Tower of Babel: A Crowdsourcing Game Building Sentiment Lexicons for Resource-scarce Languages

Yoonsung Hong[†], Haewoon Kwak[‡], Youngmin Baek[*], Sue Moon[†]
[†]Division of Web Science and Technology, KAIST, Daejon, Korea
[‡]Telefónica Research, Barcelona, Spain
[*]College of Communication, Yonsei University, Seoul, Korea
yshong@kaist.ac.kr, kwak@tid.es, ymbaek@gmail.com, sbmoon@kaist.edu

ABSTRACT

With the growing amount of textual data produced by online social media today, the demands for sentiment analysis are also rapidly increasing; and, this is true for worldwide. However, non-English languages often lack sentiment lexicons, a core resource in performing sentiment analysis. Our solution, *Tower of Babel (ToB)*, is a language-independent sentiment-lexicon-generating crowdsourcing game. We conducted an experiment with 135 participants to explore the difference between our solution and a conventional manual annotation method. We evaluated ToB in terms of effectiveness, efficiency, and satisfactions. Based on the result of the evaluation, we conclude that sentiment classification via ToB is accurate, productive and enjoyable.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation (HCI)**]: Group and Organization Interfaces—*collaborative computing, computer-supported cooperative work*

Keywords

World Wide Web, distributed knowledge acquisition, lexicon construction, sentiment labeling, online games

1. INTRODUCTION

Sentiment analysis, equivalently opinion mining in many academic literatures [4, 17], extracts and analyzes the emotions conveyed in texts. Numerous research from various domains takes advantage of the sentiment analysis; for instance, sentiment analysis is applied to detecting political stances [23], characterizing personality [11], measuring happiness [9], and even predicting stock market [3]. With the growing use of the Web worldwide, the demands for sentiment analysis are increasing, and as a response to the demands we continue to improve the science of extracting and understanding affects and emotions from texts.

Sentiment analysis is generally achieved by two classes of approaches: rule-based systems and machine learning classifiers. They differ in that the former requires sentiment lexicons while the latter needs annotated corpora for training classifiers [2]. The former is often more appreciated for having competitive advantages in convenience and generalizability. To be specific, employing sentiment lexicon is, in general, technically easier for manipulation and less constrained to the domain peculiarities of texts than training a classifier with annotated corpora.

Either case, it is costly to equip necessary linguistic resources for sentiment analysis, e.g. sentiment lexicons and annotated corpora. As a consequence, English remains the primary beneficiary of such linguistic resources [1, 4]. The popularity of English and economic circumstance of English speaking nations explain such a strong English slant in resource availability. English is, in fact, the third widely spoken language worldwide, followed by Chinese and Spanish, and five of G-20 major economies speak English as the first language. However, the rapidly growing volume of non-English contents on the Web today presents needs, challenges, and opportunities for multilingual sentiment analysis research. We give Twitter, one of the most active platforms to create textual contents in the world, as an example. A report published by Semiocast points out that as of October 2011 more than 60 percent of entire messages on Twitter is in non-English[1]. Likewise, bridging the imbalance between the excess of non-English texts on the Web and scarcity of linguistic resources available for non-English sentiment analysis sets forth an agenda demanding immediate attention from the research community.

Equipping necessary linguistic resources such as sentiment lexicons and annotated corpora is inevitably the first step toward multilingual sentiment analysis. As mentioned above, since sentiment lexicon is more convenient and generalizable, building sentiment lexicon is considered a task with a higher priority than building annotated corpora. Hitherto a few approaches are suggested for building sentiment lexicons but the manual labeling approach is considered conventional for its high accuracy [2]. However, the manual approach is often arduous, costly, and time-consuming. The success story of ESP game, a crowdsouced image labeling game for Google images, suggests that integrating crowdsourcing and game into a framework can be a solution for the shortcomings of the conventional manual approach. We can leverage crowdsourcing as a medium to reach out to a larger workforce and parallelize tasks, and game for turning the tasks fun and thus lowering the costs of achieving labeling.

In this work we propose *Tower of Babel (ToB)*, a crowdsourcing game that helps construct sentiment lexicons for any language. We show how a widely practiced design pro-

[*]Baek conducted this study while he was at KAIST.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

[1]http://semiocast.com/publications/2011_11_24_Arabic_highest_growth_on_Twitter

cess turns a boring classification task into an engaging game. We conducted an experiment with 135 volunteered participants, and quantitatively assessed the effectiveness, efficiency, and subjective evaluation of ToB. The result of the evaluation shows that ToB can achieve a similar level of accuracy for sentiment classifications to a conventional manual approach which takes the form of survey. We also show that ToB can lead to more consistent judgment across crowd workers than the conventional approach. Besides, some participants even expressed classifying words on ToB is interesting and satisfactory. We conclude that sentiment classification using ToB is accurate, productive, and even enjoyable. Our contributions are two-fold. One is the gaming framework built for a purpose of constructing sentiment lexicons. The other is the quantitative evaluation of our approach with regard to accuracy and productivity.

