Echo Chamber:
A Game for Eliciting a Colloquial Paraphrase Corpus

Chris Brockett and William B. Dolan

Natural Language Processing Group, Microsoft Research
One Microsoft Way, Redmond, WA 98052
{chrisbkt, billdol}@microsoft.com

Abstract

The problem of semantic equivalence, or paraphrase, is a fundamental one for applications that "understand" natural language. Learned approaches to this problem face a lack of colloquial training data with which to build models. This paper describes, *Echo Chamber*, a game aimed at collecting sentential paraphrases from web users. Much of our current focus is designing a framework that makes this potentially burdensome task engaging and challenging. The game draws on elements of enduring pen-and-paper games such as Battleship and Hangman, and incorporates a time component to impart a sense of urgency. In the final version, it is intended that automated validation of input will ensure that the game is scalable and can collect high quality data without editorial intervention. In this paper, we discuss design considerations and issues emerging from the prototype of this game.

Introduction

The ability to characterize superficially distinct word sequences as "meaning the same thing" is important for many software applications, including command-and-control, summarization, dialog-handling, question-answering, and search. Recent research into semantic equivalence or paraphrase (Barzilay & McKeown, 2001; Lin & Pantel, 2002; Shinyama et al, 2002, Barzilay & Lee, 2003, Pang et al., 2003; Quirk et al 2004, Dolan et al, 2004) has treated it as a machine learning problem, training statistical models on pairs of sentences or sentence fragments with parallel meanings. Paraphrase is also of likely concern in AI areas in such as commonsense reasoning: semantically equivalent statements can be used to extract inferences (Lin & Pantel, 2002).

Much of the work in this area has relied on news data, which is available in large quantities and which is rich in alternative accounts of the same set of facts, e.g.:

A child who lives near a petrol (gas) station is four times more likely to develop leukemia than a child who lives far away from one, according to a new study

Living near a petrol station may quadruple the risk for children of developing leukaemia, new research says.

News, however, is ultimately an inadequate source of data about paraphrase relationships. News tends to be sharply limited in genre and domain; the majority of articles are written about a relatively small number of topics, including acts of violence, politics, business, and sporting events. Journalists are bound by tight stylistic conventions, and must normally limit themselves to a formal diction that is rich in long, complex sentence types.

A broad-coverage solution to the semantic equivalence problem will require immense amounts of training data. This data must cover a far broader range of domains and text styles than is represented in the news genre. Corpora of colloquial text paraphrases are especially difficult to obtain. The cost of hiring humans to create a corpus manually would be utterly prohibitive, so a model in which data is contributed voluntarily by large numbers of web users is highly attractive. The question is: how can we elicit the cooperation of contributors on a large scale, and motivate them to produce, over time, data of sufficient quality and quantity to address this need?

Previous Work

The only previous attempt to collect sentence- or phrase-length paraphrases that we are aware of is Chklovski's *Paraphrase Game* at http://ai-games.org/paraphrase.html. This game involves asking users to guess paraphrases for sentences or phrases in the military medical domain. The player is rewarded for matching any previously-created paraphrase. If there is no match, the player is given "hints" in the form of partly-specified solutions. While many aspects of Chklovski's game are intrinsically appealing, we believe that his basic methodology can be improved on to make a paraphrase-eliciting game more likely to harvest the

huge numbers of paraphrase pair examples with which to construct statistical models.

Echo Chamber

In this paper, we discuss our experience with the prototype of *Echo Chamber*, a game it seeks to elicit from web users new sentence level paraphrases that can be used as training data for statistical paraphrase detection and generation models for application in Information Retrieval or automated rewriting. Our emphasis is on trying to design a game that:

- Will prove addictive to players who enjoy word games, encouraging them to donate many paraphrases
- Automates the process of validating that the contributed data is good quality

Echo Chamber currently exists in only prototype form, and we are in the process of having pilot subjects test it on different datasets in an effort to define and refine its behavior. Our initial investigation suggests that some players do find the game compelling, and that some can even be induced to continue playing for extended periods. In the near future, we hope to launch the game on the Microsoft corporate intranet, but our objective is to construct a game that can eventually gain wide exposure on the World Wide Web. It is hoped, moreover, that we will be able to make some of the data collected available to the research community. Because the game is still being prototyped, the present paper is unabashedly exploratory, and focuses chiefly on issues of design and the results of initial usability testing on a standalone prototype. Its authors hope, however, that it will contribute to discussion of the ingredients of a successful data-eliciting game.

