Online Gaming for Crowd-sourcing Phrase-equivalents

A. Kumaran
Microsoft Research
Bangalore, India
a.kumaran@microsoft.com

Melissa Densmore
University of Cape Town
Cape Town, South Africa
mdensmore@acm.org

Shaishav Kumar
Microsoft Research
Bangalore, India
v-shaisk@microsoft.com

Abstract

We propose the use of a game with a purpose (GWAP) to facilitate crowd-sourcing of phrase-equivalents, as an alternative to expert or paid crowd-sourcing. Doodling is an online multi-player game, in which one player (*drawer*), draws pictures on a shared board to get the other players (*guessers*) to guess the meaning behind an assigned phrase. In this paper we describe the system and results from several experiments intended to improve the quality of information generated by the play. In addition, we describe the mechanism by which we take candidate phrases generated during the games and filter out true phrase equivalents. We expect that, at scale, this game will be more cost-efficient than paid mechanisms for a similar task, and demonstrate this by comparing the productivity of an hour of game play to an equivalent crowd-sourced Amazon Mechanical Turk task to produce phrase-equivalents over one week.

1 Introduction

While it is fairly well known when individual words have the same meaning, it is far more difficult to determine when phrases or even sentences carry the same basic idea. While it might be possible to address this task with machine learning techniques, building a corpus of sentences from which to seed a database requires human intelligence. We suggest a *game with a purpose* (GWAP) that will serve to generate phrases with similar meanings, while simultaneously providing meta-information about the quality of the match. In this drawing game, called *Doodling*, individuals compete in groups to guess the meaning behind a given drawing that is being drawn by one designated *drawer* trying to convey a given phrase or a short sentence. The designated drawer decides when a guessed phrase matches the source phrase. For example "*How far is the airport*?" might match semantically "W*hat is the distance to the airport*?" In addition, the drawer can indicate for each partial guess how close it is on a scale of 1-3 to help the guessers converge on phrases that will match the given phrase or sentence. We then pass all of the guesses and annotations through an SVM classifier to automatically identify potential phrase-equivalents. In this study we examine several techniques for using this system to generate high quality data while also making the game more enjoyable. We measure the efficacy of each technique by comparing our results to a gold standard: using human evaluators to rate the phrase matches generated through the game manually. We also compare Doodling to a paid crowdsourcing paradigm – Amazon's Mechanical Turk – to source phrase equivalents for the same set of phrases, and we show that our approach might be cost effective for large scale sourcing of paraphrases of equivalent quality.

2 Background

In this section we define the problem we are trying to address, and discuss the various ways it has been approached in the past.

2.1 Phrase-equivalents & Evaluation of Quality

In this paper, we define phrase equivalents (PEs) as text elements – phrases or short sentences – that have same or similar semantic content, but with surface structure different from each other. PEs are similar to *paraphrases*, but broader in scope, inclusive of partial matches in meaning as well as complete paraphrases. PEs are useful for many NLP systems from simple language modelling and smoothing, to complex Machine Translation technology for generation of a surface form in the target language. Most existing corpora are hand-created, and hence they tend to be small in size, and available only in limited languages and domains. Other data driven approaches – such as, creation of paraphrases using monolingual machine translation (Quirk et al., 2004), mining inference rules from text corpora (Lin & Pantel, 2001), or paraphrase extraction from parallel corpora (Dolan et al., 2004) (Barzilay & McKeown, 2001) – were shown to be effective, but such approaches require significant seed corpora which are available only in limited domains and languages. In addition, the (Lin & Pantel, 2001) approach can generate equivalents using user defined patterns, and may not be appropriate for generating loosely related conceptual paraphrases like the human generated ones that Doodling may generate.

The criteria used for evaluating phrase equivalents differ vastly in research literature, ranging from conceptual equivalence (Barzilay & McKeown, 2001), to interchangeability (Ibrahim et al, 2003), to preservation of grammatical correctness and semantic equivalence (Callison-Burch, 2005), and the standard metric of BLEU score (Callison-Burch, 2005; Papineni et al., 2002). In general, there is no accepted standard model for measuring quality, hence we adopted manual annotation by experts.