The rest of the paper is organized as follows. We first take a visit on previous literature. Next, we explain the design process and implementation of the game. Then, we report and analyze evaluation results, and discuss future work. Finally, we close the paper with a conclusion.

2. RELATED WORK

We review previous literature to put our work in context. We first give a brief overview on subjective and sentiment analysis with emphasis on multilingual sentiment analysis, explain various approaches to construct lexicons, and finally introduce the concept of 'game with a purpose'.

2.1 Subjective and Sentiment Analysis

Subjective and sentiment analysis is a field of study that concerns with "people's sentiments, appraisals or feeling toward entities, events and their properties" [14], and "the automatic identification of private states such as opinions, emotions, sentiments, evaluations, beliefs, and speculations in natural language" [2]. Subjective analysis and sentiment analysis differ in that the former classifies texts either subjective or objective, whereas the latter classifies texts positive, negative or neutral [2].

For subjective and sentiment analysis, rule-based systems and machine learning classifiers are most widely used approaches [2]. Rule-based systems classify subjectivity and sentiments in texts pursuant to a set of predefined rules and lexical resources. For instance, OpinionFinder classifies a sentence subjective if two or more strong subjective expressions, which are found in its subjectivity lexicon, appear in the sentence. Similarly, WordNet-Affect [24], SentiWordNet [10], and SenticNet [5] are other popular tools and resources. It is worth noting that SenticNet makes distinction from the others in that it analyzes natural language texts at a semantic level instead of a syntactic level [4]. Machine learning classifiers, on the other hand, work under a different logic from the rule-based systems; it classifies subjectivity and sentiment after a training with annotated corpora.

2.2 Constructing Lexicons

Generally, three approaches are suggested for constructing subjective and sentiment lexicons for resource-scarce languages; manual annotations, translation, and bootstrapping.

First and the most accurate way to build such lexicons is through manual annotations. Linguistic Inquiry and Word Count (LIWC), one of the well-known sentiment analysis tools, contains a large-scale lexicon built by manual annotations [18]. The manual approach is often neither realized nor preferred since it is arduous, costly, and time-consuming.

Next, we can also translate lexicons from a resource-rich language (i.e. often English) to resource-scarce languages. Although this approach is convenient, translated lexicons suffer from bad quality. For instance, when a lemmatized form is used for translation, words may loose a subjective meaning [2].

Third, bootstrapping starts with a small-sized annotated lexicon, often called a seed, and expands the lexicon through linguistic relations in ontological resources; e.g. synset in WordNet and synonyms in thesaurus. Similar to the second approach, the bootstrapping comes with quality issues due to the automated nature of the knowledge acquisition process. More importantly, this approach is only feasible for the languages that have ontological resources such as thesaurus.

Besides, crowdsourcing [16] and gaming [6, 8] are newly suggested yet promising alternative solutions for constructing lexicons; they make the manual annotations attainable, practical and affordable.

2.3 Game with a Purpose

Game with a Purpose (GWAP) is not something significantly different from conventional games; however, in GWAP, players adjunctively produce outcomes useful for something else while playing the games. Often, the outcomes are neither suitable nor feasible for computers to produce.

The ESP game is one of the most successful GWAPs. The players in the game produce labels for images while playing game. As of 2008, nearly 200,000 players produced more than 50 million labels for the images on the web through well-designed game experience [26]. The game is fast-paced, enjoyable and competitive such that players are willing to perform tasks otherwise not fun. We mention a few other successful GWAPs as examples: Peekaboom [29] gamifies the task of locating objects within images, and Phetch [27] gamifies the task of adding descriptions to images. More recently, Cooper *et al.* applied gamfication to solve biochemistry problems (i.e. folding and designing proteins) [7].

According to a TED talk given by Jane McGonigal, "we invest three billion hours weekly playing online games (worldwide)" [15]. Such enormous amount of time and energy devoted to games can be utilized for doing and producing something valuable for the world.

3. GAME DESIGN: TOWER OF BABEL

We propose *Tower of Babel (ToB)*, a crowdsourcing game constructing sentiment lexicons for resource-scarce languages. We used the image of Tower of Babel as backdrop and thus named the game accordingly. The game is situated in a multilingual context, and the scenery of the game involves piling up blocks just as Tetris. ToB is a collaborative game in which a pair of players are matched to make sentiment classifications on words; the players are rewarded for making a matching classification with the partner. Below, we explain the design objectives that we set, design considerations that we took, and design goals and decisions that we made; then we also give a detailed description of ToB, and finally report on the prototyping results.

The goal of our GWAP is to design a game that builds sentiment lexicons for resource-scarce languages. As a demonstration, we build a word-level sentiment lexicon for the Korean language through our game.