Design Considerations

A key desideratum, in our view, is that the game should be compelling enough to garner a large cohort of repeat players who provide high quality contributions consistently over time. This is an especially important issue where a corporate entity offers a data eliciting game, since users can reasonably be expected to want some form of compensation for their contribution.

For our part, we seek to appeal to potential players' sense of fun and willingness to entertain a challenge, and to provide the players with personal satisfaction as a reward for their performance. An extensive literature now exists on successful game design (usefully accessible through Salen and Zimmerman, 2004); in the case of data collection games, the issues are complicated by the fact that the game has an ulterior motive on the part of the game provider that can detract from enjoyment unless care is taken to ensure a high-level of player engagement. Several considerations or assumptions, in particular, drive our design:

The Task Cannot Feel like Work.

Efforts to collect information on the web often involve directly collecting factual data that involves the contributors' overt knowledge of language or the world, for example, as in OpenMind (Singh, et al., 2002). Such tasks will certainly appeal to a limited group of web users, for example, those personally interested in AI/Common-sense reasoning who may be willing to devote large amounts of personal time to a project. However, large scale data collection of the kind that we require is unlikely to be possible on this basis. While web users might play a game briefly out of curiosity, they will in general only return to generate large amounts of data if the game is engaging and entertaining.[1]

This presents a something of a paradox when it comes to collecting paraphrases, since the task of generating rewrites turns out not to be very much fun, even for quite verbally adept contributors. Subjects in pilot experiments reported that they found it challenging and time-consuming to come up with multiple paraphrases of short sentences like:

Don't worry about it; I'll figure something out.

Dolphins are mammals that look like fish

The outcome of pilot experiments quickly persuaded us to abandon plans to ask players to donate more than one paraphrase per seed sentence. Our problem, then, is how to persuade web users to perform a task that is inherently tedious. The solution, here, is to apply the Tom Sawyer Principle: since we cannot avoid asking users to come up with a paraphrase of an input string, we seek to disguise the work aspect of the task deeply within the game framework. From the players' perspectives, the primary purpose of the exercise must be amusement; we want players to be happy to donate their words in return for an entertaining challenge.

Competition and Collaboration Motivate Players

The most successful example of a web-based data collection strategy that we are aware of is *The ESP Game* (von Ahn & Dabbish, 2004; http://www.espgame.org/). A crucial aspect of this image-labeling game is that two randomly-paired players must collaborate in order to win points. Users must form collaborative alliances, creating a sense of teamwork that is heightened the sense of urgency provided by a ticking countdown clock. Players are rewarded for both teamwork and speed, and the login names of the top scoring players are posted on the web. We believe that this sense of urgency is a crucial

[1] Casino operators have already perfected this notion; ideally the game needs to be as addictive as gambling.

component of *The ESP Game*'s success in acquiring and sustaining large numbers of players over time, and hope to replicate this in our paraphrase collection strategy.

In the case of sentential paraphrases, however, the number of variables involved is much greater than in the relatively constrained task of image identification posed in *The ESP Game:* it is somewhat improbable, for example, that two players might simultaneously converge on the same paraphrase when possibly as many as a dozen words need to be matched. If player collaboration and competition are to play a part in a paraphrase collection game, a more realistic scenario is one in which one player contributes a paraphrase that her opponent attempts to guess. This suggests that we need to have some way of ranking players, and permitting only those who have a reliable history of valid contributions ("trusted" players) to engage each other directly. In most successful games there is little or no challenge or enjoyment in being set up against an incompetent opponent.

Overt Data-Vetting will Annoy Contributors

Unfortunately one problem likely encountered by any effort to collect data from self-selecting populations on the World Wide Web is that not all contributors will in fact provide reliable data. Epithets, random character strings, failed efforts at humor, and well-meant but incorrect or irrelevant inputs all present challenges.