2.2 Crowdsourcing & Games with a Purpose (GWAP) for Computational Linguistics

Many flavors of crowdsourcing paradigms exist for the creation of language data. From the *for-pay* model where the contribution is for monetary rewards (Callison-Burch, 2009; Irvine & Klementiev, 2010; Chen & Dolan, 2011), to the *for-recognition* model, where the contribution is made for individuals' visibility in a community (e.g., SourceForge), and the *common-good* model, value is produced for the benefit of some community (Kumaran et al., 2009). In this paper, we explore the *for-fun* model (Cooper et al., 2010; Law et al., 2007; Von Ahn & Dabbish, 2004; Von Ahn et al., 2006), in which data is a by-product of some gameplay, often referred to as "Games with a Purpose" (Von Ahn & Dabbish, 2008), which have been shown to be very successful in many domains.

Specifically with respect to generation of paraphrases or phrase equivalents, (Chen & Dolan, 2011) present their paraphrase collection using video annotations, focusing primarily on viability of establishing Mechanical Turk for providing paraphrases in a productive way. (Barzilay & McKeown, 2003) posited that multiple translations of a foreign text may be a naturally occurring source for paraphrases as each is authored by a different translator; our approach is analogous to this approach, though our source phrases/sentences are not from a foreign language. (Chklovski, 2005) presents an online paraphrase collection tool and studies the incentive model for responsible contributions by volunteers. Paraphrases generated by Doodling would be similar to paraphrases labelled under class "Phrasal" and to a lesser extent class "Elaboration" in (Chen & Dolan, 2011). In our earlier work (Kumaran et al., 2012) we focused on a proof-of-concept methodology using a Pictionary-based approach for generation of paraphrases. In this paper, we expand our concept for generating phrase equivalents in scale inexpensively, using several game and UI/UX features, and also compare it with a realistic for-pay baseline using Mechanical Turk. The power of our methodology is its self-verification mechanisms (by drawer annotating the response for convergence, and the final acceptance) that validates the generated paraphrases.

3 Doodling as a Game

In this section, we present the design elements and the game flow of the Doodling game.

3.1 Game Design

In the Doodling game, the games are played in rooms with one player (designated as the *Drawer*) sketches an assigned concept - as phrase or sentence - while other players in the game room (*Guessers*) attempt to guess the assigned concept from the drawing that is being replicated to all screens. The Guessers typically start guessing the words first (based on the concept that the Drawer starts sketching on the screen); while the game will automatically indicate exact partial matches (for example, "Taxi" as a guess for the given phrase, "Taxi Driver"), the drawer also has the ability to provide feedback using annotations. The Drawer may annotate partial guesses as incorrect (red), on the right track (orange), or partially correct (green), to guide the convergence. All the guessers' guesses and the drawer's annotation are broadcast to all the players in the room. Such broadcasting provides a mechanism in which players can build on the top of other’s guesses, gradually building up the phrase or the sentence. At some point, if one of the guessers guess the right phrase exactly, the game is closed automatically. In addition, if the drawer judges the guess as having the same meaning as the assigned concept (for example, “Cabbie” for “Taxi Driver”), he/she can end the round by marking the guess as correct, rewarding the guesser with game points. If the timer runs out before a correct guess happens, then the game times out. Figure 1 shows the UI during the progress of a game (the given text element being “*taxi driver*”).