3.1 Design Considerations

Here we give a comprehensive overview of design considerations that we took for designing our game ToB.

3.1.1 Gamification

Gamification refers to a process of turning a non-game task to a game. While a good gamification results in mutual benefits for both players and task givers, it is hard to satisfy both sides simultaneously mainly because transforming boring tasks to games requires more than just intuition; it requires systematic thinking.

The first step toward gamification is to answer where or what to gamifiy. To answer this question, we need to first identify the components of the task. However, tasks are often multifaceted that it is hard to conceptualize and break them down into definite components. The Black Box theory, extensively used in computer science and electrical engineering fields, offers a simple method to disambiguate task abstraction. Black Box theory formalizes a task by three components: input, process, and output. The theory examines the causal relationship between input and output, and illuminates the logical flow of the process.

Once the process is clearly identified, we have only one last step left: transforming it into a game. Here we either designing a new game from scratch or borrowing ideas, insights, and gaming mechanisms from successful games, and incorporating them in a new game The latter is often preferred since it learns from success.

3.1.2 Motivation

Designing an engaging game that motivates players is the most essential step toward making a successful game. What are gaming elements and features that enhance player engagement? It is not easy to answer because there are neither definite nor widely accepted theories in both the best practice in industry and academic literature. Nevertheless, although limited, some findings in psychology, instructional game, and human computer interaction give us useful tips.

To begin with academic literature, Yee suggests an empirical model for explaining motivations of playing online games; his model consists of three main concepts: achievement (advancement, mechanics, and competition), social (socializing, relationship, and teamwork), and immersion (discovery, role-playing, customization, and escapism) [30]. Similarly, a theory proposed by Tekofsky also suggests achievement, recognition and satisfaction as motivational factors of online gaming [22]. Yee's model and Tekosky's theory suggest that achievement, social, and immersion are important factors to consider when designing a motivational game.

In practice as a response to the aforementioned theories, a set of useful gamification features and techniques are suggested to help design motivating GWAPs [26]. According to von Ahn and Dabbish, timed responses and randomness can enhance immersive gaming experience, score keeping can trigger feelings of satisfaction, and recognitions; score keeping and score list facilitate competitions, which in turn lead to satisfaction. Similarly, Siering suggests gaming features, such as setting clear goals, progress indicators, level up, badges, and leader board, can enhance motivation in games [20].

3.1.3 Quality Control

Motivating players is only the first part of making a successful GWAP. Once we meet the needs of players, we must consider how to meet a task giver's goal which is, in our case, achieving high-quality sentiment classifications. Previous GWAPs such as the ESP game suggest that collaborative game is an effective gaming format for controlling quality of labeling tasks. In a collaborative game, it is easy to arrange players such that they mutually validate each other's labels.

We reference several quality control mechanisms from previous collaborative GWAPs. We explain each mechanism briefly: (i) In *output-agreement games* such as ESP Game [25], two players each produce outputs given an input, and get rewarded when their outputs match. (ii) In *input-agreement game* such as Peekaboom [29], and others [27, 28], two players compare the output produced by their partner's with their own, and guess whether they are given the same input. (iii) In *inversion-problem game* such as TagATune [13], player B is given the description of an input made by player A, and the players are rewarded when player B makes a correct guess on the input given to player A.

For the advancement of the quality control, we consider several options. Although random matching is the most common among collaborative GWAPs such as ESP game, matching two unleveled players can bring down players' motivation and engagement. Since poor motivation and engagement of players diminish classification quality, we expect that a player matching mechanism that takes the variation of players' skills into account can promote quality control. Also, a reputation system that profiles players with trust scores or a algorithm that detects ill-intended players can be integrated for better quality control. Additionally, repeating classifications and placing taboo outputs can also enhance the quality and diversity of labels.

3.2 Design Decisions

Here we explain design decisions that we made based on the design considerations outlined in the previous sections.

First, we decide to learn from success. Particularly, we decide to follow Tetris model because (i) Tetris is widely known, (ii) Tetris is simple, intuitive and engaging, and (iii) block placing routine in Tetris resembles our classification task that the task can be integrated into the game format with minimal modifications.

Second, we decide to focus on three motivational aspects for our gamification: achievement, social, and immersion. (i) Following Tetris model turns out to be an apt choice for adding immersive experience to our game because Tetris inherits immersion-evoking elements such as timely responses and randomness. (ii) We make our game collaborative and add social features with the helps of social networking services to generate competitions and cooperations. (iii) We incorporate a score system and leader board into our game to provoke the feelings of achievement.

Finally, we decide to employ *output-agreement* mechanism for quality control because it is simple and adaptive to Tetris-like games. For the purpose of a small-scaled experiment, however, we decide to limit quality control to random matching in this work.

In summary, we decide to build a collaborative game similar to Tetris, with a conventional quality control mechanism and social features.