In the case of paraphrases, as we have already noted, the number of contributions what would need to be garnered for use in statistical models is so large that it is uneconomic to subject them to human review. One conceivable strategy might be to have users directly vet the contributions of others. On the basis of our experience with paraphrase identification and writing tasks in a variety of contexts, however, hand verification and hand editing are likely to irritate or bore users even more than generating paraphrases themselves, thereby defeating the whole point of the game.

Here again, *The ESP Game* provides an excellent example of how a quality-vetting procedure can be invisibly integrated into the data collection process. Contributions to *The ESP Game* are only deemed relevant if both users happen to hit on the same label. This creates a sense of cooperation between players and provides robust assurance of data quality.

In a similar manner, we believe that it should be possible to validate paraphrase contributions on the basis of the data itself. Because paraphrasing involves implicit, often unconscious, knowledge of language and the real world, it has the virtue of being amenable to validation using statistical techniques of the kind used, for example, in the Speech Recognition and Information Retrieval communities. By integrating such techniques with models of players who of players who provide reliable input, we expect to be able develop scoring methods that will allow us to determine which contributions are most likely to constitute good paraphrases. By bootstrapping on these models, we expect that it will be possible to promote players to a "trusted" status where they can interact directly with other "trusted" players, thus providing the necessary element of collaboration and competition.

Familiar, Enduring Games Offer the Best Model

The final problem is how to make the game inherently enjoyable. Many successful board games, computer games, and television game shows are built from combinations of features of popular pencil-and-paper games. These time-tested game elements provide a library of design features that can be used for new game; we assume that a game constructed from such elements has the best chance of being both playable and enjoyable. *Echo Chamber*, in its present instantiation, can be thought of as a hybrid of Battleship and Hangman. Battleship is a game in which players attempt to guess which unseen square have been filled. In Hangman, the purpose is to guess the word before the other player completes a stick figure of a man being hanged, one pencil stroke for each incorrect character.

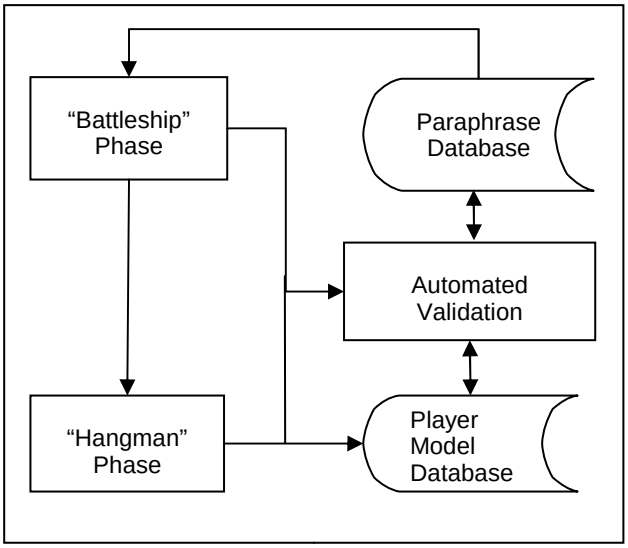
Design Features

Game Structure

Each round of *Echo Chamber* proceeds in two distinct phases: a "Battleship" phase and a "Hangman" phase. "Battleship" phase is principally a data elicitation phase, the player's input being collected for evaluation. The second, or "Hangman", phase is played largely as a reward to players for their contribution, and provides an opportunity to collect additional player data for developing a model of the player's reliability as a contributor.

In the "Battleship" phase, the player is asked to paraphrase a source sentence, with no hints provided, while a countdown clock ticks away. The player is awarded points for the number of words that match a target paraphrase, weighted by an automated metric that gauges string-edit distance from the original seed sentence. Bonus points are also awarded if the sentence matches another paraphrase already in the database. Future versions will also include such features as the probability of the sentence assigned by a language model and the number of times that the sentence has been submitted by other players. This weighting, which is intended to provide some measure of both the originality and well-formedness of the contribution, is not displayed to the user, but is used in computing any points assigned in the subsequent stage of the game.

At the end of the "Battleship" phase, those words that the player successfully guessed at displayed on the screen, and players proceed the second stage (the "Hangman" stage).



