A Game-based Evaluation Method for Subjective Tasks Involving Text Content Analysis

Keun Chan Park, Jihee Ryu, Kyung-min Kim and Sung Hyon Myaeng
Department of Computer Science
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
{keunchan, zzihee5, kimdarwin, myaeng}@kaist.ac.kr

ABSTRACT
Standard test collections have remarkably supported the growth of research fields by allowing direct comparisons among algorithms. However test collections only exist in popular research areas. Moreover constructing a test collection needs huge amount of time, cost and human labor. On that account, many research fields including newly emerging areas are evaluating their result by manually constructed test sets. However test sets are unreliable because they often use small number of raters. It is even more unreliable when the task is subjective. We define subjective task as a task where the judgment may differ from individuals due to various aspects such as preference and interest but still preserving a sense of commonality. We address the problem of evaluating subjective task using a computer game. Playing the game, as a side effect, performs subjective task and utilizing the piles of game result lead to an objective evaluation. Our result outperforms the baseline significantly in terms of efficiency and show that evaluating through our approach is nearly the same evaluating with a gold standard.

Categories and Subject Descriptors
H.1.2 [**Information Systems**]: User/Machine Systems – *human information processing*. H.3.4 **[Information Systems**]: Systems and Software – *performance evaluation (efficiency and effectiveness).*

General Terms
Algorithms, Experimentation.

Keywords
Evaluation, Online Game, Contextual Advertising.

1. INTRODUCTION
Research fields in computer science, such as information retrieval, natural language processing and data mining, have benefited significantly from the availability of standard test collections which allow performance comparisons between systems and algorithms. However, constructing a test collection requires a huge amount of resources. Because of the heavy cost, the scope or coverage of test collections may be limited or needs to be incremented gradually (e.g., ad hoc retrieval in TREC).

On the other hand, the growth of some research fields has been hampered with the lack of standard test collections. This is particularly true for newly emerging fields. Without test collections it is difficult to make an objective comparison among different systems and algorithms. Following the norm and expectation of the field for objective evaluations, researchers attempt to evaluate their ideas with a small set of test data or user studies. Often times, however, the data are too ad hoc and/or a small-scale to warrant repeatability. Consequently a new research result for new functionality, for which no reliable test collections exist, is not likely to be published in a major conference or journal. Reviewers tend to discount the new effort and idea without solid experimental validation.

While small-scale experiments with ad hoc test data need to be more tolerated for new tasks at least at an early stage, they do have well-known limitations. First, the volume of the data is too small a sample to represent the reality. Second, the quality of the test data is often doubtful because it is very likely that the human resources for the judgments were scarce. Especially when the task calls for a subjective judgment, reliability becomes critical. For example, determining the ground truth for a summarization task is not easy because one can summarize a piece of text from different perspectives, for different purposes, and with different levels of abstraction (and hence lengths). Another example, which is even worse, is evaluation of contextual ad searching. The task of judging whether or not an ad is appropriate for a given document (e.g., news article or a web page) is heavily dependent on individuals who tend to have their own interests, preferences, and needs. Like as not, building a test collection for ad searching by hiring a few judges would be heavily biased.

We define the term "subjective task" as a task where the judgments of individuals may vary depending on their backgrounds, interests, preferences, and interests, but still tend to converge with a large number of people. Making a judgment about relevance of a document for a given topic or appropriateness of an ad for a news article belongs to this task. While ads on fresh vegetables and organic restaurants would be relevant for an article about healthy foods, for example, an ad on fat-free chicken breast may not be appropriate, especially for those whose religion forbids them from eating any animal meat. Relevance judgments in traditional information retrieval also have subjective aspects [15]. A task such as choosing one's favorite color or choosing favorite ice cream from an ice cream shop is not considered a subjective task in our definition because no answer would converge as the number of judgments increase.

Because of high likelihood of individual variations, building a test collection for a subject task by hiring a few judges is error-prone. A viable way to evaluate a system or method for such a task is to conduct a field test in which a large number of potential users can participate. But this method is not only costly but also difficult, if not infeasible, for academic research. Inability to conduct a sound

evaluation of a new method for a new subject task would discourage new innovations from being pursued and presented in a scientific way.

In this paper, we propose an evaluation method for subjective tasks for which no reliable, large-scale test collections have been made available and no field test can be conducted or is economically viable. Our approach is to use the human computation paradigm [2], which has been introduced as a way to utilize distributed human brain power for solving problems that computer cannot solve yet. Treating human brains as processors in a distributed system, the method induces a person to perform a small part of a massive computation. Since humans require some incentive to become part of a collective computation, unlike computer processors, an online game is an attractive method for elating people to participate in the process. We feel that human computing is an appropriate way to alleviate the problem of evaluations for subjective tasks in that it essentially simulates a field test without explicit cost for deploying a system and hiring people. As long as an online game is devised in such a way that it is played by a sufficient number of people, it is an ultimate testing method because the quality of the final system will be evaluated by the actual users in the end.