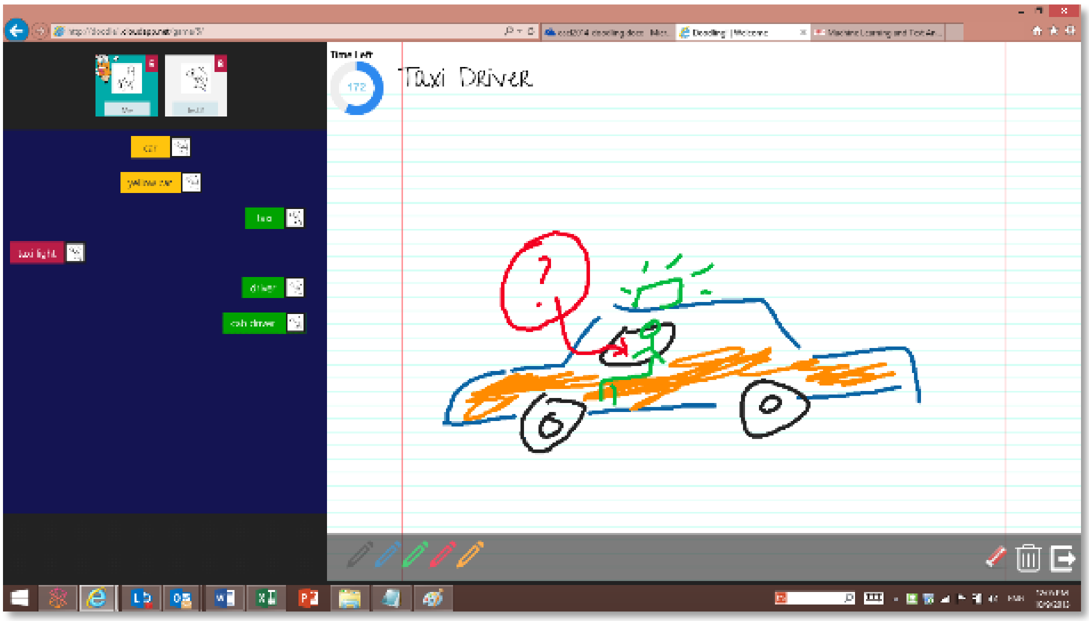

Figure 1: Doodling Game

Our primary intuition is that the sketches provide a language-independent means of communication of concepts that is effectively employed for the generation of phrase equivalents. Thus, we leverage a fun drawing-guessing game to fulfill the linguistic purpose of generating phrase-equivalents. An important aspect of making Doodling effective was to make it engaging to play. We underwent multiple user studies followed by changes to the game’s UI/UX. Earlier trials had revealed the need for additional feedback from the drawer, leading to the introduction of 3-stage annotations of guessed phrases. From a usability standpoint, the UI and gestures were optimized for use with touchscreen capable devices, including of the use of swipe gestures for annotating incoming guesses.

The Doodling game subscribes to the Inversion Problem (Von Ahn & Dabbish, 2008), where one of the players produces an output in the form of a sketch for a given input phrase. The other players attempt to guess the given input. The game may produce multiple surface forms of a single semantic intent that have a relationship similar to that of the input-output pair in the “noisy-channel” model.

3.2 Game Elements

While the game dynamics promote the resolution of the underlying computational problem (i.e., the generation of phrase equivalents), we made certain modifications to the basic *sketch-and-convey* metaphor – in the formation and constitution of the game rooms, in the assignment of roles to players in a round-robin fashion, and the drawer’s feedback using annotation, in exposing every player’s guess to the entire game room, and the winning strategy that encourages building on each other’s guesses – in order to help the rounds finish successfully, converge faster, and be more competitive. Above all, the game dynamics and the UI were designed to make Doodling enjoyable as a game.

Roles: Users may join existing game rooms, or can create a new private game room after logging in to the Doodling portal. In a game room, one of the users is assigned – randomly – the role of Drawer (D), and the others the role of Guessers (G). At the end of a given game round, the role of drawer cycles among the game room participants. All G’s both compete (the first guesser to guess right – either fully or partially – is rewarded), as well as collaborate (each builds on other’s guesses to build longer phrases for bigger rewards) in guessing the text element being conveyed by the D.

Game Round: Like the sketches, the individual guesses of a given G are broadcast to the entire room, along with any annotation from D on each of the guesses (red/orange/green). While the right guesses (either lexicographic match, or as judged by D) gives the game point to the specific G, the broadcast of guesses and feedbacks from D to the entire game room provides a transparent mechanism to help each player build on the guesses of the others. The game round closes with exact reproduction of the source phrase by one of the G's, or by D accepting a full semantic equivalent by double tapping a tile. As an incentive for the role of the drawer, the D is also rewarded with some game points.