Fig, 1. Data collection and evaluation in *Echo Chamber*

This is the phase that is intended to provide the most meaningful challenge and the bulk of the entertainment (reward) for the player. As in the *Paraphrase Game*, hints are now provided; in our game, however, the hints are random letters distributed across the string, rather than words. New letters appear at 3.5 second intervals as a countdown clock ticks away and the number of available points for a correct guess steadily decreases. Players are thus under time pressure to solve the puzzle before the game solves it for them, resulting in massive loss of points. To increase tension, it is possible to increase the speed with which hints are introduced as more rounds are played. The highest scores are awarded for those paraphrases that share few words with the original string and yet are guessed quickly, in fewest attempts and with fewest hints. A screenshot illustrating the "Hangman" phase, taken from the standalone prototype, is shown in Fig. 2.

Data Collection

At present, the direct collection of paraphrase samples is conducted only during the "Battleship" phase. We are not currently considering doing this during the "Hangman" phase since inputs during this phase may be somewhat incremental, making it difficult to distinguish dirty data from valid.

The "Hangman" phase furnishes additional information that may be interpreted as indicative of the verbal competence and reliability of the player. In addition to the final score, this data might include the mean score over a series of games, how long the player took to complete each game, and the number of hint characters and guesses the player required. In general, we expect a player who reliably comes up with a solution quickly and with few attempts to be a better candidate for a "trusted" player, than one who is slow and requires many attempts before succeeding. Players are informed that their input is being collected, and, importantly, that the data may be used for purposes other than the game itself.

The "Hangman" phase also provides additional validation of the data itself, an important consideration when player supplied data is used. For example, the time frame within which players can complete the partially-specified sentence is probably a function of the plausibility of the target paraphrase. Paraphrase sentences that multiple players can come up with quickly and easily are more likely to be contain high-frequency elements that are of interest in building statistical paraphrase models, while over time, those paraphrases that are consistently guessed less successfully can be identified and dropped out of the models.

Evaluating the Prototype

This project has undergone several iterations with a simplified standalone prototype in anticipation of posting a web-enabled version on internal and eventually on external websites. Experimentation is taking place offline to avoid technical issues relating to website management.

The most recent version was tested on a group of in-house volunteer players, who were invited to play unsupervised until they were ready to quit. Five individuals attempted the game in a 24 hour period and returned their logs by email. Although this sample is very small, it presents sufficient data to establish baseline expectations of player behavior that can help frame future investigation, and affords some sense of what issues might lie ahead, particularly when rendering the game interactive on the World Wide Web.

The prototype was seeded with 119 sentences, each matched against 2-4 hand-created target sentences, for a total of 331 sentence pairs. Of these, 104 source sentences were taken from an elementary school science text (Victor and Kellough, 1989), the remainder being extracted from recent news articles. During the game, source sentences were selected randomly from this data set for presentation to the players, who were then asked to produce paraphrases. The most intrepid of our volunteers lasted a promising 48 minutes, playing a total of 38 games; the least persistent abandoned the effort after only 2 minutes. The other three volunteers played between 7 and 18 successful games.

Although we have some distance to go before the game could be described as "compelling," we are nonetheless encouraged by informal comments from the volunteers (to say nothing of the apparent presence of one bona-fide addict among their number) to think that, given a large enough base population of experimenters among web users, a future version of this game might eventually succeed in

