

Leveraging the Wisdom of the Crowds for the Acquisition of Multilingual Language Resources

Arno Scharl, Marta Sabou, Stefan Gindl

MODUL University Vienna, Department of New Media Technology
Am Kahlenberg 1, 1190 Vienna, Austria
{arno.scharl, marta.sabou, stefan.gindl}@modul.ac.at

Walter Rafelsberger

Holzweg e-Commerce Solutions
Sillgasse 12, 6020 Innsbruck, Austria
walter@rafelsberger.at

Albert Weichselbraun

University of Applied Sciences Chur; Faculty of Information Science; Pulvermühlestr 57, 7004 Chur, Switzerland
albert.weichselbraun@htwchur.ch

Abstract

Games with a purpose are an increasingly popular mechanism for leveraging the wisdom of the crowds to address tasks which are trivial for humans but still not solvable by computer algorithms in a satisfying manner. As a novel mechanism for structuring human-computer interactions, a key challenge when creating them is motivating users to participate while generating useful and unbiased results. This paper focuses on important design choices and success factors of effective games with a purpose. Our findings are based on lessons learned while developing and deploying Sentiment Quiz, a crowdsourcing application for creating sentiment lexicons (an essential component of most sentiment detection algorithms). We describe the goals and structure of the game, the underlying application framework, the sentiment lexicons gathered through crowdsourcing, as well as a novel approach to automatically extend the lexicons by means of a bootstrapping process. Such an automated extension further increases the efficiency of the acquisition process by limiting the number of terms that need to be gathered from the game participants.

Keywords: crowdsourcing, language resource acquisition, sentiment detection.

1. Introduction

Games with a purpose leverage collective intelligence, which is described as combining “behavior, preferences, or ideas of a group of people to create novel insights” (Segaran, 2007). Collective intelligence from groups of people often produces better results than individual domain experts (Surowiecki, 2004). Games with a purpose have been used successfully to solve problems that computers cannot yet solve, such as tagging images (Ahn, 2006) and annotating content (Siorpaes and Hepp, 2008). The main challenges when creating such games are motivating users to play the game while generating useful data, and ensuring that the process yields unbiased results.

This paper investigates game design choices that ensure solving these challenges. It builds upon the lessons learnt from *Sentiment Quiz*,¹ a Web-based social verification game for sentiment detection that was released as part of the *US Election 2008 Web Monitor* (Scharl and Weichselbraun, 2008).² This election monitoring project aimed at gaining new insights into information diffusion via interactive online media, and into the interdependence of news media coverage and public opinion. To capture the editorial slant of news media coverage, the system automatically measured media attention (frequency of candidate references) as well as media sentiment (positive versus negative).

The Sentiment Quiz initially served two main purposes (Phase 1): firstly, the acquisition of a large set of manually tagged sentences to evaluate the accuracy of sentiment detection algorithms; secondly, the analysis of *hostile media effects* (i.e., the different perception and interpretation of Web content depending on the reader's political orientation) in conjunction with extensive user polling between January and November 2008.

In Phase 2, the Sentiment Quiz was changed to query users for their assessment of terms instead of sentences, with the aim of creating sentiment lexicons in multiple languages (Figure 1). Such lexicons are a prerequisite for most sentiment detection methods and will serve as an example of language resource (LR) acquisition throughout this paper.

The screenshot displays the 'US08 Sentiment Quiz' interface. At the top, there is a navigation bar with 'facebook', 'Friends', 'Applications', 'Inbox', 'Home', and a search field. Below this, the quiz title 'US08 Sentiment Quiz' is shown with sub-links for 'Game', 'About', 'Awards', and 'Feedback'. The main content area features a question: 'Does this sentence use positive or negative language?' with a quote: 'We are headed down a path that is certain to end in the destruction of our experiment in democracy.' Below the quote are buttons for 'Yes' (red minus) and 'No' (green plus), with a note: 'Hover over the buttons to see what they stand for.' To the right, a 'Hall of Fame' table lists top performers: 1. Walter (535), 2. Julia (361), 3. Hermann (211), 4. Reinhard (205), 5. Norberto (204), 8. Fabio (154), 9. Isabella (130), 10. You (127), 11. Lord (121), 12. Sandra (102). Below the table are 'Tips & Tricks' and a 'Status' section showing a score of 127 points and a level of 5. At the bottom, there are sections for 'Spread the Word' and 'New Media MBA'.