Data: In our current experiments, we used standard phrases from a generic WikiTravel (http://wikitravel.org/en/wikitravel:phrasebook template) tourism phrase book as input elements. The authors subjectively classified each text element as Easy or Hard, depending on the potential difficulty to express it as a sketch; though such annotation implies additional preparatory work, it may be well worth the investment as such tagged corpora forms the seed for many variations. We plan to add text elements in many domains (Celebrities, Movies and Idioms), to provide diversity to the players.

Text Element	**Diff.**	**Granularity**
Cheese Omelette	Easy	Phrase
Museum of Modern Art	Hard	Phrase
I would like a bowl of soup.	Easy	Sentence
I am not feeling well!	Hard	Sentence

Table 1: Sample of text elements used in the initial seed corpus

In order to understand the dynamics of the game, and to improve the quality and quantity of the phrase equivalents generated in Doodling, we incorporated many features.

Number of Players: The application supports 2-4 players per game room, to measure the effect of room size on convergence rate and the player enjoyment. We hypothesize that those game rooms with more players will lead to better completion primarily due to higher productivity in phrase generation.

Hints & Reminders: we provided hints to all guessers at the beginning of the game to prime then on what to expect about the guess phrase. Hints are simple text elements, such as "Short Phrase" or "Hard Sentence", etc. In addition, we also provided some reminders periodically for improving the game dynamics, especially for the new players, including a reminder to the drawer that they can accept non-exact phrases with the same meaning by double-tapping on the guess tile. Reminders appear on the screen, and fade away unobtrusively. Some game rooms were provided the hints, while others are not, in order to measure how helpful the hints are for game completion.

Soft Matches: Exact lexicographic guesses (full or partial) are automatically rewarded by the game engine. However, as the primary mechanism for gathering paraphrases, soft matches were allowed and rewarded at the discretion of the drawer (either by the double-tap action that accepts a guess as a correct phrase equivalent, or by the swipe-right action that which indicates a potential partial match). Yet, to discourage collusion or cheating, a reporting mechanism is provided: The final accepted guess along with the input text element are shown to all participants, to report any unsatisfactory acceptance.

Metrics: For measuring the effectiveness of the Doodling game, we define many metrics ranging from completion statistics (completion rate and completion time), to quality by comparison with gold data (true positives as compared with user-annotated data, precision and accuracy of automatically classified data), to qualitative user feedback (fun factor).

4 Doodling: Experimental Evaluation

Doodling is an HTML5 app that is accessible from most devices - touchscreen laptop or tablets - and deployed in the cloud (http://doodle1.cloudapp.net/). After deployment, we recruited volunteers (primarily graduate students) to log in and play the game for one hour. As the volunteers entered the game server, they were assigned to different game rooms; each room was instrumented for a specific configuration (game room size - between 2 and 4 players, and availability of hints and reminders). Each room was given the same set of 38 phrases in the same sequence, to keep the variability to a minimum. After an hour, the games were closed and the players asked to fill in an online questionnaire.

In these trials, the 14 volunteers played a total of 112 games, in different game rooms. Most players had previously been exposed to the *sketch-and-convey* metaphor through Pictionary-type games.

4.1 Quality of the Generated Data

Basis for evaluation: We first extracted all of the text elements annotated as a potential match (green, orange, or winning) by the Drawer. Each of the three authors then independently classified according to the relevance of the match. The following five classes were used for annotating every annotated text element: EF (Exact Full Match), EP (Exact Partial Match), TF (True Full Match), TP (True Partial Match) or NM (Not a Match). Partial matches entailed guesses which captured some sub-element of the seed text, but not the entire meaning. We then measured inter-annotator agreement of author's annotations using a Fleiss Kappa measure (Fleiss, 1971), which stood at 0.7424, indicating substantial agreement among our annotations. Hence, we used our annotation (using majority voting for resolving any conflicts) as the gold data set for validating automatically the user generated paraphrases, in subsequent sections.

Quality of the generated data: Of the 112 games played, 98 of them completed successfully. Games were considered incomplete if the timer expired before successful completion. Of the 98 completed games, 15 of the final guesses were false positives (i.e. NM, wrong answers accepted erroneously), 42 games closed with guessers reproducing the exact text element given to the drawer (i.e. EF), and 19 games closed with Drawer correctly accepting a guess that is semantically equivalent to the given text element (i.e. TF, a true phrase-equivalent), and the remaining producing various degrees partial semantic matches (i.e. TP, true partial phrase-equivalents). The average time of completion for successfully completed games was 160 seconds.