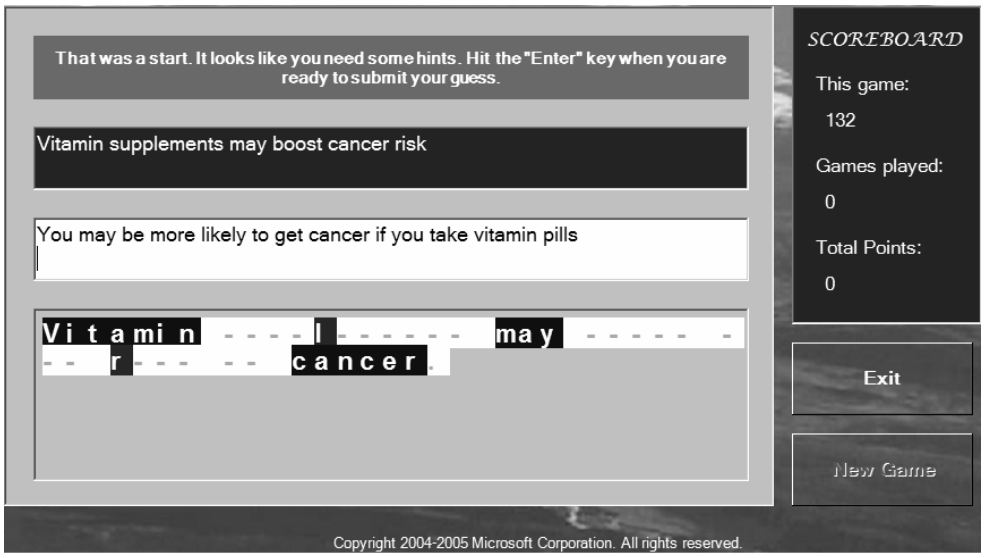


Fig. 2. Screenshot of Echo Chamber prototype depicting the "Hangman" Phase. Full words shown were successfully guessed during the preceding "Battleship" phase. Single letters are being dropped in to provide hints to the player.

attracting sufficient numbers of returning players to sustain a data collection effort

.

Even the miniscule dataset provided by our volunteers elicited 60 new unique and valid paraphrases, of which only one exactly matched the target paraphrase during the "Battleship" phase. Only two of the full sentences that were submitted were completely unrelated to the source or target sentence. If the game can be made attractive enough, it should be possible to acquire a large number of acceptable paraphrases in this manner, although data sparsity probably mandates that significant body of paraphrase pairs will need to be collected before it will be possible to subject input to automated validation techniques with any reliability.

Zero or Ill-formed Responses

The data submitted in the volunteers' logs reveal a number of issues, both expected and unexpected. One question we had was how much fragmentary input we would receive. Within our volunteer group, input was overall well-formed.

One outlier volunteer, however, exhibited a high rate of fragmentary or zero responses as the rounds progressed, that player having apparently decided to forgo the *Battleship* phase entirely and to participate only in the *Hangman* phase. Some 44.4% of this player's 18 rounds involved zero responses. It would seem that for this player, the *Battleship* phase is not compelling enough, and that its data collection function will not succeed unless players can be more strongly motivated to submit input—perhaps by having the player forfeit the round, or requiring the player to input before continuing.

Single and multi-word fragmentary responses, of which there were a smaller number, need to be properly detected so that they can be rejected. In the final version of the game, "spam filtering" techniques, including stop word lists and use of language models, will eliminate sentences that contain obscenities, random strings, or strings that are multiply submitted as paraphrases of unrelated sentences.

Spelling errors are an obvious problem that may be addressable using a contextual spelling checker algorithm. More problematic, however, are syntactically ill-formed strings that reflect typing errors on the part even of native speakers of English. In the following, for example, the error is not caught by the grammar checker in Microsoft Word: it is not clear that input like the following can be successfully rejected, especially if players (perhaps non-native speakers of English) eventually come up with identical input.

SOURCE: *The voice box really is like a box, and it is made out of cartilage*

INPUT: *the larynx is shaped like box [sic], and is formed from cartilage*

Filtering Irrelevant Same-topic Sentences

In our volunteer logs, we encountered input sentences that are superficially similar with the source and target in that

they share some words in common, but which are semantically unrelated and do not constitute paraphrases:

SOURCE: *Hackers wasted little time in exploiting the flaw*

INPUT: *Exploiting weaknesses in code is the passion of hackers.*
TARGET: *Hackers wasted little time in taking advantage of the flaw*

A variant of the above and possibly harder to detect automatically, are sentences that contain a proposition that is completely antithetical to that of the source and target sentence.

SOURCE: *Rumors of the death by piracy of the global music industry may have been exaggerated*

INPUT: *It is an exaggeration to say that piracy is dead in the global music industry.*

TARGET: *Rumors that the global music industry may die from piracy may have been exaggerated.*

Even highly educated, verbally competent individuals may introduce noisy data of these kinds under the pressure of the game, and on the World Wide Web we would expect to run into many more problems of this ilk. However, the problem that input like the two examples above presents may turn out to be self-correcting: while these particular examples might well slip through a paraphrase detection module, they would likely be discarded as long as other players failed to make identical errors.

Will the Game Elicit Colloquial Forms?