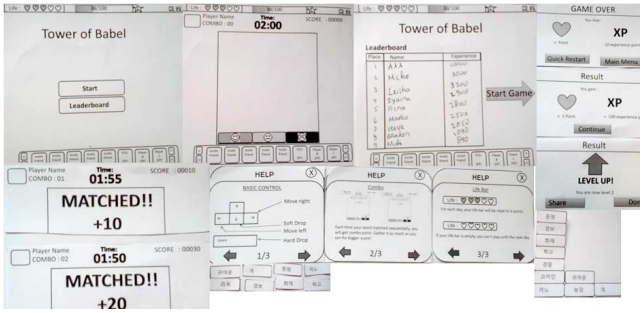

Figure 1: Paper prototype interfaces

3.3 Game Description

Probably, the best way to describe our game ToB is comparing it with Tetris since Tetris serves as a basis and inspiration for ToB. In Tetris, one of seven different shapes of blocks, known as tetrominoes, are drawn randomly for placement. As soon as a block appears on the play, the block begins to fall from the top to the bottom along the playing field (i.e. a rectangular vertical shaft) as if there exists gravity. When a block starts to fall, a player manipulate the direction and orientation of the block, and places the block in a right place. When blocks are piled up such that they make horizontal alignment and leave no space between them, the blocks are removed from the playing field and the player is rewarded with points. The flow of ToB and Tetris is similar; players need to place blocks in appropriate places to get rewarded.

The differences, however, come from several places: (i) blocks in ToB have the same shape and size, (ii) each block carries a different word, (iii) the bottom edge of the playing field is broken into three stacks representing different sentiment classes (i.e. positive, neutral, and negative), (iv) players need to place each block on a stack that matches with the sentiment of the word carried by the block, (v) ToB is collaborative game in which a pair of players are matched to play together, and (vi) a block is removed from the playing field and points are rewarded when a pair of players place the same block in the same stack (i.e. output agreement); unlike Tetris, making a spaceless horizontal alignment of blocks does not lead to rewards or winning of the game. If the pair of players, for example, places a word-block 'good' on the 'positive' sentiment stack, the block will be removed from the playing field for both players. The players are rewarded with points for making an agreement on the sentiment classification.

Similar to Tetris, players in ToB lose the game when blocks stack up above the top of the playing field. The players win the game if they make placements of the predefined number of word-block without reaching the top.

3.4 Rapid Prototyping

Rapid prototyping is a type of prototyping strategies practiced widely in the field of human computer interaction, and in the context of developing user interfaces [19]. The core principle of rapid prototyping is in fast iterations [19]. Paper prototyping is one of the best rapid prototyping methods that complies with the principle. One can easily build a mock-up or proof-of-concept interface with pieces of paper, makers and scissors. The point of paper prototyping is to build an minimal functional-oriented interface for quick testing with users [19, 21]. The paper prototyping is often used in an early stage of a design process since it is fast to identify usability problems and iterate for improvements.

We made a paper prototype based on the aforementioned design considerations and game description, as shown in Figure 1. We conducted a prototype evaluation with five participants (i.e. students from KAIST) to validate the design of the game before implementing it. At the beginning of the evaluation, a facilitator gave the participants the background information and instructions. Then, the facilitator asked the participants to start interact with the prototype as if they are actually playing a game. When participants made interactions with the prototype, the facilitator, taking the role of machine, made corresponding changes on the interface with his hands. For instance, when participants pressed 'Start' button on the first screen, the facilitator replaced the first screen with the second one. The participants were asked to start, complete the game, and explore as many features as possible. During the evaluation, an observer made notes of problems confronted by the participants. Finally, at the end of the evaluation session, we asked participants to share their experiences with us.

Although we did not find significant problems from our game design in the evaluation, a few minor issues were discovered. To begin with, we observed a few participants having trouble grasping how to play the game; they asked questions about how to play the game during the game. One participant suggested that the instruction of the game should be written on the screen, and presented to players before the game starts. In implementation, as a response to the recommendation we added a written instruction right next to the playing field so that players can look at it when needed. Second, some participants suggested removing a navigation menu which always appears on the bottom of the screen because it was useless and even annoying. Third, one participant complained that blocks being dropped and blocks already in place should have different color so they are distinguishable at a glance. Also, some participants asked about how the actions taken by a pair of player will be synchronized in the real version of the game–whether player A needs to wait for a new word when player A makes the placement of word-block ahead of player B, and what happens when there is no one to match with. These were crucial points that we missed to consider beforehand. We addressed these issues in our implementation by adjusting the speed of falling blocks and creating a virtual player whose action is probabilistically determined.

Overall from the prototyping evaluation, we received valuable feedbacks. We incorporated the lessons learned from the prototyping at the stage of the implementation.

4. IMPLEMENTATION

In this section we give detailed description of our ToB implementation. We begin with explaining the system architecture, authentication and partner matching, game interface, and finally ranking and social features.