A key element of human computation is that the underlying task (evaluation of a system for a subjective task in our case) is hidden and performed indirectly by the game-playing users. As they play the game, the hidden purpose is accomplished. Our hypothesis is that by gathering the results of games played by unidentified users, we can evaluate the underlying system (e.g., ad searching system) with reasonable accuracy. The online game we propose is casual and easy for anyone to play without much burden. Recall that making annotations or judgments for subjective tasks are very much burdensome cognitively and can be biased by the small number of judges. In this case, anybody can access the game online from anywhere as long as they are connected by Internet. The sheer number and diversity of the participants can minimize the artificial factors caused by the traditional annotation efforts.

In order to show the feasibility of the proposed approach, we implemented an online multi-player game called mADtch whose objective is to accumulate points by selecting the ads other people would choose for a given news article. When played, it accumulates the results that are used to evaluate one or more underlying contextual ad searching systems. The game asks the users to drag and drop ads into appropriate categories given a news article. The players are rewarded when their decisions match others'. Note that while the side effect of the game playing is evaluation, the users' goal is to maximize their scores by correctly identifying what other would consider relevant ads, not just based on their own preferences. Besides, they are not in working mode but read news articles at their leisure and play a game naturally.

Our contribution lies in: 1) exploring a new evaluation paradigm using human computation, which would save much of the efforts and resources required for constructing test collections or conducting field tests, 2) suggesting a method that alleviates annotator bias problems with subjective evaluations, and 3) demonstrating the feasibility of the proposed method.

The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 describes mADtch and details of the evaluation method. In section 4, we present results with experimentations. In section 5, we evaluate our method with other evaluation schemes. Finally we conclude with future work in section 6.

2. RELATED WORK

2.1 Test Collection

The notion of a reusable test collection is central to modern information retrieval (IR) research, dating back to the Cranfield experiments [9]. A test collection consists of a set of documents, a set of topics, and a set of relevance judgments. A topic represents a formalized information need while the relevance judgments specify the set of documents within the collection that satisfy the information need, as assessed by the person issuing the request. Relevance judgments require most of the human efforts as the number of judgments to be made is equal to the Cartesian product of the queries and documents contained in the collection.

In order to reduce the prohibited amount of human judgments for a large-scale text collection, the pooling method [11] has been exploited in TREC and other subsequent efforts like NTCIR and CLEF by exploiting a number of participating systems that can return their own search results. While the number of judgments can be drastically reduced and the sample is reasonable enough to reliably rank-order various algorithms and systems for retrieval effectiveness, the amount of human efforts is still massive when all the efforts are added up for the entire collection built across multiple years.

To reduce the number of required judgments, the Interactive Searching and Judging method [10] was introduced. It was shown that the number of required judgments is reduced to less than one-quarter than that of using the original pooling method, using TREC-6 data. However, they could not reduce the cost for hiring, educating, and supervising judges, which should always take place when one constructs a new test collection.

Test collections for other subjective tasks have been developed along with those constructed for IR. Some are in a small scale as in for genre classification, for sentiment analysis, and for information distillation, for protein-to-protein interaction, just to name a few. Others are in a relatively large scale as in the Question Answering (QA) Track in TREC [16]. An additional difficulty in finding answers is that it is difficult to determine the correct size of the answer string even for factoid questions. The problem becomes even more difficult when two systems return answers for a description question for which an answer could be of arbitrary length with a great variability in sentential structures and vocabulary. There was an effort to build a test collection for summarization tasks but as mentioned in Introduction, it is difficult to compare system results with a gold standard. To the best of our knowledge, there has been no attempt to construct a reusable test collection for contextual ad searching tasks.

2.2 Contextual Ad Searching

Contextual ad searching or context match (CM) refers to the placement of commercial textual advertisements within the content of a generic web page. Since relevance of an advertisement can be affected by the article subscriber's background knowledge, preference, or confronted situation, it is hard to evaluate a contextual advertisement system. In general, there are three methods being used widely to evaluate contextual advertisement systems.

Firstly, A/B testing is a method to apply a new system to a small proportion of existing system. This method is mostly used by big

companies with their own systems. When it is applied to a contextual advertisement system, one can evaluate a new system by observing how CTR (Click-Through Rate) or CPR (Cost per Click) changes. There are two disadvantages of this method. The first one is that only the organizations or research groups which have the commercial system can use this method. Another one is that there is a risk of damage on business if there is a defect on the new system.

