



People Watcher: A Game for Eliciting Human-Transcribed Data for Automated Directory Assistance

Tim Paek, Yun-Cheng Ju, Christopher Meek

Microsoft Research, One Microsoft Way, Redmond, WA 98052 USA

{timpae|yuncj|meek}@microsoft.com

Abstract

Automated Directory Assistance (ADA) allows users to request telephone or address information of residential and business listings using speech recognition. Because callers often express listings differently than how they are registered in the directory, ADA systems require transcriptions of alternative phrasings for directory listings as training data, which can be costly to acquire. As such, a framework in which data can be contributed voluntarily by large numbers of Internet users has tremendous value. In this paper, we introduce *People Watcher*, a computer game that elicits transcribed, alternative user phrasings for directory listings while at the same time entertaining players. Data generated from the game not only overlapped actual audio transcriptions, but resulted in a statistically significant 15% relative reduction in semantic error rate when utilized for ADA. Furthermore, semantic accuracy was not statistically different than using the actual audio transcriptions.

Index Terms: game, automated directory assistance

1. Introduction

Automated Directory Assistance (ADA) allows users to request telephone or address information of residential and business listings using speech recognition [5][6][7]. The main objective of ADA is to reduce the need and cost for human operators who have traditionally provided this service. Automating this task, however, has proven to be quite challenging because callers frequently express listings differently than how they are registered in the directory [6][7]. For example, the listing *Kung Ho Cuisine of China* can be expressed in abbreviated form (e.g., *Kung Ho*), or include words not in the registered title (e.g., *Kung Ho Chinese Restaurant*), or be pronounced in a different way (e.g., *Hung Ho*). In order to deal with user variations, ADA systems require transcriptions of alternative phrasings for the listings as training data. Unfortunately, there are over 18 million listings in the US Yellow Pages alone. Collecting and transcribing data of that scale would involve enormous effort and cost. As such, a framework in which data can be contributed voluntarily by large numbers of users on the Internet for different local areas has tremendous value.

This paper concerns the problem of obtaining human-generated, alternative phrasings for proper nouns. To address this problem with respect to ADA, we introduce *People Watcher*, a computer game that elicits transcribed, alternative expressions for places; in particular, businesses listed in the telephone directory. The paper is organized as follows. After discussing related research in Section 2, we describe the game format, rationale, and uses in Section 3. In Section 4, we evaluate the usefulness of game by comparing data generated from a trial deployment to actual audio transcriptions and by

experimentally validating whether the data can improve performance on ADA.

2. Related Research

The problem of obtaining human-generated, alternative phrasings for proper nouns is not unique to ADA. For example, in natural language understanding, it is important to know that multiple proper names can refer to the same person (e.g., *Bush*, *George W.*, *Mr. President*). However, because the primary task of ADA is to ultimately find a unique residential or business listing, alternative phrasings pose a serious problem to ADA usability. As such, researchers have pursued various methods, all of which require data that our game could provide. We survey a few here.

Recent research on ADA systems has focused on a search approach in which recognized text is used to match against the set of business listings [7]. This approach uses n-gram language models which, like dictation models, compress and generalize across listings and their observed expressions. Whether alternative expressions are included in the grammar or used to train n-grams, our game could provide transcribed training data.

Another method for dealing with alternative phrasings is to generate user expressions using transduction rules applied to directory listings [9]. Again, those rules must be induced from initial training data. Furthermore, validating that the generated expressions are indeed possible user expressions is a problem, which our game circumvents by using human users to generate the data.

The problem of obtaining alternative phrasings for proper nouns can be viewed as part of a larger problem of leveraging human computation to address tasks that cannot be easily automated. One way to leverage human computation is to utilize a software platform like *Mechanical Turk* to programmatically access and incorporate paid human intelligence into an application [1]. Another way is to exchange human computation for entertainment in the context of a computer game. In fact, since the introduction of *The ESP Game* [11], researchers have sought to use games to tackle machine learning problems such as image classification [11], generating natural language descriptions for images [12], and paraphrasing journalistic sentences for machine translation [2]. To our knowledge, no games to date have been utilized for speech recognition problems.

3. Game

In order to address the problem of obtaining alternative phrasings for proper nouns, and in this case, official business listings, we sought to leverage human computation and transcription with a computer game. In this section, we describe the format of the game, the rationale behind the design, and how it can be used for ADA.