4.1 System Architecture

For the front-end, we used HTML and Javascript for building the gaming interface and user-experience-level functionalities. Thanks to Jake Gorden who shared Tetris Web application in open source, we saved enormous amount of

time and energy. We used PHP, an popular open source server side scripting language, for implementing the back-end functionalities such as managing players, processing requests from the front-end, and making queries to a database system. Ajax was used to make the communication between the front-end and back-end codes possible. We used MySQL, an open source database managment system, for storing the list of words to give to players, the system traces, and the classifications made by the players. We integrated Facebook Graph API to ToB to enable features for enhancing motivation and attracting more audience. For example, features such as 'Like' and 'Share' were added to allow players to let their friends know about the game, and even send invitations. We also built a ranking system with the help of Facebook friend list to promote social motivation. We requested the players to give permissions to access necessary information from their Facebook accounts.

4.2 Authenticating and Partner Matching

When players entered ToB web page, we first asked them to login with their Facebook credentials. Our system matched players with a partner when they pressed the 'Start' button. Although matching players with similar levels can lead to more meaningful results and enjoyable gaming experiences, for the purpose of a small-scaled experiment we adopted random matching as our matching policy; in fact, many others successful GWAPs (e.g. ESP game) used random matching. If there was no available player for matching, a player was paired with a virtual player. On top of the random matching, we modified the code to support virtual player mode. In the experiment, in fact, all players faced with the virtual player whose action was probabilistically determined. We expect that the player availability will be less of problems when the game is actually deployed to the general public. Rather, managing players with ill intentions open up an interesting research problem.

4.3 Game Interface

Figure 2 is a screenshot of ToB during a game. The screen was mainly composed of an information pane on the left, and playing field on the right side. The information pane included a score board, preview for the next block, and rules & instructions. Each block contained a word. The rules & instructions section provided information such as how to play the game, how to get rewarded, and the wining condition. The visual effects of the game took place in the playing field. The bottom edge of the playing field was further subdivided into three subsections; the subsections stand for the stacks for positive, neutral and negative word-block.

With the start of a game, blocks began to fall from the top of the playing field. Then, players had to manipulate a falling block with arrow keys on keyboard to place it to a stack that matches with the sentiment of the word in the block. For instance, if a block contained 'Good', the player needed to place the block on the left most stack since the left most stack represented positive sentiment. When the player placed a block upon a stack, the system checked whether the partner had placed the block in the same sentiment stack. If two players made the same classification on the word-block, the block was removed from the playing field; at the same time, a message 'Hit' appeared in the center of the screen. If the classification made by the players did not agree, the block turned gray and remained on the stack for the rest of

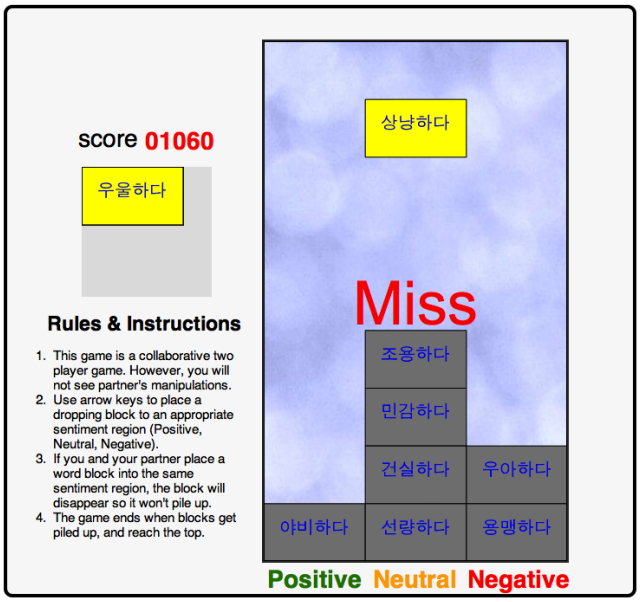

Figure 2: In-game snapshot of ToB

the game; in this case, we showed a message 'Miss' in the center of the playing field. Note that when only one player made the placement of a block, the block stayed gray in the players's plying field until the player's partner too made his or her placement; however, the block was removed from the playing field as soon as the partner made the identical classification on the block with the player.

4.4 Ranking and Social Features

Figure 3 shows the screen shown to players when they lose a game. The same screen was shown for winning except that a message for victory replaced one for loosing. The screen was composed of four sections. We explain each from the top box to the the bottom most one respectively. The first box announced the result of just-ended game–whether the players won the game or not. Also, a button for starting a new game was added in the section. In the second box, we reported score statistics; both cumulated score and the score from just-ended game were shown. Also, we reported the rankings for players. In third box we listed the rankings of all players and their cumulated scores. For the purpose of a small-scaled experiment, we only provided the global ranking; the ranking among players' Facebook friends was not implemented. The forth box contains social plugins for Facebook. In this section, players were able to make comments, press 'Like' to express their satisfaction, let their friends know about new activity, or use 'Send' to invite their friends to ToB.