The second method is to use retrospective data collected from a large contextual advertising system. Chakrabarti et al. [8] used data collected from Yahoo! System during 15 days to evaluate CTR prediction system. Since the evaluation process is done off-line, one can evade from the risk of damaging business when using this method. However, one cannot evaluate the performance of system for new advertisements which were not available for the past data. Generally, since the data itself is much valuable, it is rarely shared publically. Research groups with no access to this kind of data have to find out other methods for evaluation of their ideas and methods.

The last method is to annotate page-ad pair on a discrete scale (usually binary or 3 ~ 5 scales) with human judges' efforts on a relatively small size collection. The created gold standards are used as a source of calculating evaluation metrics like precision-recall or rank correlation (e.g., Kendall's τ). Broder et al. [7] judged page-ad pairs by three or more human judges on a 1 to 3 scale: relevant, somewhat relevant, and irrelevant. In their experiment, the inter-annotator agreement among judgments was 84%. Such existence of disagreements is due to the subjective characteristics of the task. Averaging results from a few judges is risky since they can be biased by the judges' generational or cultural inclinations. Moreover, it is costly to employ judges. The cost includes money, time, and human labors to hire, educate, and supervise the judges.

2.3 Human Computation

Human computation is used in tasks where it is trivial for humans but still challenges even the most sophisticated computer. So far, human computation has focused on data annotation and knowledge acquisition. In [3], [12], and [14], the authors attempted to collect common sense using a computer game. The work in [1] and [4], renowned as the Google image labeling game, was purposed to annotate images.

There have been attempts to enhance the quality of optical character recognition (OCR) using human efforts. Captcha [5] is the representative security tool asking the user to input poorly OCR-ed characters when making an account on a website. The typed results are used to enhance and correct the recognition. TypeAttack[1] is another game where players are asked to type in poorly OCR-ed characters. However, as far as we know, there is no work using human computation for relevance evaluation.

Crowdsourcing [6] is also somehow related to our work. Crowdsourcing aims to outsource tedious tasks (e.g., finding relevant page given a key word) to the netizen by rewarding them money for accomplishing the task (e.g., Amazon's Mechanical Turk[2]). On the other hand, human computation games motivates user because it is fun (no financial rewards). Also human computation games hide their goal (e.g., to evaluate ad system) whereas the purpose of crowdsourcing is directly shown to the user.

3. PROPOSED SCHEME

3.1 Game Description

Our proposed game mADtch is a casual online multi-player game. The title "mADtch" is formed from the two words "match" and "ad". The current implementation of mADtch only provides two modes: single player and two players. However, we designed our scheme considering more than two players for generalization and further expansion. Despite the game is limited to two players, we describe our scheme considering multi-player situation involving more than two players for the remainder of the paper.

When a player logs in, the game randomly groups a number of players waiting and lets them play a game. For a particular run of the game, all players whose real identities are not known see the same news article with same sets of ads. Ads come from one or more of the underlying contextual ad searching systems that we aim to evaluate. Players are asked to drag and drop each of the ads into its appropriate category the player thinks. In order to reduce selection variations caused by individual differences, they are instructed to choose ads that are likely to be selected by other players. We provide three categories: relevant, somehow relevant in some sense (ambiguous) and irrelevant, which are the labels used in contextual ad searching [7]. When a selected ad category is matched with that selected by the majority of the players, the player gets points.

If the players want to obtain points, they need to consider not only their own understanding and preference, but also the judgment of others. The player has to think about the commonality of the task. This output agreement game partially verifies that the output is correct, since identical output among many players mean that the selection is not erratic. At the same time, by trying to match the way other players think, the players can consider this game an enjoyable social interaction. One round is for 5 minutes, and the players must press the finish button to end the round. When the game is over, the server updates accumulated scores, skill levels, the number of played games, and the list of seen articles in the user profile.

3.2 Game Design

Our game is designed based on the "games with a purpose (GWAP)" guidelines [2]. The guidelines introduce enjoyment factors, apparatus to make the game result accurate (e.g., prevent cheating) and evaluation measures for GWAP.

To increase player enjoyment, we adapted timed response (i.e., setting time limits), score keeping, player skill levels and randomness (i.e., random news articles).

To ensure output correctness and prevent player collusion, we applied random player matching, player testing, repetition and heuristic rules. Random player matching prevents cheating with a known partner. Player testing is a way to prevent abnormal single players. We test single players with test data (i.e., article and ad set with correct answer we already know). If his game result doesn't exceed a certain threshold, the game ignores the player's result and regards it as a noise.