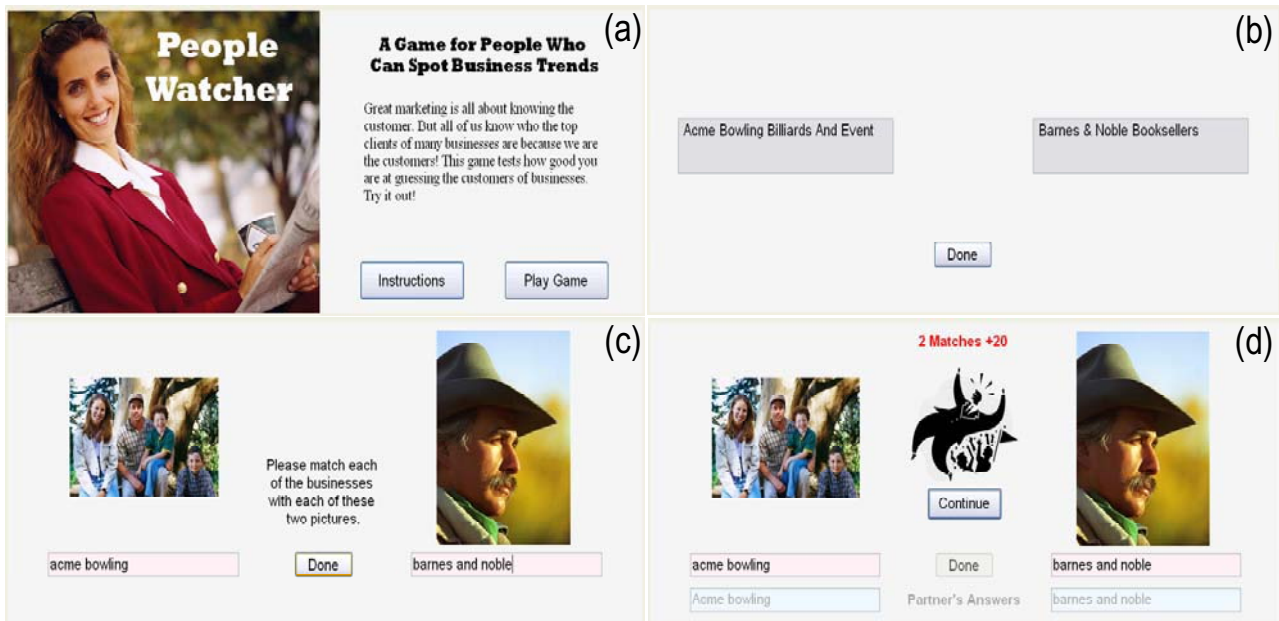


Figure 1(a-d). Four screenshots of the People Watcher game showing the process of how a player matches business listings to images.

3.1. Format

People Watcher is a multi-player game in which players are randomly paired with one or more partners on the Internet. One of the partners can also be a “bot” that emulates a live player by using previously recorded user expressions. The game is framed as a marketing game that tests people’s ability to “spot social trends” by having them identify who they think would be likely customers of various businesses. The customers are shown in photographs of all kinds of people – from individuals to families, bakers to zoo keepers, Americans to New Zealanders. The businesses are directory listings submitted as input to the game. Figure 1(a) shows the message about marketing that players receive when they first start the game. In order to make the game more relevant to the players and the data more targeted, players can be paired with people of the same local area and use business listings for that local area as input.

The game proceeds as follows. Once players have logged in, they are paired with a random partner and shown two business listings as in Figure 1(b). When they are ready, they click the ‘Done’ button. Then they are shown two photographs of people as in Figure 1(c). These photographs vary along different attributes such as group, occupation and nationality. Photographs are paired so that they differ in at least one attribute. The player’s job is to match each of the photographs with one or the other of the two businesses they saw. They do this by typing in their guesses underneath the photographs. Finally, they are shown their partners’ answers as in Figure 1(d). Players receive points for every matching answer. To be added to the list of the top players, they need to garner as many points as possible in the least amount of time.