5. EVALUATION

As evaluation, we experimentally and quantitatively assessed our game ToB in terms of effectiveness, efficiency, and subjective evaluation. Below, we explain method and results obtained from our experiment and quantitative analysis.

A corpus of 80 words was constructed for experiment and analysis. Out of 80 words, 60 were obtained from an established dictionary, and annotated into one of positive, nega-

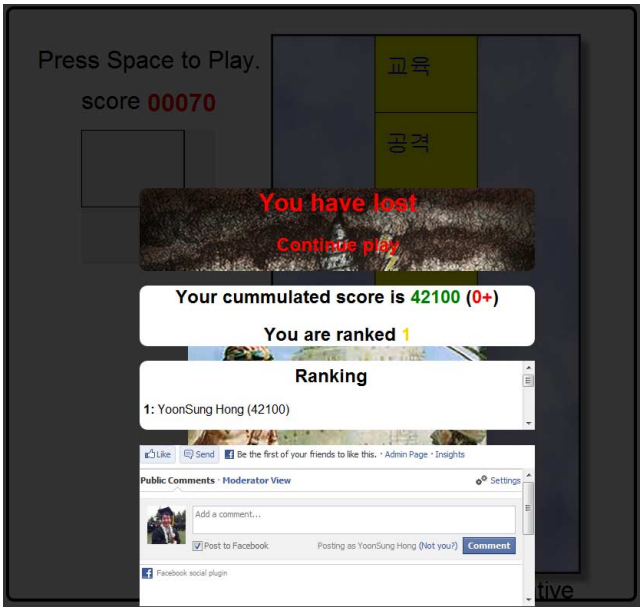

Figure 3: Ranking and Social Features

tive, or neutral class by human judges. The remaining 20 words were intentionally left ambiguous to assess the limitations of the sentiment class system and the consistency among judges. Here we call the former 60 words as *known set*, and the latter 20 words as *unknown set*.

We recruited 135 volunteer participants through viral marketing on Twitter. In experiment, we randomly assigned the participants into either *game* or *manual* classification condition. The participants in former condition were instructed to play ToB for sentiment classifications. The participants were matched with a virtual player for the purpose of a small-scaled experiments. The virtual player correctly classified the sentiments for the words in *known set*; however, for the words in *unknown set*, it took 70% chances to agree with a real player. In contrast to the *manual* condition, the participants in latter condition were instructed to fill out a common web survey consists of a bunch of words and radio buttons. Among 135 participants, 79 people finished the classification task ($N_{game} = 32$; $N_{manual} = 47$). The demographics of the participants showed that more than a half of participants were male (60 percent), and most of participants were in late 20s ($Mean_{age} = 28.87$, $SD_{age} = 4.21$).

5.1 Accuracy of Classification

We quantified the accuracy of a participant classifying words using precision score, a extensively used measure in information retrieval and classification system. We computed the average precision scores for both conditions. The result showed that the propensity of accurate classification under both settings is almost the same ($Mean_{game} = .80$, $SD_{game} = .124$; $Mean_{manual} = .81$, $SD_{manual} = .127$). Statistical testing also confirmed that difference of precision score between two conditions is negligible ($t_{(77)} = -.42$, $p = .67$). In sum, we achieved sentiment classifications through ToB as accurate as ones by the conventional manual annotation method.

5.2 Consistent Judgment across Judges

A gold-standard data set is hardly available in reality. Instead, we measure the consistency among judges as a proxy to asses the quality of the knowledge base acquired. Here we computed Krippendorff's α [12], a well-known measure for assessing inter-judge agreement in social sciences. Krippendorff's α quantifies the degree of consistent judgment across multiple judges, indicating that α becomes 0 with no consistency and 1 with complete consistency. To estimate which condition leads to higher α coefficient among judges than the other, we calculated 95 % confidence interval of α coefficients, based on bootstrapping by drawing 1000 random sample from each condition [12]. While game condition resulted in $\alpha = .19$ and its 95 % CI = [.14, .27], manual condition resulted in $\alpha = .11$ and its 95 % CI = [.06, .18]. We here make two interesting observations. First, very low α coefficient for the manual condition indicates that the words in *unknown set* fails to achieve consistency across judges. Second, however, game condition achieves slightly higher consistency across judges while the difference between two conditions is not distinguishable according to conventional statistical criterion. In other words, sentiment classification via ToB might lead to more consistent agreement across judges than by manual system, although its level is unsatisfactory or fails to seduce statistical certainty.