[1] http://apps.facebook.com/typeattack

[2] http://www.mturk.com

Figure 1. Game screen of mADtch

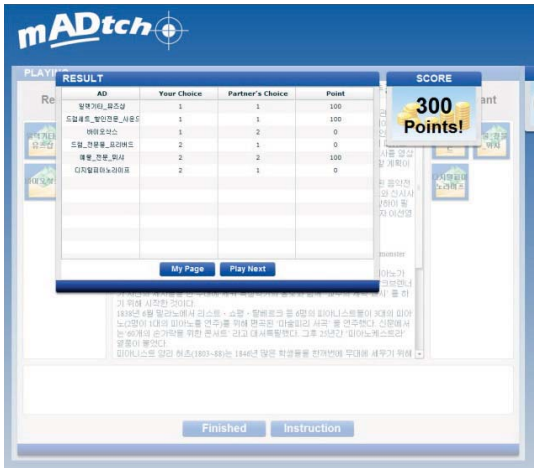

Figure 2. Result screen of mADtch

Though we make sure the player does not see the article that he/she already has seen, we make other players (who haven't seen the article) play the game until a certain number of game results have been gathered. This repetition has two advantages. The more results we get for a single article, the more reliable the evaluation would be with attenuation of possible noises. We also provide several heuristic rules such as games finished within 10 seconds and games that have exceeded the time limit are not counted.

The former assumes that players cannot comprehend the article and choose a relevant ad within 10 seconds. The latter is based on assumption that after five minutes, the player is not on the game anymore. In either case, the player is instructed to read the same article again and play the game. Any result with no ad moved to the relevant or irrelevant area is not stored as a result. When there is no player waiting, the current player can still play a game with the old data in the database.

Figure 1 shows a snap shot of the game screen. The main panel shows the actual game being played by the user. On the right is the timer used to indicate how much is left before the game is over. A small envelope shape contains the title of an ad whose description appears when the mouse is over the shape. When the game begins, the article and a list of ads at the bottom, which is the somehow relevant category, are shown to the player. On the left and right of the article are the areas to which relevant and irrelevant ads should be moved, respectively.

Table 1. Description of the symbols in equation (1)

Symbol	Meaning
C	the category of the ad: 1, 0, or -1
U	all users who played for the document
U_C	the users who selected the category C for the ad
O_u	the order by which the user u have selected the ad
T	the whole play time for the document
t	the time taken to agree on the dominant category

```
<game id= "000001">
  <article id="joins-001"/>
  <users id="user01"/>
  <ads>
    <ad id="1" label="1" time="10" order="1"/>
                    ⋮
    <ad id="12" label="-1" time="20" order="7"/>
  </ads>
</game>
```

Figure 3. Game result of user

The bottom box is where the ads that have not been moved to one of the boxes remain. We do not show more than 12 ads per game, because too many ads might be a burden for the player and might decrease the enjoyment of the game. Also exceeding more than 12 ads generates a scroll bar for some browsers depending on the resolution. If there are more than 12 ads to be displayed, the game randomly selects 12 from the pool and records the history to make sure that all ads are played at least once. The rectangle surrounding each of the area flashes when an ad is placed in it.

Figure 2 shows a screen shot after the game is over. The pop-up box includes the list of ads, the selections made by the players, and the scores. It shows the result of a two-player game.

3.3 Behind the Scene

We record all clicks in XML format and store the data with the following information: game id, participants' ids, article id, ads with each ad id, the category the player has dropped an ad to, drop time, and the selection order among ads. The XML schema is described in Figure 3. The server calculates the score of an ad (*AdScore*) from this result.

Given an ad, we consider: the category that the majority of the users have chosen (i.e., 1, 0, -1 for relevant, somehow relevant and irrelevant, respectively), how many people have agreed on the most popular category, and the selected order and time when the category became the dominant one. We assume that an obviously relevant or obviously irrelevant ad would be chosen faster (i.e., prior selection order and faster selection time) than ambiguous ads. The value of *AdScore* for an ad 'a' is computed as follows with the meanings of the symbols in Table 1.

$$AdScore(a) = C \cdot \frac{|U_C|}{|U|} \cdot \frac{\sum_{u \in U_C} \frac{1}{o_u}}{|U_C|} \cdot \frac{T-t}{T} = \frac{C}{|U|} \sum_{u \in U_C} \frac{1}{o_u} \frac{T-t}{T} \quad (1)$$

The formula consists of four parts: the first part is the label of the dominant category. The second part measures the proportion of the users who chose the dominant category for the ad.

```
<game id="000001">
  <article id="joins-001"/>
  <ads>
    <ad id="1" mean="0.7742" reliability="0.0283"/>
                    ⋮
    <ad id="12" mean="-0.3333" reliability="0.0951"/>
  </ads>
</game>
```

Figure 4. Result of aggregation

The third part indicates the relative importance of the ad among those chosen for the category, and the fourth how quickly the particular category was chosen by the user.