One issue that was of interest to us was whether *Echo Chamber* is actually capable of harvesting the more casual kinds of expression that we are seeking. This is evidently the case: informal game-playing conditions do have the desired effect of eliciting colloquial forms even though the volunteers were not explicitly instructed to provide these. For example:

SOURCE: *The Canadian government is negotiating to secure passage out of China for 44 North Korean refugees*

INPUT: *The Canadians are trying to get 44 North Korean refugees out of China*

The science text sentences, on the whole, proved less amenable to colloquial paraphrase by our volunteers, though we do find evidence of colloquial forms making their way into the input: the word “your”, for example, in the following input is characteristic of a style typical of spoken language.

SOURCE: *The gullet, or esophagus, leads to the* stomach

INPUT: *Your esophagus is connected to your stomach.*

Importance of Seed Data Selection

One significant issue that emerged in the course of evaluating the prototype, however, relates to the nature of the data used to seed the game with source and target sentences. In practice, the source sentences from the science text (Victor and Kellough, 1989) often turned out to be extremely difficult (and tedious) for our contractor to paraphrase—whether in more colloquial language or otherwise—leading to many instances of target sentences that involved trivial reordering of words. This had two undesirable side effects. First, it was apparent that players quickly became trained by the seed paraphrase pairs to provide simple rewordings that involved little or no new lexical material.

SOURCE: *In a gas, molecules move very fast and are far apart*

INPUT: *Molecules in gases move very fast and are far apart*

TARGET: *In a gas, molecules are very fast moving and are far apart.*

The problem of players being trained to the game directly contradicts our effort to elicit imaginative linguistic insights. We expect, however, should to some extent be alleviated in later versions, with a different seed data set, and again when the game becomes interactive on the World Wide Web, since a greater variety of opponents will be encountered. By the same token, however, since the target paraphrases were created by humans in the first place, this issue also demonstrates that in an interactive environment the creativeness of an opponent’s input has the potential limit a player’s performance in a given session. This assigns even greater importance to the need to rank players and ensure that they are well matched.

The second, related side effect is that players reported that they found it discouraging to come up with an insightful or imaginative paraphrase, only to be presented with a mundane reordering of the phrases in the sentence as a solution. Looking at an example like the following, it is easy to see why the game would not be particularly exciting in these circumstances, and why some players might quickly get bored, resentful, or even rebellious.

SOURCE: *Fine performances from a stand-out cast punctuate this cinematic offering.*

INPUT: *This movie offers great performances by a terrific cast*

TARGET: *Excellent performances from a stand-out cast punctuate this cinematic offering.*

Additional research is needed into what kinds of example sentences are most likely to be amenable to paraphrase and to be of engaging interest to players. Since it would appear that many definitional sentences of the kind found in the scientific text, for example, are relatively limited in their potential for rewriting, such instances might have to be presented to players sparingly if at all. Ensuring that the seed sentences are entertaining and that source and target are sufficiently dissimilar to be intellectually challenging is likely to be a major factor if the game is to succeed in the future. This will probably necessitate the development of some form of lexical distance metric between source and target so as to ensure that players are properly and consistently rewarded with a reasonably intriguing puzzle.

User Interface

In providing a game environment, the quality of the user interface is itself a critical component in sustaining user enjoyment of, and interest in, the game. All our volunteers indicated that they found the crude prototype interface counterintuitive and that the transition between the two phases was especially difficult. In "Hangman" mode, several volunteers indicated that they wanted to insert directly into the text box in which hint characters were being displayed. Clearly further interface improvements will be in order.

Future Directions

Our next step will be to move our prototype onto the corporate intranet for large scale testing. As we do so, we will be addressing in greater detail a number of issues that remain to be implemented.

Automated Validation

Our plan is recycle as target sentences only those sentences independently proposed by multiple players, and to automatically evaluate their goodness both in terms of paraphrase and general well-formedness against a variety of features including a language model, and the players' histories, represented as confidence scores on the input. Only those sentences that meet a certain threshold are added to the game as new source or target sentences.

Fully viable automated validation of paraphrase will not be possible, however, until certain additional technologies are in place. In the context of other projects, we are currently working on developing statistical models of paraphrase detection that can be adapted to support automated validation. It may be helpful to have players facilitate the development of paraphrase detection systems by providing spot judgments on the quality of the targets at the end of the game: this information can then be utilized to create tagged data sets with which to bootstrap models.