3.2. Rationale

The design of People Watcher is premised on the psychological notion that in free *recall* of text, adults are likely to reproduce the *gist*, or “essence” of the text instead of a verbatim reproduction [3]. The gist can be propositional in

nature, as seen for example, in the way that people sometimes add the word *restaurant* to a business listing. By showing and removing the business names from the screen, players must commit them to memory. Furthermore, because the game is timed, players are encouraged to provide short names to identify the businesses. It is important to note that in the instructions players are told that they do not have to write the verbatim name of the business, but enough so that we can compare their answers to their partners’.

Another important consideration is the social aspect of the game. To succeed, a player must not only guess which businesses are likely to have the people shown in the photographs as customers, but they must also match their guesses with their partners’. Similar to *The ESP Game*, this kind of collaboration can motivate players to provide reasonable answers (i.e., it takes two or more people to win) and to play multiple times [11]. By keeping track of the top players, we also add the desire to achieve reputation in the game as motivation.

3.3. Uses

The game can be used for ADA in many ways. First, the game can be used to obtain alternative phrasings for all business listings. For example, in our trial deployment, in one session which lasted about 10 minutes, a player was able to provide alternative phrasings for 30 business listings. Assuming this rate holds on the Internet, if we have at any given time 100 concurrent players, it would take just 42 days to generate alternative phrasing for all 18 million listings in the U.S. Yellow Pages. Second, even if it never achieves widespread popularity, the game can be used for active learning of ADA systems [4] by inputting, for example, those listings whose recognitions have had low confidence scores [10], or listings with high call interest. Finally, the game can also be used as a low-cost replacement for audio transcription. Because transcription of audio data can be boring and prone to error, particularly when it is done by hearing only without a reference listing, it may be cheaper to pay someone to play the game than it is to hire a professional transcription service,

especially if the data generated by the game yields similar performance as compared with actual audio transcriptions.

The game can also be used for other purposes as well. As discussed earlier, all the photographs used in the game contain attributes that identify the people presented, such as occupation and nationality. These attributes can be used to generate a user model that can be used to set priors on likely businesses. For example, the data may identify restaurants people with families are likely to patronize, which can then be used as priors. The data could also be used to help businesses understand public perception of their clientele. For example, a local restaurant might be able to find other businesses with a similar actual or perceived clientele which might be valuable in advertising scenarios.

4. Evaluation

Although the game generates alternative phrasings for business listings, it does so within the context of a “marketing” game where players recall listings while matching them with photographs of people. In order to assess whether this data can be useful for the ADA task, in this section, we describe the results of a trial deployment.

4.1. Trial deployment

In anticipation of evaluating the usefulness of the game, before deploying it, we sought data from a prototype ADA system described in [13]. Because the system had only been deployed internally at Microsoft for 3 months, we initially did not have any audio transcriptions. We transcribed those cases where the user tried multiple times to find a business listing and was eventually satisfied with a returned result. We call these cases *long successes*. As suspected, many long successes did in fact contain alternative phrasings, though many were also due to noise. In all, we obtained audio transcriptions for 58 business listings. We then split those listings to create 29 pairs of businesses and submitted them to People Watcher as input. Because organizing a time in which people can simultaneously play with each other was logistically difficult, we deployed a single-player version of People Watcher where a player’s answers were compared to those of someone who had previously played the game. For the trial deployment, 22 participants, all employees of Microsoft, agreed to play in exchange for a latte coupon. The game lasted about 10 minutes, including the self-paced instructions.

Before comparing the human-transcribed alternative phrasings from the game to actual audio transcriptions, it is important to note that the data did require cleaning. In the game, we indicate that players do not have to use the exact name of the business listing in matching businesses with photographs. One consequence of this is that users sometimes made spelling errors. This was easy to fix. A more insidious consequence was that users sometimes put just enough words to distinguish one business from its pair but not from all other businesses. For example, if one of the businesses was a restaurant and the other was not, some users simply wrote *restaurant*. This is a design flaw that needs to be fixed in future versions.

In cleaning up the game transcriptions, we noticed a few interesting trends in the way that people remembered business listings. First, restaurant listings had more alternative phrasings than any other business type, perhaps because of the considerable variability in restaurant names. Second, people often used the possessive form to express any part of

	Directory Listings	Audio Transcriptions	Game Transcripts
% of Directory Listings That Overlaps with	100%	25.9%	100.0%
% of Audio Transcriptions That Overlaps with	13.8%	100%	36.1%
% of Game Transcripts That Overlaps with	29.9%	64.2%	100%

Table 1. Percentage of overlap between the directory listings, audio transcriptions and game transcriptions with each other the listing that seemed like a proper name (e.g., *Keg’s* for *Keg Steakhouse*).