AdScore ranges from -1 to 1. It gets higher as the number of agreed users increases, the individual category selections are made faster, and the category selection agreement is reached more quickly. The user gets points from the sum of the *AdScore* values for the ads they have correctly selected (which was also chosen by the majority). As the results get piled up, the server aggregates them and generates the final output. At this point, the *AdScore* values over all the ads get averaged into μ, and each ad gets a *reliability* score based on the standard deviation of *AdScore* among played games. The equation is as follows:

$$reliability(a) = |U_G| \left(\frac{1}{|G|} \sum_{i=1}^{|G|} \left(AdScore(a_i) - \mu \right)^2 \right)^{-\frac{1}{2}} \quad (2)$$

Where *a* is an ad in a specific article, $|G|$ is the number of games played with *a,* and $|U_G|$ is the number of participants played in game *G*. A high *reliability* value means that the selections of the players are consistent whereas low *reliability* indicates the score varies among games. The category (i.e., relevant, somehow relevant, irrelevant) of an ad with high *reliability* can be more trusted than low *reliability* ads. The final output is stored in the format of figure 4.

3.4 Evaluation of Underlying Systems

Based on the final output generation method as in the previous section (we call it mADtch collection), we can evaluate the underlying ad searching systems. Since *AdScore* is computed over all the categories *C* as in Equation (1), an ad searching system that categories ads into one of the three categories (i.e., relevant, somehow relevant, irrelevant) can be evaluate naturally. By allowing *C* to be a rank order, on the other hand, this game-based evaluation scheme can be used for ad searching systems that rank-order the ads for a document. That is, the evaluation scheme can be used for two types of underlying systems: one with categorical judgments of ads (e.g., relevant vs irrelevant) and the other with numerical relevance values for ads.

If the target system to be evaluated outputs category labels, the evaluation result will be the same as "precision" by comparing the labels of mADtch collection and those of the target systems. If the target system outputs relevance values for ads (e.g., 0.87% relevant), we can evaluate the system by sorting the values to rank the ads and calculating correlation between two rankings using, for example, Kendall's tau (τ) rank correlation coefficient.

If there is only one target system to evaluate, we can produce an absolute value for precision or τ coefficient. If there is more than one system, we need to ensure that the systems use the same set of documents (i.e., news articles) in order to make the evaluation valid. Without using the same set of documents, the results will not be directly comparable. While using the same pool of ads for individual documents is also important for direct comparisons of ad selection algorithms, this requirement may not be absolutely necessary. Unlike ordinary document retrieval system evaluations, where the documents are given, the performance of an ad searching systems can be determined by different kinds of ad pool as well. That is, two systems with different ad pools can be compared although they are not using the same ad pool.

In this situation, the mADtch system can simply take both outputs of the ad searching systems and mix them up to generate a single pool of ads that can be presented to the players. This is based on an assumption that two different ad pools don't affect each other. We call this *ad independence assumption*. In other words, even though the ads of two or more systems are mixed, we assume that the ads of one system don't affect the other. We have conducted experiments to prove that this assumption is valid as discussed in Section 5.

Although the ads from two underlying ad search systems are mixed for evaluation based on the ad independent assumption, singling out the result for a particular system can be done easily by simply ignoring the other set of ads. Since each system can get its own precision and τ coefficient, we can see which system as a whole is more effective by comparing the values obtained from their own ad pools. This flexibility helps evaluating a new system in comparison with those evaluated in the past. When a new system is to be evaluated, the new ad pool generated by the new system is mixed with the outputs of the old systems to create a mixed pool and proceed as if multiple systems are compared concurrently.

4. EXPERIMENT

In this section, we discuss the experiments we conducted to validate our approach. Our goal was to demonstrate that the cost needed for our approach would be significantly lower than that of the conventional approach while the effectiveness of the system evaluation performance is not compromised. We begin by explaining not only the data but also the procedure employed to generate the gold standard and describe relative benefits of using the proposed approach.

4.1 Data Set

To demonstrate the efficacy of our approach, we conduct a series of experiments on two sets of advertisement placement data. The first set is real-world data from a commercial advertisement system, Joins [3]. It is operated by one of major newspaper publishing companies in South Korea. The online newspaper site suits our need in that several (5 ~ 6) ads are placed for each news article. We selected 100 recent news articles from diverse news categories and crawled the news content including the title, news text, and ads associated with the news content. All of the news articles and ads are in Korean, which is the native language for the game players.