Figure 1. Sentiment Quiz Sentence Evaluation

¹ www.modul.ac.at/nmt/sentiment-quiz

² www.ecoresearch.net/election2008

For deploying the Sentiment Quiz, an application framework has been developed (Rafelsberger and Scharl, 2009), which is compatible with a range of developer platforms such as *Facebook*,³ *iGoogle*⁴ and *Netvibes*.⁵ It acts as a wrapper that gives developers more flexibility and enables them to implement applications on multiple social platforms.

This paper shows how games with a purpose can address existing bottlenecks in language resource acquisition and evaluation (Section 2), including a discussion of incentive schemes and differences in extrinsic and intrinsic motivation of participants. Section 3 outlines principles of game design and how they relate to crowdsourcing tasks. Section 4 exemplifies the use of games for language resource acquisition by building seed sentiment lexicons based on aggregated assessments from Facebook users. Section 5 summarizes the evaluation results and describes the subsequent expansion of seed lexicons through an automated bootstrapping process. Section 6 concludes the paper and suggests an extension of the current approach to support both the acquisition and sharing of language resources.

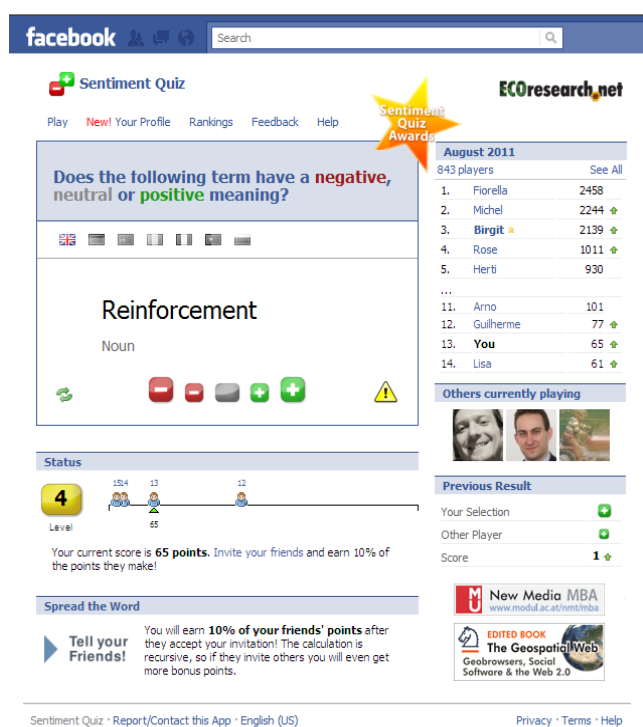


Figure 2. Sentiment Quiz Dictionary Extension

2. Language Resource Acquisition

The *Sentiment Quiz* has successfully demonstrated the potential of using Facebook for building and evaluating sentiment lexicons. Other language resource acquisition tasks that could be addressed through crowdsourcing include corpus annotation with facts, word senses, and events; the creation of annotated corpora of special resource types – e.g., Twitter feeds, medical text; the provision of parallel corpora, question-answer sentence pairs,

³ www.facebook.com

⁴ www.google.com/ig

⁵ www.netvibes.com

speech transcriptions, and bi-lingual entailment corpora. Such language resource acquisition tasks are currently addressed either by using manual (high-quality, but slow and costly) or automated approaches (fast and cheap, but error-prone and language-specific).

Making use of the collective intelligence of online users, crowdsourcing is currently gaining popularity as an attractive alternative to combine the advantages of manual and automated approaches. In this section we discuss the most popular crowdsourcing platforms and compare them in terms of the application domain they serve, the incentive scheme they rely on, the technical platforms used to distribute them, whether they exploit synergies between the hosted projects as well as whether new crowdsourcing projects can be easily added by third-party practitioners.