4.2. Overlap with audio transcriptions

After cleaning up the game transcriptions, the first question we sought to answer was how the alternative phrasings generated from the game compares to actual audio transcriptions. For the comparison, we had 58 directory listings, 109 distinct audio transcriptions and 194 distinct game transcriptions. By “distinct,” we mean a unique alternative phrasing. The audio transcriptions had on average 1.9 alternative phrasings per listing, whereas the game had on average 3.3 alternative phrasings. Unfortunately, because our audio transcriptions were limited to the long successes, it is difficult to tell if this average difference would be significant across all listings. As a representative example of the kind of data we obtained, the distinct alternative phrasings we obtained for the listing *Pabla Indian Cuisine Sweet* were *Pabla*, *Pabla Indian*, *Pabla Indian Restaurant*, *Pabla Restaurant* and *Pabla Indian Cuisine Sweet*.

Table 1 displays the percentages of matches between the directory listings, audio transcriptions and game transcriptions with each other for the long successes. 100% of the directory listings matched the game transcriptions, but this may have been due to the first few participants who had cut-and-pasted the business listings. They did not claim to have used it exclusively and the feature was subsequently disabled. Including this data would not of course help ADA performance. In contrast, only 14% of the audio transcriptions matched the listings in the directory, indicating that people rarely used the full business names. However, this is likely to reflect sampling bias, given that we only transcribed long successes. On the other hand, within the long successes, 64% of all game transcriptions matched the audio transcriptions. In other words, more than half of the data generated from the game overlapped the transcriptions of the long successes, which seemed like a promising start.

4.3. ADA Experiment

In order to evaluate whether the game transcriptions could be useful for improving ADA, we conducted a recognition experiment where we compared the performance of a prototype voice search ADA system [13] before and after training its n-gram language model with the game transcriptions. Because we did not have enough data from the game transcriptions to have reliable priors, to ensure a fair comparison, we made sure that in training the n-gram language models on alternative phrasings, the prior probabilities for the business listings remained the same. Indeed, an interesting avenue for future research is to see if the game can generate useful priors with enough data.

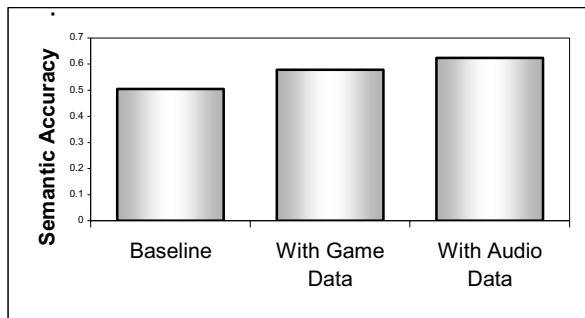


Figure 2. Semantic accuracy of an ADA system before training, and after training on the game transcriptions or on the audio transcriptions.

Figure 2 displays the semantic accuracies achieved by the prototype system on the 109 user utterances that made up the audio transcriptions. The ‘Baseline’ is the prototype ADA system without any training data. The semantic accuracy of the ADA system using the audio transcriptions was 62%. As shown in the Figure, utilizing the game transcriptions for training resulted in a 15% relative reduction in semantic error rate, increasing accuracy from 50% to 58%. This accuracy difference was significant by McNemar’s test ($p < .05$). Interestingly, the accuracy difference between using the game data and the audio transcriptions was not statistically significant by McNemar’s test ($p = .68$). Because the recognition test was performed on the utterances which constituted the audio transcriptions, this means that the semantic accuracy using the game data was not statistically different from what is essentially the upper bound.

In summary, data generated from a game which elicits transcribed alternative phrasings from users in the context of a marketing game can indeed be used to improve ADA performance.