We collected the second data set from an existing commercial contextual ad searching system, gmail.com, which attaches ads to individual e-mail messages automatically. To obtain a set of ads for the same news articles used for the first data set, we sent an e-mail message whose content is identical to a selected news article to an e-mail address we have and collected the ads attached to the

[3] http://joins.com

e-mail on the receiving side. In order to ensure the resulting data collection match well with the first data set, we gathered six ads from each e-mail. If the number of ads is larger than six, we selected the top six. If it is smaller than six, we repeated the crawling process by opening the e-mail several times at two-day intervals to capture new ads. Gmail somehow shows slightly different set of ads when the same e-mail is opened at different times. The ads collected from Gmail were also written in Korean. All the results reported in this paper are based on the two data sets.

4.2 Game Setting

Before making mADtch available for game players, we composed three sets of games. The first one (dataset 1) is based on the dataset crawled from Joins.com. The second one (dataset 2) is the result of crawling ads from email messages via gmail.com. The last one (dataset 3) is based on the union of the ads in dataset 1 and dataset 2. Each news article in dataset 3 is associated with about 12 advertisements.

A pair of news article and a set of associated ads form a game stage. With the three data sets, the number of game stages is 300. This means a person can play a maximum of 300 unique games although the number of articles is 100. A person playing the mADtch game is given a game stage chosen randomly among the 300 stages without repletion every time the old game is cleared.

We haven't officially launched the game in public, but the beta version of mADtch is available on the web[4]. The beta version was launched in March 29th. For the experiment, we took the game result for two days. The total number of participants was 20, and they played 356 games (17.8 games per person on average).

4.3 Gold Standard

To compare the proposed method with the conventional approach of constructing a test collection, we constructed a gold standard through manual annotations. We recruited three annotators, those who are the participants of the online game, one undergraduate and two graduate students from the CS department in the university where we conducted our research, to annotate dataset 3, which includes the article-advertisement pairs of dataset 1 and dataset 2. The annotators' average age was 25.6. Each annotator was required to make a relevance judgment for each ad after reading each news article. The annotation policy is borrowed from Broder, et al.'s work [7]: each ad is annotated on a -1 to 1 scale: 1) relevant, 0) somewhat relevant, and -1) irrelevant. The initial inter-annotator agreement of judgments was 77.5%. To obtain a score for an article-ad pair we averaged all the scores and then rounded to the closest integer. We then used these judgments as a baseline to evaluate the two contextual advertisement systems described above.

4.4 Efficiency

During the annotation process, we recorded the time taken to annotate each article. After the annotators had completed all the tasks, we gave them a survey to obtain their feedback. We asked them to rate the easiness and joyfulness of the annotation, using a five-point scale with 1 being lowest (hard) and 5 being highest (easy). We also recorded the time spent for playing mADtch at the background of the game for comparison (the game with same articles used for gold standards).

[4] http://zzihee.kaist.ac.kr/madtch.html

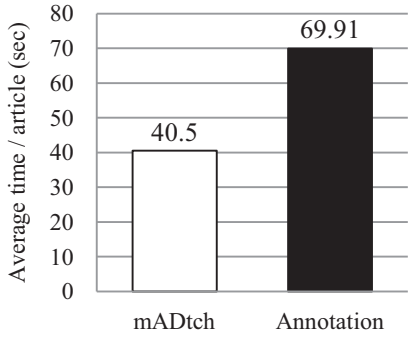

Figure 5. Time taken for annotating articles

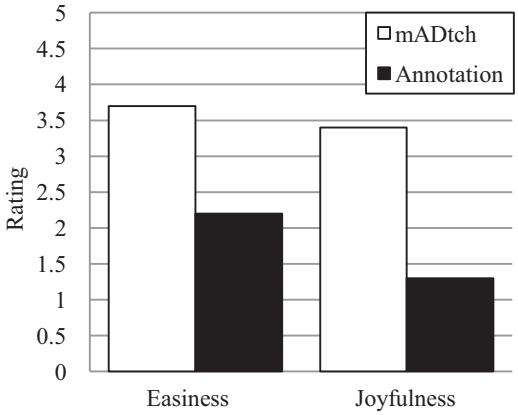

Figure 6. mADtch vs. Annotation: easiness and joyfulness

Again, we gave all the players an online questionnaire to obtain their feedback identical to the survey mentioned previously.

Figure 5 shows the average time required for users to annotate advertisements per news article. The charts demonstrate that the effort of a person on mADtch is much less than on annotation methodology. For average time over the 100 news articles, the mean on mADtch is 40.5 seconds – again, much better than 69.9 seconds on annotation methodology.

4.5 Game Benefits

Figure 6 compares user feedback in terms of easiness and joyfulness on the two evaluation methods. The annotators gave a very low easiness score to the annotation method. Although they did not feel that mADtch was absolutely easy in deciding whether an advertisement was relevant to an article or not, they gave it a much higher score than the annotation method.