2.1 Crowdsourcing Marketplaces

Most research projects that acquire LRs through crowdsourcing make use of crowdsourcing marketplaces such as *Amazon Mechanical Turk*⁶ and *CrowdFlower*,⁷ which are dedicated portals to extrinsically motivate participants by economic incentives. A key benefit of *Mechanical Turk* is the low setup cost as new projects can be easily created through an API, deployed on *Mechanical Turk*'s infrastructure and benefit from its large user base. Synergies between projects remain unexploited, however, and there is a high possibility of obtaining low quality output due to users' economic motivation and financial incentive to cheat.

2.2 Games and Virtual Communities

To reduce the incentive to cheat, other crowdsourcing approaches leverage the intrinsic motivation of a community interested in a domain. Game examples from the language technology area include *PhraseDetectives*⁸ to acquire anaphorically annotated corpora, *Minefield* to transcribe images of Arabic text (Dahab and Belz, 2010), and *Sentiment Quiz* to elicit sentiment terms and assessments of political statements.

Compared to crowdsourcing marketplaces, games with a purpose promise superior results due to intrinsically motivated players and making better use of sporadic, explorer-type users. A critical mass of players can be achieved by (i) leveraging social networking sites as a distribution mechanism (*Sentiment Quiz*) and/or (ii) building a community of committed players whose expertise and trustworthiness can be assessed based on their game history (*Zoonivers*).

Another significant trend of games with a purpose, in general, is the creation of GWAP platforms which bundle together multiple games as opposed to individual games being published in a stand-alone fashion. For example, *GWAP.com* bundles together seven games designed at Carnegie Mellon University but it is not open for usage by others. The *OntoGame platform* (Siorpaes and Hepp, 2008) focuses on the deployment of games that support the ontology lifecycle and has been used to deploy five

⁶ aws.amazon.com/mturk

⁷ www.crowdfower.com

⁸ anawiki.essex.ac.uk/phrasedetectives

content annotation games. The *Social Application Development Framework* (Rafelsberger and Scharl, 2009) of MODUL University Vienna currently hosts two games for acquiring sentiment detection data, with a third game targeting the climate change domain currently under development as part of the “Climate Change Collaboratory” project (www.ecoresearch.net/triple-c).

3. Designing Language Games

Games with a purpose must be designed in a way that they engage users while delivering valuable information to solve the underlying problems. In this section we rely on lessons learned from the *Sentiment Quiz* game to discuss the main design decisions a game developer must consider: the type of game, incentive schemes, task complexity and result validation.

Different LR acquisition games might benefit from different game types. Game designers need to consider the best fitting game type in terms of players (e.g., individuals, pairs, or groups) and game mechanics (e.g., selection, Q&A, output agreement, input-agreement (Ahn and Dabbish, 2008)). Both *Sentiment Quiz* games are designed for pairs (although, players' responses might be matched against cached input from other players, if not enough players play simultaneously) and use the output-agreement paradigm (i.e., players are given the same input and they win if their assessment of this input is the same). In general, output agreement games work best with tasks that require evaluating certain features of the input (e.g., the polarity of a term). However, they are rather restrictive collecting broad and diverse information from users (e.g., annotating an image with all possible tags). In these cases, input-agreement games are more suitable – players must describe presented inputs to each other until they can determine, based on these descriptions, whether the two inputs differ or not. As a side effect of this process, various descriptions of input artefacts are obtained.

Besides these basic types of games, designers might consider combining various games into a workflow: for example, using an input-agreement style game to generate

artefact annotations and then channelling the output of this game into an output-agreement game in order to select the most relevant annotations.

Sentiment Quiz games make use of a variety of incentive mechanisms including score boards (right side of interface) and game levels (bottom part of interface). These are generic mechanisms that can easily be provided by the application platform for future games. As a more complex incentive structure, currently we are experimenting with implementing expert-league games which combine tasks from different games running on the application platform in an arbitrary fashion. These games can only be played by those that have completed all levels of all games and are aimed to retain players beyond completing all games. By implementing games on social networking platforms, game creators can exploit viral mechanisms specific to these platforms to attract new players: a player receives a certain percentage of the points gained by players he has recruited among his social network acquaintances (in our games this bonus accounts to 10%). Finally, the citizen science flavour of these games has a strong motivational value on its own for many participants.