In addition, most of the games, irrespective of whether closed correctly or not, produced partial equivalents to the given text element as intermediate guesses, thus providing valuable data for research. These include all the potential matches which were not accepted as the final answer for a game, but were marked as green or orange via the drawers' swipe-based annotation. Table 3 shows the breakdown of the gold classification of all of the potential matches.

4.2 From Game to Corpus

Once assured the quality of the generated data, we devised a methodology for automatically detecting phrase equivalents (full or partial) from the user generated data, so that the game would be able to scale without the need for human annotators to verify individual guesses. We designed a classifier for automatically validating phrase equivalents (partial or full), based only on the game meta data, and very shallow text level features, and not based on any linguistic (such as, dictionaries, thesauri, etc.) or other specialized corpora (such as, parallel or paraphrase corpora). Our basic premise is that if such a classifier can identify good paraphrases with simple features, then we will be able to identify the phrase equivalents automatically, in new domains or languages.

Our classifier uses only simple game and text-level features: hardness of the input text element (easy/medium/hard), status of completion flag and cheating flag at completion, order and time of the guess, drawer's annotation (green/orange/red), cross-game evidence, substring similarities to the input text element and orthographic overlap with the input text. First, we extracted any exact matches (EF or EP) by removing any text elements that were a substring of the original guess, leaving us with a training corpus of 122 potential phrase-equivalents. We trained the classifier using a 5-fold cross-validation this corpus. Some paraphrases thus extracted are shown in Table 2.

Source phrase	**Paraphrases extracted**
Police Officer	Policeman, Police Inspector, Police Superintend
I lost my luggage.	I need to find my bag at the lost-and-found counter, Lost-and-found luggage counter.
School Teacher	Class Teacher, Teacher teaching in school.
Railway Station	Railroad Station, Railway Platform

Table 2: Automatically Extracted Paraphrases

	Doodling	Doodling + SVM			MTurk
	Raw Corpus	Training Corpus	SVM = NM	SVM = TF|TP	Corpus
Size	234	122	73	49	92
Exact Full (EF)	42	EF and EP Data automatically removed from Corpus using String and Substring Match			0
Exact Partial (EP)	71				21

True Full (TF)	30	30	2	28	53
True Partial (TP)	11	11	5	6	13
Not a Match (NM)	81	81	66	15	5
Precision (TF+TP/Size)	17%	34%	10%	69%	72%

Table 3: Comparison of corpora produced by Doodling and MTurk to gold data. SVM numbers are an average of the results generated during the 5-fold cross-validation.

The classifier reduces the burden on expert hand-annotators, by automatically filtering out text elements that are likely to not be a match. As can be seen in Table 3, only 17% of the raw corpus constitutes useful data. Removing exact and substring matches (EF and EP) increases the precision to 34%. The usable corpus produced by the classifier (SVM=TF|TP) has a precision of 69%, with only 10% of the remaining corpus (SVM=NM) constituting false negative, or "lost" data. The overall accuracy of the classifier (% of true positives + true negatives) is 82%.

This methodology provides a viable means of generating paraphrase corpora, with a small amount of hand-crafted corpus in a new domain. The classifier can be fine-tuned either for accuracy of prediction (precision) or productivity (recall); in our experiments we fine-tuned it for precision. Also, we believe that given that these features used are devoid of linguistic or domain information, our results may provide a lower bound on the quality of automatic identification of phrase equivalents; this may be improved substantially by use of appropriate linguistic resources or specialized corpora.

In addition to phrase-equivalent data, many of the guesses relate semantically to the input text element, in varying degrees. Using similar features as used in the classifier, the annotation data can be used for identifying sets of related words for given input text elements, creating valuable resources for search query expansion.