There was a significant gap between two methods' joy scores, and the scores for mADtch were surprisingly higher. We suspect that the participants felt more comfortable and significantly less stressful with mADtch. Since the players of mADtch are unpaid volunteers, it is a fundamental requirement to make a game stress-less and enjoyable to attract people for participation. Overall, the experimental result shows mADtch has a sufficient merit for attracting public's participation.

5. EVALUATION

In this section, we examine our proposed scheme of using a GWAP game, namely mADtch, as an evaluation method.

Table 2. Evaluation with gold standard and mADtch

Comparison	Precision
mADtch against the first gold standard	0.9667
Ad systems against first gold standard	0.8607
Ad systems against mADtch result	0.8500

Table 3. Comparison of results of two evaluation methods

Gold Standard	System A		System B	
	Pre @ 1	Pre @ 3	Pre @ 1	Pre @ 3
mADtch	0.97	0.73	0.94	0.67
Annotator	0.93	0.79	0.87	0.62

Table 4. Intervention between two systems

Evaluation Metric	System A	System B
Precision	0.9833	0.9667
Rank Similarity (Union)	0.4933	0.3067
Rank Similarity (Relevant)	0.9445	0.8889
Rank Similarity (Irrelevant)	0.3067	0.2733

To do this, we first compared each of the two evaluation methods and the ad systems' result. We then directly compared the results of the two ad systems to see if there are any differences caused by the two different gold standards (i.e., result from mADtch and annotation). We also present evaluations dealing with ad independence and reliability correlation.

5.1 Comparison with Gold Standards

We compared the results provided by two distinct evaluation methods, manual and game-based, with ad system's result. As mentioned above, the first gold standard was constructed by human judges while the other one was obtained indirectly by running the mADtch game.

Table 2 shows that the gold standard and the mADtch result share 96.7% in common. Moreover, there exists only 1.07% difference between the precision values measured for the ad systems using the gold standards obtained by the two evaluation methods. The result shows that mADtch provides approximately similar judgments compared with the annotator-generated gold standard.

5.2 Effectiveness

For the next step, we compared the performances of two ad systems based on the two gold standards in terms of the widely used evaluation metric, Precision @ n. This experiment was done to compare and verify the performance of mADtch, on evaluating ad systems in practice.

Table 3 shows the observed performances through the two evaluation methods. Pre (precision) @ n indicates that the ratio of relevant ads within top *n* ads retrieved by an ad system. In case of Pre @ 1, Annotator (the gold standard created by the annotators) marked 0.93 (0.87) for System A (B), while mADtch marked 0.97 (0.94) for System A (B). We can see that the judgments made through mADtch were slightly more lenient than those made by the annotators for both systems. In case of Pre @ 3, mADtch marked more strictly for System A, while marking more leniently for System B. For both cases, Pre @ 1 and Pre @ 3, the relative rankings of the two systems are consistent regardless of the evaluation methods we employed.

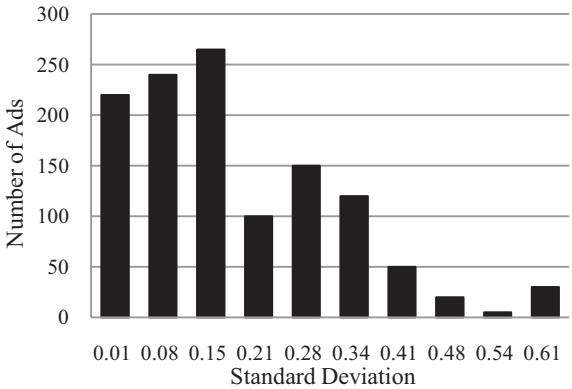

Figure 7. Judgment distributions based on standard deviation

The result is very encouraging that mADtch has a possibility of replacing the traditional human annotator-based method for generating a gold standard. The game-based approach has an additional merit that the accuracy of the judgments is likely to get increased with additional game players over time. This is even more important for subjective tasks for which inter-annotator agreement is low.

5.3 Ad Independence

It can be a problem if a system's performance measured by an evaluation method is affected by unexpected factors. In running the mADtch system, the ads originated from different ad searching systems are given to the user simultaneously at the same game stage. We assumed that the relevance of each ad is independent of other ads in the previous section. The following experiment was intended to indirectly prove the ad independence assumption. Table 4 shows that the result of precision and rank similarity between the games made out of a single system (System A) and the integrated (ads mixed) game (System B). The precision values indicate the extent to which the two judgment sets (one ad pool vs. a combined ad pool) agree to each other for a particular document. Ad System A (single pool) recorded 98.3% of precision, while Ad System B (combined pool) recorded 96.7%. The results show that there is little difference between the two cases, indicating that the ad independence assumption is reasonable.

We can also see the similar result when rank similarity is used as in Table 4. The rank similarity is measured by Kendall's τ metric. The rank similarity of relevant ads is much higher than that of irrelevant ads. It seems that there is a strong tendency that the players pay more close attention to selecting relevant ads in order than irrelevant ones, which seem to be selected in random order.