A crucial task when applying games with a purpose is to make sure that the games yield unbiased results. Result verification can be achieved by resource pre-production and post-production methods. A number of simple measures can be taken to ensure output of high quality: (i) hide the identity of the other player; (ii) analyze the temporal distribution of answers; (iii) assign trust values to each player, which in turn determine the impact of their answers – e.g. insert questions with known answers into the exercise queue and identify users who tend to score low on these questions; (iv) avoid exploitable patterns in the sequence of answers, since users who identify the pattern could quickly earn credits without solving the puzzle; (v) using resource-specific aggregation strategies to handle partial human contributor agreement (e.g. full agreement, majority vote, expert and multi-level reviews and average); (vi) defensive task design to encourage users to put in genuine efforts to carry out the tasks.

Platform	Domain	Incentive Scheme	Distribution Mechanism	Project Synergies	Add New Project/Game
MTurk	Generic	Economic	Web Portal	No	Yes
CrowdFlower	Generic	Economic	Web Portal	No	Yes
Zoonivers	Astronomy	Intrinsic	Web Portal	Yes	Yes
GWAP.com	Image Annotation	Intrinsic	Web Portal	Yes/Implicit	No
OntoGames	Semantic Web, Ontologies	Intrinsic	Web Portal	Yes/Implicit	No
Sentiment Quiz	Politics, Sentiment Detection	Intrinsic	Social Media	No	No

Table 1. Overview of generic and domain-specific crowdsourcing platforms

4. Sentiment Detection

Sentiment detection is a challenging language processing task. Lexical approaches assume that there is a conceptual connection between words and their adjacent text (Giora, 1996). They calculate sentiment towards a target term by

measuring the co-occurrence between the term and words from a sentiment lexicon – i.e., a specific LR type that correlates words with their perceived polarity on a range between -1 (negative) and +1 (positive). The acquisition of such lexicons is problematic for several reasons:

- Word sentiment is a **subjective feature**, difficult to obtain via automated algorithms.
- Manual methods usually require ratings from multiple subjects for each word (which are averaged to obtain a final polarity value), thus leading to **high annotator costs and long acquisition times**.
- Generating sentiment lexicons through simple translation does not provide accurate results, as words often have **different polarities across languages**; e.g., translating from English to German: ‘abolish’ (-1) vs. ‘beseitigen’ (-0.3), ‘dirt’ (-1) vs. ‘schmutz’ (-0.38), ‘excitement’ (1) vs. ‘aufregung’ (-1). This issue, among others, results in the currently limited availability of such lexicons in less-spoken languages.

To address these limitations, we describe how to build and evaluate user-generated sentiment lexicons in multiple languages through games with a purpose, and how to extend these lexicons automatically by a bootstrapping process based on the analysis of semantic associations in various corpora. Leveraging the wisdom of the crowds by engaging users in online games addresses the scarcity of human resources to tackle such tasks.

Results from the *Sentiment Quiz* reflect the potential of games with a purpose for research projects in general and for acquiring sentiment lexicons in particular: more than 3,500 users provide about 325,000 evaluations in seven different languages (English, German, French, Italian, Portuguese, Spanish and Russian) and according to a five-point sentiment scale (very negative, negative, neutral, positive, very positive). This yielded a number of compact sentiment lexicons in multiple languages, whose preliminary evaluation showed promising results compared to lexicons compiled by experts. These lexicons served as the basis for the extension processes outlined in the next section.

5. Extending Sentiment Lexicons

We have built a bootstrapping method to automatically detect both sentiment terms and indicators (= terms that occur in polar texts and indicate either a positive or negative sentiment) from archives of domain-specific documents. The three-step process starts with the calculation of sentiment values for all documents in the archive using the seed lexicon from *Sentiment Quiz*. Based on these values the system subsequently compiles a corpus consisting of the k strongest positive and negative documents in the second step. The frequency distribution of each term in the positive and negative section of this corpus is decisive for the integration of the term into the seed lexicon. We then use the Naive Bayes algorithm to compute candidate terms for inclusion into the sentiment lexicon and obtain their probability values based on the following formulas:

$$P(\sigma(t_j)|C^-) = \frac{n(t_j|C^-)}{n(t_j)}$$

$$P(\sigma(t_j)|C^+) = \frac{n(t_j|C^+)}{n(t_j)}$$

The terms with the highest probabilities are included into the seed lexicon. Multiple iterations of this process generate a considerable number of new terms, while the reduction of the used documents by half in each iteration guarantees high reliability of the new sentiment terms.