5 Mechanical Turk Experiments

To understand the quantitative difference between Doodling and a paid crowdsourcing model for generating paraphrases we designed a "Data Collection" Mechanical Turk task using the same phrases that were used in our user experiments. Based on previous work relating to designing of Turk experiments and accepted best practices, we kept the task description simple: Each task asked a respondent to generate five unique and semantically equivalent phrases for a given source phrase. The respondents were chosen based on their familiarity with English as their first language, and each phrase was to be annotated by 20 respondents over one week duration; this duration was chosen to keep the respondent population size roughly equal to that of our user experiment. Reward for completing the generation of five phrase-equivalents for a single given phrase was fixed at $0.10USD, in line with the rewards given out for tasks with similar levels of difficulty as cited in published literature (Callison-Burch et al. 2009; Dolan et al, 2011). Though the time frame was a larger than the duration of our experiments (one hour) significantly, the overall time taken for task is comparable to the time spent in gameplay.

At the end of the one-week duration of the experiment, 14 out of 38 phrases got at least one set of valid paraphrases, leading to a completion percentage of 37%. Most of the submitted phrases were annotated only by one respondent; the average number of respondents per phrase was 1.23. The annotation data was judged by the authors in the same scale as outlined in Section 5.1, and the Fleiss Kappa measure for the annotation was 0.74, signifying significant agreement between their judgments. Overall, 72% of the MTurker generated paraphrases were accepted as full or partial alternatives (See Table 3). While the quality of data is very good, any misunderstanding of the task generated results that are significantly off the mark: For example, "*How do I get to the nearest international airport?*" was generated for "*International Airport*" as the source phrase. Since the participation and completion was low, we extended the duration of the task by another week, but the second week yielded only 2 additional completed tasks indicating that the duration of the experiment was not the sole factor in the relative low rate of task completion; perhaps it is the nature of the task that did not attract significant participation.

6 Discussions

6.1 Viability of Doodling as a Game

The 85% successful completion (98 out of 112) of the games is encouraging, and indicates the viability of the game to complete successfully. At the end of the experimental session (wherein 30 rounds of the game had been completed by each player on an average), the players were asked to fill in an online survey to measure various qualitative metrics on effectiveness of Doodling as a game. A wide variety of questions were asked, ranging from specific input (*How did [a specific feature] affect your ability to guess the right phrase?*) to generic qualitative measures (*Would you play this game again?*). Among the questions were three specific questions on how much the players enjoyed the game as a drawer, as a guesser and overall, in a scale of 1 (*Hated it*.) to 5 (*Loved it*!). From the 10 respondents, the enjoyment factor averaged at 4.7 overall. Such high score validates the game design and UX as a viable mechanism for an enjoyable game. Further, 9 out of 10 respondents said that they would *definitely* play the game again, with comments such as "It was very interesting and fun" and "This game is kind of addictive", indicating attraction of the game for subsequent engagement

6.2 Use of Hints & Reminders

We find no evidence for the hints or reminders to be valuable either in improving the quality of the result, or helping the time for convergence/completion. We note that several gamers resorted to other means of indicating the structure of the guess phrases, such as drawing out a number of dashes to indicate the size of the guess phrase, with some of them requesting us to do the same.

6.3 Scaling Up: Comparison with Mechanical Turk for crowd-sourcing phrase equivalents

GWAPs have been criticized for their complexity, long time-to-market, and hidden running costs (Wang et al., 2012). Paid crowd-sourcing methods, by comparison, are simpler to set up, and have lower initial costs. While a concrete, direct comparison is not possible, Table 3 lays out some of the differences between the two methods, especially with reference to our metrics.

	Mechanical Turk (MT)	Doodling
Experimental Operating Costs	US$82	US$90
Ongoing Costs	US$0.10/source phrase	US$90/month
Setup Costs	Minimal	3 man-months
Players/Workers	9	14
Time	2 weeks	1 hour
Completion (Games with ≥ 1 TF generated)	14/38 (37%)	38/38 (100%)
Quantity (# of Unique TFs)	53	28
Precision (% of usable data)	72%	69% (with Classifier)

Table 4: Comparison of MTurk and Doodling experiments for generation of phrase-equivalents