The overall results in Table 4 support the ad independence assumption, which helps conclude that two different ad systems can be evaluated simultaneously without interfering each other's evaluation result.

5.4 Reliability of the Game Results

Analyzing the game results, we realized that some games produced the relevance judgments that differ among players. This variance corresponds to the lack of inter-judge agreements when annotators are employed in the conventional method. We analyzed the game results further to investigate the extent to which the inter-player variances would influence the ad system evaluations.

Table 5. Effect of difference in reliability

Group	Precision
Low reliability ads	0.9333
High reliability ads	1.0000

We divided the game results (relevance judgments obtained from the games) into two groups, high reliability ads and low reliability ads, and then calculated precision values for the groups separately based on the gold standard created by the annotators.

Figure 7 shows that there are much more results with high reliability (i.e., low standard deviation) in the first place. When they were partitioned into two groups of the equal size with the threshold 0.176, we observed a slight difference between the precision values measured separately for the groups. Table 5 shows that the result with large standard deviation records lower precision. From this result, we can conclude that we should obtain a result with small standard deviation to construct a reliable test collection.

6. CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel evaluation method for subjective tasks. Our contribution is summarized as follows: 1) the quality of the evaluation is almost the same as that using the gold standard constructed out of human judgments, resulting in 96.67% in precision. 2) The efficiency of using mADtch is remarkable, compared to the time required for annotators. 3) It is more enjoyable, inexpensive, and easier than the traditional annotation based evaluation approach.

Some factors we have considered in the design don't seem to reflect the real user's behavior, however. For example, we expected that the player will drag and drop an ad that is the most relevant at first. On the contrary, most of the users ended up dragging and dropping ads from the right to the left or vice versa, ignoring the relevance order. Recoding observations from the real users and incorporating them into the game is an important future work.

As seen in Figure 6, mADtch outperforms the annotation approach in terms of ease and joy, but still users don't find mADtch as an attractive game they want to continue in a natural setting. We need to investigate on possibilities of adding more enjoyment factors. Since the evaluation method depends on the number of played games, it is important to make the player more engage into the game.

7. ACKNOWLEDGMENTS

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency). [NIPA-2010-C1090-1011-0008]

8. REFERENCES

[1] Ahn, L. V. and Dabbish, L. 2004. Labeling Images with a Computer Game. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[2] Ahn, L. V. and Dabbish, L. 2008. Designing Games with a Purpose. In Communications of the ACM.

[3] Ahn, L. V., Kedia, M. and Blum, M. 2006. Verbosity: A Game for Collecting Common-Sense Facts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[4] Ahn, L. V., Liu, R. and Blum, M. 2006. Peekaboom: A Game for Locating Objects in Images. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[5] Ahn, L. V., Maurer, B., McMillen, C. and Blum, M. 2008. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. In Science.

[6] Alonso, O., Rose, D. E. and Stewart, B. 2008. Crowdsourcing for Relevance Evaluation. In ACM SIGIR Forum.

[7] Broder, A., Fontoura, M., Josifovski, V. and Riedel, L. 2007. A Semantic Approach to Contextual Advertising. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval.

[8] Chakrabarti, D., Agarwal, D. and Josifovski, V. 2008. Contextual Advertising by Combining Relevance with Click Feedback. In Proceedings of International Conference on World Wide Web.

[9] Cleverdon, C., Mills, J. and Keen, E. M. 1968. Factors Determining the Performance of Indexing Systems. ASLIB Cranfield Research Project, Cranfield, England.

[10] Cormack, G. V., Palmer, C. R. and Clarke, C. L. A 1998. Efficient Construction of Large Test Collections. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval.

[11] Jones, S., Van Rijsbergen, C. J. 1975. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266. Computer Laboratory, University of Cambridge.

[12] Lieberman, H., Smith, D. A. and Teeters, A. 2007. Common Consensus: a Web-Based Game for Collecting Commonsense Goals. In Proceedings of the International Conference on Intelligent User Interfaces.

[13] Lin, J. and Katz, B. 2006. Building a Reusable Test Collection for Question Answering. Journal of the American Society for Information Science and Technology.

[14] Speer, R., Krishnamurthy, J., Havasi, C., Smith, D., Lieberman, H. and Arnold, K. 2009. An Interface for Targeted Collection of Common Sense Knowledge Using a Mixture Model. In Proceedings of the International Conference on Intelligent User Interfaces.

[15] Voorhees, E. M. 2000. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. Information Processing and Management.

[16] Voorhees, E. M. 2003. Overview of the TREC 2003 Question Answering Track. In Text REtrieval Conference.