10-fold cross-validation on three lexicons (the seed lexicon, the bootstrapped lexicon as well as an expert lexicon derived from the General Inquirer) showed significant improvements achieved through the bootstrapping process (Weichselbraun et al., 2011) – in terms of precision (measure of exactness), recall (measure of completeness), and the F-measure (a hybrid metric that combines both aspects). The results also revealed that the performance of semi-automatically compiled sentiment lexicons is comparable to lexicons compiled by experts (particularly when used for lexical analysis, as compared to machine learning approaches), although they contained less than half the number of terms.

6. Conclusion and Outlook

The wisdom of the crowds contained in social evidence sources can be used in several ways to acquire and evaluate language resources. This paper presented crowdsourcing in the tradition of *games with a purpose* as a reliable and cost-effective method for language resource acquisition and evaluation. The crowdsourced sentiment lexicons were extended by means of a bootstrapping process. The improved accuracy achieved through this extension process is particularly useful in situations where comprehensive lexicons compiled by linguists are not available –in the case of less-spoken languages, for example, or when processing unusual expressions found in content from social media sources.

A successor of the Sentiment Quiz currently being developed will target not only the acquisition of language resources, but also their sharing and distribution among various stakeholders (e.g. researcher centers, companies, game participants, etc.). It will also integrate Linked Data repositories⁹ as structured evidence sources to enrich and validate concepts and relations identified in unstructured coverage from news and social media sources.

7. Acknowledgement

The methods and results presented in this paper are being developed within DIVINE¹⁰ and Triple-C,¹¹ two research projects funded by the *Austrian Climate Research Program* of the Austrian Climate and Energy Fund (klimafonds.gv.at), and *FIT-IT Semantic Systems* of the Austrian Research Promotion Agency (www.ffg.at).

8. References

- Ahn, L.v. (2006). “Games with a Purpose”, *Computer*, 39(6): 92-94.
- Ahn, L.v. and Dabbish, L. (2008). “Designing Games with a Purpose”, *Communications of the ACM*, 51(8): 58-67.

⁹ www.linkeddata.org

¹⁰ www.weblyzard.com/divine

¹¹ www.ecoresearch.net/triple-c

- Dahab, K. and Belz, A. (2010). A Game-based Approach to Transcribing Images of Text. *International Conference on Language Resources and Evaluation (LREC-2010)*. N. Calzolari et al. Valletta, Malta: European Language Resources Association.
- Giora, R. (1996). "Discourse Coherence and Theory of Relevance: Stumbling Blocks in Search of a Unified Theory", *Journal of Pragmatics*, 27: 17-34.
- Rafelsberger, W. and Scharl, A. (2009). Games with a Purpose for Social Networking Platforms. *20th ACM Conference on Hypertext and Hypermedia*. C. Cattuto et al. Torino, Italy: Association for Computing Machinery: 193-197.
- Scharl, A. and Weichselbraun, A. (2008). "An Automated Approach to Investigating the Online Media Coverage of US Presidential Elections", *Journal of Information Technology & Politics*, 5(1): 121-132.
- Segaran, T. (2007). *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. Beijing: O'Reilly.
- Siorpaes, K. and Hepp, M. (2008). "Games with a Purpose for the Semantic Web", *IEEE Intelligent Systems*, 23(3): 50-60.
- Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. London: Little, Brown.
- Weichselbraun, A., Gindl, S. and Scharl, A. (2011). Using Games with a Purpose and Bootstrapping to Create Domain-Specific Sentiment Lexicons. *20th ACM Conference on Information and Knowledge Management (CIKM-2011)*. Glasgow, UK: Association for Computing Machinery. 1053-1060.