In the case of the Doodling game, the development of the game took 3 man-months, while Mechanical Turk's (MT) setup time was minimal. Both the Doodling and MT experiments had similar operational costs, at US$90 and US$82 respectively. This cost of $90 for Doodling consists of hosting and bandwidth charges incurred for two virtual servers running on a commercial cloud platform. However, once we scale Doodling up to permit more users and higher productivity, we expect the costs to remain fixed, whereas MT costs will scale proportionally to the productivity at US$0.10 per source. In addition, even with approximately equal investment, one hour of Doodling game play is more productive than the two weeks of MT task. As discussed in Section 5, we encountered a significant limitation of paid crowd-sourcing: workers may not choose to do tasks they consider uninteresting. While it is possible to increase the pay rate to increase the completion rate, this entails additional costs, with deteriorating completion rates. While we expect the productivity of Doodling to scale with the number of users, MT's productivity is low even for our limited experiment, and may not scale at all.

To put this in perspective, the time taken to generate useful data using Mechanical Turk varies highly depending on the task: (Chen and Dolan, 2011) reported a duration of 2 months, whereas (Callison-Burch et al., 2009) reported 2 days for their experiments. In our Doodling experiment, the task completion rate for the game (one hour, 14 players) is faster than the equivalent Mechanical Turk task (two weeks, 9 workers). We argue that for scalable data collection, a fixed recurring cost for a reliable completion rate may be preferable over a variable recurring cost. Furthermore, the Doodling game setup is easily scalable to large user base with little marginal cost, and hence we hypothesize that the economy of scale will make Doodling cheaper than MT for diverse domains. Finally, while MT workers tend to be transient, gamers tend to be loyal, particularly if the game is perceived to be interesting. Such a user base may be likely to participate and be productive in other (perhaps related) GWAPs for the generation of useful language data.

6.4 Cheating

Doodling depends on fair gameplay in order to generate reliable phrase-equivalent. Although we did not have many cases of cheating during the trials, cases of cheating will be unavoidable as the game scales to more users. The drawer scribbling answers to the canvas is a most obvious form of cheating, which may require sophisticated image recognition algorithms to weed out automatically. However, we opted for a low-cost approach of allowing any guesser to mark a certain game round as cheating, if they find the drawer scribbling on the canvas. Any guesser can also mark a game round as cheating, if he/she finds the drawer concluding a game round with guesses that are not equivalent phrases. All guessers in a room other than the guesser who provided the accepted guess, are given three seconds to report cheating in case the guess was not found as a suitable equivalent phrase. While this methodology may not work in a two player room, we expect that in larger rooms the competitive nature of the players will keep a game honest. Frequent offenders may be penalized. Proposed penalties would be banning from game rooms, disabling certain roles or introducing harder authentication protocol to prune out offending players.

Along the same lines, we intend to introduce an “inappropriate or offending” flag, to be flagged for a drawing or a guess, by any of the players in the room. Such flags, once set, may need to be investigated offline, and the players penalized in order to discourage misuse or abuse of the game environment.

7 Conclusions and Future Work

In this paper, we explored gaming as a methodology for generating paraphrase data that is useful for NLP or IR research and development of practical systems. Specifically, we outlined a *game-with-a-purpose* – Doodling – that is based on *sketch-and-convey* metaphor, where a sketch by a Drawer was used as a mechanism for abstracting a concept (the source phrase) which was then surfaced by different guessers in the game room, potentially producing paraphrases. We showed that our online multiplayer game was effective in generating paraphrase data, by mining user guesses in the familiar *sketch-and-convey* paradigm, and rewarding phrase-equivalents in addition to exact phrase guesses. Our experiments for just one hour with volunteers have shown that this game can generate high quality data in scale. Most importantly, our volunteers rated the game “very enjoyable”, even after an hour of continual play. In addition, we presented a classification mechanism to automatically identify good partial or full phrase-equivalents from the user guesses, using only the meta-level features of the game and shallow text features, opening an avenue for data generation in diverse domains, with a small seed corpora. We believe the quality of such identification may be improved significantly with addition of linguistic resources, such as, dictionaries or thesauri. Finally, our experiments with Amazon’s Mechanical Turk indicated that our game is comparable to and potentially more scalable than paid crowd-sourcing. We believe such a game may be a viable mechanism for generating paraphrase data in diverse domains and languages, cheaply.