5.3 Task Completion Time

If one condition demands less time than the other for completing a task with a comparable level of accuracy, such a condition is considered efficient since it accomplishes the same amount of work in a shorter time. We measured the time taken (i.e. in seconds) for a human judge to finish the entire classification tasks. We found that the task completion time under game condition ($Mean_{game} = 174.97$, $SD_{game} = 47.91$) is definitely shorter than that under manual condition ($Mean_{manual} = 242.40$, $SD_{manual} = 103.38$). Statistical testing also assured that the difference of completion time is substantial ($t_{(77)} = 3.44$, $p < .001$). In sum, the result demonstrated that sentiment classification by ToB demands judges less time and thus is more productive than the ordinary manual annotation method.

5.4 Subjective Evaluation of the Task

After completing classification tasks, we asked participants how much (i) interesting, (ii) engaging, and (iii) satisfactory their tasks were. Questions were given with a conventional 5-point Likert-type scale, indicating that higher number denotes more interesting, engaging, or satisfactory experiences with the experiment. Since three evaluations were highly inter-correlated (Cronbach's $\alpha = .64$), we averaged the three measures and called the average subjective evaluation. When comparing subjective evaluation of the task between two conditions, we added two covariates related with the task (i.e., subjectively estimated length of the task and effectiveness). Results via ordinary least squares are reported in Table 1. We observed that people under game condition evaluated the task more positively than those under manual condition ($\beta = .31$, $p < .05$). In short, the result clearly shows that sentiment classification via ToB led to more positive evaluation than conventional manual annotation method.

Source	Parameter Estimates	t-ratio
Intercept	1.54^{**} (0.50)	3.111
Subjectively estimated length of the task	$-.14^{*}$ (0.07)	-2.019
Effectiveness (precision score)	2.38^{***} (0.57)	4.201
Experimental condition (Game = 1; Manual = 0)	0.31^{*} (0.15)	2.035

$^{*}\ p < 0.05$ $^{**}\ p < 0.01$ $^{***}\ p < 0.001$

Table 1: Result of ordinary least squares testing showing that people under game condition give more positive evaluation than those under manual condition

5.5 Evaluation Summary

With 135 participants we assessed the effectiveness, efficiency, and subjective evaluation of ToB and conventional manual classification method. Results showed that ToB achieved classifications of sentiment as accurate and consistent as conventional manual method. Also, ToB took less time to achieve the same amount of classification tasks. Further, participants perceived more positive feeling with ToB. In summary, sentiment classification by ToB is accurate, productive, and enjoyable.

6. DISCUSSION AND FUTURE WORK

We discuss about the challenges faced by the current research, and suggest future directions.

6.1 Toward Richer Knowledge Bases

Sentiment is often classified into three classes such as positive, neutral, or negative; however, sometimes three classes are not adequate to accurately capture the sentiment perceived by human judges. A number of participants in the experiment answered that some words were hard to be classified into one of the three classes. They described some words as 'both positive and negative', or 'neither positive or negative, and nor neutral'. We can partially resolve the problem by adding granularity to sentiment classes (e.g. sentiment scores in real numbers). Introducing extra dimensions of sentiment is, however, a better solution because sentiment expressed by humans is more complex than the extent of a polarity-based system can connotate. LIWC, for instance, has 64 behavioral and psychological dimensions on top of positive and negative affects, such as 'anxiousness', 'anger', and 'inhibition' [18]. For future work, we would like to expand the dimensions of sentiment on the game in order to produce richer sentiment lexicons.

Some participants in the after-experiment survey mentioned that some words can imply different sentiments depending on their contexts of use. In fact, domain-specific lexicons can enhance the quality of sentiment analysis. For future work, we will refine ToB so that domain specific sentiments can also be collected. One possible solution is showing relevant pictures or words that trigger the sense of a targeted domain along with a word to be labeled.

In this research, we only focused on GWAPs constructing lexicons, but we can challenge making GWAPs for annotating corpora in the future. Annotated corpora will offer an alternative option for sentiment analysis, and complement the shortcomings of the lexicon-based approach, especially in resource-scare settings.

6.2 Large-scale Deployment and Evaluation

Amazon Mechanical Turk and CrowdFlower are a few of the most popular crowdsourcing platforms. Although these platforms are suitable for most of general tasks, they do not meet the requirements for a GWAP intended to construct non-English lexicons; they lack social features and access to global population. In this research, we thus integrated our game ToB with Facebook API. Beside social features, we can also target mobile users for the future work. According to newzoo, mobile games are played by 100 million U.S. alone, and it jumped 35% from the last year[2]. Also, it is reported that (i) people spend more time playing with mobile apps than web browsing on desktops[3], and (ii) people spend most of time playing game when using mobile[4]. Further, social networking services started to offer social features to mobile applications. Since mobile games attract a large audience and offer convenient features to make promotions, we will consider developing mobile version of ToB in the future.

