuo, Y.-l.; Lee, J.-C.; Chiang, K.-y.; Wang, R.; Shen, E.; Chan, C.-w.; and Hsu, J. Y.-j. 2009. Community-based game design: experiments on social games for commonsense data collection. In *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation*, 15–22. New York, NY, USA: ACM.

Lenat, D. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 11:33–38.

Lieberman, H.; Smith, D.; and Teeters, A. 2007. Common Consensus: A web-based game for collecting commonsense goals. *Workshop on Common Sense for Intelligent Interfaces ACM* . . . .

Luis von Ahn, M. K., and Blum, M. 2004. Verbosity: A game for collecting common-sense knowledge. In *Proceedings of ACM Conference on Human Factors in Computing Systems, CHI*.

Singh, P.; Lin, T.; Mueller, E. T.; Lim, G.; Perkins, T.; and Zhu, W. L. 2002. Open Mind Common Sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems*, 1223–1237. London, UK: Springer-Verlag.

Speer, R.; Krishnamurthy, J.; Havasi, C.; Smith, D.; Arnold, K.; and Lieberman, H. 2009. An interface for targeted collection of common sense knowledge using a mixture model. In *Proceedings of the Intelligent User Interfaces Conference*.

Speer, R.; Havasi, C.; and Lieberman, H. 2008. AnalogySpace: Reducing the dimensionality of common sense knowledge. *Proceedings of AAAI 2008*.

Speer, R. 2006. *Learning Common Sense Knowledge from User Interaction and Principal Component Analysis*. Ph.D. Dissertation, MIT Media Lab.

von Ahn, L., and Dabbish, L. 2008. General techniques for designing games with a purpose. *Communications of the ACM* August.