7.1 Future Work

Currently, we are in the process of developing and releasing Doodling as a multiplayer game app, providing a potential opportunity to study its uptake in the Internet, and the quality of data generated. In our experiments we measured, through a post-game survey, the potential for Doodling being a fun game, and we obtained a score of 4.7 out of 5 for “fun-factor”, in addition to many verbal comments on how enjoyable the game was. Such user feedback amply indicate Doodling's potential for scaling well as a game in diverse domains, such as sports, entertainment and idioms. Also, while the current implementation of Doodling game works well for phrases, we have ample evidence that it works for short sentences (such as, "My luggage is lost", "Where is the nearest post office?" etc.). We hope to extend it to complex sentences as future work.

One of our goals long term is to explore the game's potential for generating parallel data – perhaps through a game being played between two players conversant in two different languages. While this multi- and cross-lingual game poses significant challenges, it provides for an interesting exploration into generation of parallel data through games. Significantly, it may also provide opportunities for language learning and/or cross-cultural awareness, as many of the idioms and culture-specific phrases are not readily conveyed by the surface forms in one language or another. If successful, this may pave way for cost-effective generation of parallel data between many languages of the world.

Reference

Barzilay, R. 2003. Information Fusion for Mutli-document summarization: Paraphrasing and Generation. *Ph.D. thesis @ Columbia University*.

Barzilay, R., and McKeown, K. 2001. Extracting paraphrases from a parallel corpus. *39th Annual Meeting of the Association for Computational Linguistics*.

Callison-Burch, C. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. *EMNLP'09*.

Callison-Burch, C., Cohn, T., and Lapata, M. 2008. ParaMetric: An Automatic Evaluation Metric for Paraphrasing, International Conference on Computational Linguistics, 2008.

Chen, D.L. and Dolan, W. 2011. Collecting highly parallel data for paraphrase evaluation. *49th Annual Meeting of the Association for Computational Linguistics*.

Lin, D., and Pantel, P. DIRT - Discovery of inference rules from text. *Proceedings of the seventh ACM SIGKDD* International conference on Knowledge discovery and data mining. ACM, 2001.

Chklovski, T. Collecting paraphrase corpora from volunteer contributors. *Proceedings of the 3rd K-CAP*. International conference on Knowledge Capture, ACM, 2005.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fey, A., Baker, D., Popovic, Z. and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature (466), Aug 2010*.

Dolan, W., Quirk, C., and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. *20th International Conference on Computational Linguistics*.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378–382.

Ibrahim, A., Katz, B., and Lin, J. 2003. Extracting structural paraphrases from aligned monolingual corpora. *Second International Workshop on Paraphrasing (collocated with ACL 2003)*.

Irvine, A. and Klementiev, A. 2010. Using Mechanical Turk to annotate lexicons for less commonly used languages. *NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.*

Kumaran, A., Jauhar, S. K., and Basu, S. 2012. Doodling: A Gaming Paradigm for Generating Language Data. *Human Computation Workshop 2012*.

Kumaran, A., Saravanan, K., Datha, N., Ashok, B. and Dendi, V. 2009. WikiBABEL: a wiki-style platform for creation of parallel data. *ACL-IJCNLP 2009*.

Law, E.L.M., Von Ahn, L., Dannenberg, R. B. and Crawford, M. 2007. Tagatune: A game for music and sound annotation. *ISMIR'07*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. 2002. Bleu: A method for automatic evaluation of machine translation. *40th Annual Meeting of the Association for Computational Linguistics*.

Quirk, C., Brockett, C., and Dolan, W. 2004. Monolingual machine translation for paraphrase generation. *Empirical Methods in Natural Language Processing (EMNLP-2004)*.

Von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. *CHI'04*.

Von Ahn, L., Kedia, M. and Blum, M. 2006. Verbosity: a game for collecting common-sense facts. *CHI'06*.

Von Ahn, L. and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM, Vol 51*.

Wang, A., Hoang, C. D. V. and Kan, M. 2012. Perspectives on crowdsourcing annotations for natural language processing. Language Resources & Evaluation Conference, 2012.