In this work we evaluated the gaming approach for building sentiment lexicons. For the future work, we would like to evaluate the quality of the lexicon achieved by the gaming approach. One way to assess the quality of a lexicon is comparing it to well-established lexicons. However, since established lexicons are hardly available to non-English languages, we can consider building one by bootstrapping from a seed set or translating an English lexicon. Alternatively, we can also apply the lexicon achieved from deployment to non-English texts such as news or blog articles for quality evaluation of the lexicon.

7. CONCLUSION

We propose a language independent sentiment lexicon generating framework, *Tower of Babel (ToB)*. ToB aims to help resource-scarce languages achieve high quality lexicons at low costs. For realization, we adopted a conventional design procedure widely practiced in human computer interaction field, and various Web development technologies as leverage to design and implement ToB. We conducted a quantitative study to assess ToB in terms of effectiveness, efficiency, and subjective evaluation. The evaluation results shows that sentiment classification via ToB is accurate, productive and even enjoyable. As a next step, we will deploy ToB, and conduct further studies to better support non-English languages facing resource scarcity problems.

8. ACKNOWLEDGEMENTS

We thank professor Uichin Lee for inspiration for the game design, and Satria Hutomo for performing paper prototyping. We thank Jake Gordon for sharing Tetris game in open source. We also thank all participants of the experiment. Finally, we thank all members of Advanced Networking lab at KAIST for their cheerful supports.

[2]http://www.newzoo.com/trend-reports/mobile-games-trend-report/

[3]http://techcrunch.com/2012/12/05/flurry-mobile-apps-television/

[4]http://www.emarketer.com/Article/Time-Spent-With-Mobile-Apps-Rivals-TV/1009533

This work is supported by WCU (World Class University) program under the National Research Foundation of Korea and funded by the Ministry of Education, Science and Technology of Korea (Project No: R31-30007).

9. REFERENCES

[1] A. Balahur. *Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types*. PhD thesis, Universitat d'Alacant, 2011.

[2] C. Banea, R. Mihalcea, and J. Wiebe. Multilingual sentiment and subjective analysis. *Multilingual Natural Language Processing*, 2011.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011.

[4] E. Cambria and A. Hussain. *Sentic Computing*. Springer Netherlands, 2012.

[5] E. Cambria, R. Speer, C. Havasi, and A. Hussain. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI Fall Symposium Series*, 2010.

[6] E. Cambria, Y. Xia, and A. Hussain. Affective common sense knowledge acquisition for sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012.

[7] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, Z. Baker, David Popovic, and F. players. Predicting protein structures with a multiplayer online game. *Nature*, 466, 2010.

[8] A. Das. Psychosentiwordnet. In *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[9] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS One*, 6(12):e26752, 2011.

[10] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LRECąŕ06)*, 2006.

[11] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *Privacy, security, risk and trust, 2011 IEEE third international conference on social computing*, oct. 2011.

[12] K. Krippendorff. *Content analysis: An Introduction to Its Methodology*. Sage, 2004.

[13] E. Law, L. von Ahn, R. Dannenberg, and M. Crawford. Tagatune: A game for music and sound annotation. In *In Proceedings of the Eighth International Conference on Music Information Retrieval*, 2007.

[14] B. Liu. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2010.

[15] J. McGonical. Jane mcgonigal: Gaming can make a better world. `http://www.ted.com/talks/jane_mcgonigal_gaming_can_make_a_better_world.html`, 2010. [Online; accessed 21-Feb-2013].

[16] S. Mohammad and P. Turney. Crowdsourcing a word-emotion association lexicon. *To Appear in Computational Intelligence*, 2013.

[17] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2(1-2)*, July 2008.

[18] J. W. Pennebaker, F. E. Martha, and B. J. Roger. *Linguistic Inquiry and Word Count*. Erlbaum Publishers, 2007.

[19] M. Rettig. Prototyping for tiny fingers. *Commun. ACM*, 37(4), Apr. 1994.

[20] G. Siering. Gamification: Using game-like elements to motivate and engage students. `http://citl.indiana.edu/news/newsStories/dir-mar2012.php`, 2012. [Online; accessed 21-Feb-2013].

[21] C. Snyder. *Paper Prototype*. Morgan Kaufmann, 2003.

[22] S. Tekofsky. Theory of gaming motivation. `http://www.thinkfeelplay.com/theory-of-gaming-motivation/`. [Online; accessed 21-Feb-2013].

[23] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the fourth international aaai conference on weblogs and social media (ICWSM)*, pages 178–185, 2010.

[24] R. Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004.

[25] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, New York, NY, USA, 2004. ACM.

[26] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8), Aug. 2008.

[27] L. von Ahn, S. Ginosar, M. Kedia, and M. Blum. Improving image search with phetch. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, april 2007.

[28] L. von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, New York, NY, USA, 2006. ACM.

[29] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, New York, NY, USA, 2006. ACM.

[30] N. Yee. Motivations of play in online games. *Journal of CyberPsychology and Behavior*, 9, 2007.