4.4. Questionnaire

In addition to playing the game, we had participants fill out a post-hoc questionnaire. On a 5-point Likert scale, where ‘5’ is very fun and ‘1’ and no fun at all, the average user rating was 2.5 ± 1.1 . In short, they rated it as neither fun nor not fun. As for difficulty, on a Likert scale where ‘5’ is very difficult and ‘1’ is not difficult at all, the average difficulty rating was 2.4 ± 1.0 , which again is right in the middle. In addition, on a zero-sum scale, users were asked if having a live partner would enhance the fun of the game (‘1’), detract from the game (‘-1’), or not affect the fun at all. The average user rating was 0.4 ± 0.7 , indicating that people prefer to do this in a multi-player setting. Finally, we gave participants an open-ended question to tell us how they thought we could enhance game play. The majority of people provided an answer that pointed to social aspects of the game, such as being able to discuss with their partners why they made their choices.

5. Conclusions & Future Directions

In this paper, we addressed the problem of obtaining transcribed, alternative phrasings for proper nouns by introducing a game that leverages human computation in exchange for entertainment. In particular, we introduced People Watcher, a game for generating transcribed corpus data for ADA. People Watcher is framed as a “marketing” game where players recall business listings while matching them with photographs of people who they think might be

likely customers. Because players type in their guesses, the game provides human-transcribed alternative expressions for whatever business listings are submitted as input. Results from a recognition experiment indicate that using the game data can indeed improve ADA performance. In our experiments, it led to a 15% relative reduction in semantic error rate. Furthermore, its performance was not statistically different from using audio transcriptions, which constituted an upper bound for our experiment.

We have been exploring several extensions to this research. First, instead of typing in business listings, we are exploring having users speak their answers to a telephony system. This would provide audio data that could be used for training of the pronunciation model for both recognition and text-to-speech (TTS). TTS is a problem for ADA systems because even when the system finds the correct listing, users may not recognize that it did if the TTS pronunciation is difficult to understand. Second, we are exploring how to design the game so that it is more “fun” based on social game design principles [8], especially considering the feedback we received from the questionnaire. Finally, we are considering how best to utilize the game in an active learning approach to generate training data for ADA systems.

6. References

- [1] Barr, J. & Cabrera, L. (2006). “AI Gets a Brain”, *ACM Queue*, 4(4): 24-29.
- [2] Brockett, C. & Dolan, W. (2005). “Echo Chamber: A Game for Eliciting a Colloquial Paraphrase Corpus”, in *Proc. AAAI 2005 Spring Symposium, Knowledge Collection from Volunteer Contributors (KVC05)*.
- [3] Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics*. New York: Harcourt Brace Jovanovich.
- [4] Hakkani-Tur, D., Riccardi, G. & Gorin, A. (2002). “Active Learning for Automatic Speech Recognition”, in *Proc. ICASSP*.
- [5] Kellner, A., Rueber, B., & Schramm, H. (1998). “Using Combined Decisions and Confidence Measures for Name Recognition in Automatic Directory Assistance systems”, in *Proc. ICSLP*, pp. 2859-2862.
- [6] Levin, E. & Mane, A.M. (2005). “Voice User Interface Design for Automated Directory Assistance”, in *Proc. Interspeech*, pp. 2509-2512.
- [7] Natarajan, P., Prasad, R., Schwartz, R., & Makhoul, J., (2002). “A Scalable Architecture for Directory Assistance Automation”, in *Proc. ICASSP*, pp. 21-24.
- [8] Salen, K. & Zimmerman, E. 2004. *Rules of Play: Game Design Fundamentals*. Cambridge, MA: The MIT Press.
- [9] Scharenborg, O., Sturm, J., & Boves, L. (2001). “Business Listings in Automatic Directory Assistance”, in *Proc. Eurospeech*, pp. 2381-2384.
- [10] Tur, G., Rahim, M. & Hakkani-Tur, D. (2003). “Active Labeling for Spoken Language Understanding”, in *Proc. Eurospeech*.
- [11] von Ahn, L. & Dabbish, L. (2004). “Labeling Images with a Computer Game”, In *Proc. CHI*, pp. 319-326.
- [12] von Ahn, L. Ginosar, S., Kedia, M., Liu, R. & Blum, M. (2006). “Improving Accessibility of the Web with a Computer Game”, In *Proc. CHI*.
- [13] Yu, D., Ju, Y., Wang, Y., Zweig G. & Acero, A. (2007). “Automated Directory Assistance System – from Theory to Practice”, in *Proc. Interspeech*.