In addition to the potential cost savings in evaluating the elicited data, automated validation is expected to constitute an important component of realtime feedback to users in *Echo Chamber*. By letting players know that a recently submitted paraphrase is in some way distinctive or matches paraphrases submitted by others, we expect that their experience of the game will be enhanced and that they will be encouraged to contribute more interesting variant sentences in the future.

Eliciting New Paraphrases through Player Interaction

In the early stages of data collection, the game is presented as a simulated multi-player game, using seed sentence pairs that have been hand generated. Inputs from other players will not be employed until a sufficient quantity of data has obtained and validated, and a cadre of "trusted" players has been identified who consistently supply reliable input.

One major issue that the *Echo Chamber* project faces is the question of how to elicit reliable new paraphrases where no target paraphrase data is already available. As described so far, *Echo Chamber*, is entirely dependent on data for which paraphrases have already been identified in corpora or manually constructed. At some point, it will be necessary to shed this limitation and to allow "trusted" players to contribute entirely new data.

In its present instantiation, in which the player engages the machine, *Echo Chamber* does not yet meet our primary desideratum of exploiting players' collaborative and competitive needs. Since we wish to avoid the multiplayer solitaire effect, in which the players do not truly interact with each other, we will be exploring ways to create a genuinely interactive multiplayer environment where trusted users can play each other by offering their own paraphrases for entirely new sentences.

Conclusions

At the time of writing, the game environment that we have described remains under development and is not yet ready for web release. It is expected that in its final form it will have evolved considerably from what is presented here. In this paper we have addressed some of the issues are arising in attempting to develop a compelling game environment for eliciting unconscious linguistic knowledge about the relationships between sentences. Numerous additional usability issues will doubtless need to be resolved before a

public release of *Echo Chamber* is possible. We anticipate, however, that implementation of automated paraphrase detection techniques and other data verification methods will permit scalability and minimize the need for human editorial intervention. Once these and related issues are resolved, and with an appropriate level of publicity, we believe that this game environment offers the promise of collecting large corpora in a comparatively short time, and may provide a model for other volunteer contribution tasks. By providing entertainment and appealing to the users' sense of fun and playfulness (along with perhaps a certain grim determination to persist in solving the puzzle) we expect that *Echo Chamber* will offer players emotional rewards that will obviate the need to rely on altruistic contributions.

Acknowledgements

We would like to thank Anthony Aue, Simon Corston-Oliver, Deborah Coughlin, Chris Quirk, Lucy Vanderwende, and other members of the Microsoft Research NLP group for their feedback on this project. Our appreciation also goes to Monica Corston-Oliver for contributing much vital advice and to members of the Butler Hill Group for providing the seed data.

References

von Ahn, L. and Dabbish, L. 2004. Labeling Images with a Computer Game. In *Proceedings of CHI 2004*.

Barzilay, R. and McKeown, K. R. 2001. Extracting Paraphrases from a parallel corpus. In *Proceedings of the ACL/EACL*.

Barzilay, R. and Lee, L. 2003. Learning to Paraphrase; an unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL2003*.

Dolan William. B., Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Lin, D. and Pantel, P. 2001. DIRT - Discovery of Inference Rules from Text. In P*roceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*, pp. 323-328.

Pang, B., Knight, K., and Marcu, D. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of HLT/NAACL, 2003*.

Quirk, C., Brockett, C., and Dolan, W. B. 2004. Monolingual Machine Translation for Paraphrase Generation, In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 25-26 July 2004, Barcelona Spain, pp. 142-149.

Salen, K. and Zimmerman, E. 2004. *Rules of Play: Game Design Fundamentals.* Cambridge, MA: The MIT Press.

Shinyama, Y., Sekine, S., and Sudo, K. 2002. Automatic Paraphrase Acquisition from News Articles. In *Proceedings of NAACL-HLT*.

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 *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems. Lecture Notes in Computer Science*. Heidelberg: Springer-Verlag.

Victor, E. and R. D. Kellough 1989. *Science for the Elementary School*, Seventh edition. New York: Macmillan Publishing Co.