A Novel Human Computation Game for Critique Aggregation

Claudiu Cristian Musat and **Boi Faltings**
École Politechnique Fédérale de Lausanne
Route Cantonale
Lausanne, Switzerland

Abstract

We present a human computation game based on the popular board game - Dixit. We ask the players not only for annotations, but for a direct critique of the result of an automated system. We present the results of the initial run of the game, in which the answers of 15 players were used to profile the mistakes of an aspect-based opinion mining system. We show that the gameplay allowed us to identify the major faults of the extracted opinions. The players' actions thus helped improve the opinion extraction algorithm.

Introduction

Human computation games have proven to be a reliable way to model incentives in crowdsourcing tasks. Starting from the ESP image annotation game (Von Ahn and Dabbish 2004), the diversity of such games steadily increased. For instance, *Verbosity* (Von Ahn, Kedia, and Blum 2006) gathers knowledge about words, while *Phrase Detectives* (Chamberlain, Poesio, and Kruschwitz 2008) was used to construct a corpus for NLP tasks. But so far, human computation has largely been seen as a method to generate labeled data. Defining a wider range of human computation operations and leveraging them for effective acquisition of commmonsense knowledge remain largely open problems.

We present a different type of game, in which we **gather critiques** of the way a complex system performed on individual items. We evaluate an aspect based sentiment analysis system. Previous games for sentiment classification have been created (Weichselbraun, Gindl, and Scharl 2011; Al-Subaihin, Al-Khalifa, and Al-Salman 2011; Musat, Ghasemi, and Faltings 2012). In this work, however, the focus is no longer simple labeling. There are two goals: to **evaluate the performance** of an automated system by finding the mistakes it makes and to elicit critiques from the players and use them as *classification features* to determine **which subcomponent must be improved**, and how.

We split the aspect-based opinion extraction system into two subsystems - one that defines the aspects and one that retrieves opinions about them. The game then incentivized the players to critique the subsystem that makes the most mistakes. By recording this critique, we were able to determine that the initial aspect definition method was inappropriate. Moreover, an in-depth look at what faults the players noticed showed that the negation detection was faulty, which affected the quality of the opinion extraction subsystem.

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Gameplay

We base our game on a popular round-based board game, Dixit.Different playing cards are distributed to each player. Each round the player who goes first, *the storyteller*, selects one of her cards, which becomes *the target* card. She says aloud a *description* of it and places it face down, Each other player selects one of their own cards, that best fits that description, and form a pile. **The goal** of these other players is to trick the rest into thinking their card was actually the *target*. Once the played cards are shuffled and revealed, everyone except the storyteller picks one of the played cards, which they believe was the target. Players who pick the real target card earn points, *and* so do those whose card is mistaken for the target by others.

The storyteller needs to get some, but not all of the others to identify his card. Something in between earns her points. The description must be **not too specific, nor too broad**. This is the key of the applicability of Dixit to critiquing, as the *descriptions* become good classification features.

Fig. $1a$, $1b$ are two sample cards. Each card contains two phrases, and for each phrase a relevant section is highlighted, for instance *slow service - N*. This means that the aspect discussed is *service* and our algorithm extracted a negative (N) opinion about it, because of the modifier *slow*. Each card must contain two phrases that discuss the same aspect and have the same polarity. If not, then the card is a mistake. For instance, if the player believes that *birthday* and *wedding* don't represent the same aspect, then card $1b$ is a mistake. If he believes that *slow* is positive, card $1a$ is a mistake.

In a first, **selection**, stage, the players draw cards from a deck. They discard the cards that contain no mistakes and keep the ones that do, until they hold $\tau = 12$ cards. In the second stage, they follow the Dixit rules above, using the remaining cards. In the selection process, they get acquainted with the types and relative quantities of mistakes the system makes. They are then able, when it is their turn to be a *storyteller*, to utter a description of one of the cards they hold, which will also fit cards that others hold. Fig. 1.e and 1.f show examples of definitions that were given by the players.

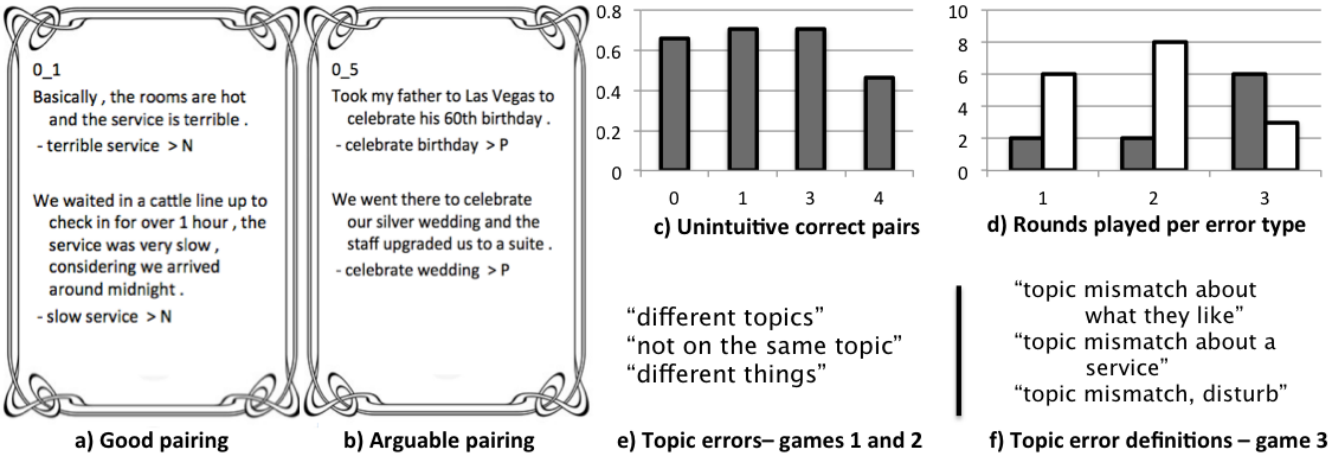

Figure 1: Sample game cards and evaluation results

Experiments and Results

We present the results of the offline pilot study, using 15 players, separated in 3 groups of 5 players each. Each group played a separate instance of the game. We created two decks of cards, that differed in the way the aspects were constructed. Aspects are relevant traits of the object being discussed - in this case hotels, for instance the cleanliness or location. The first deck of cards contained words that were manually selected and separated into aspects. The second deck of cards contained aspects based on topics extracted using LDA (Blei, Ng, and Jordan 2003). Attaching a given polarity to the aspects was done using a dictionary approach (Wilson et al. 2005), based on lexical dependencies. In games 1 and 2 we used the first deck of cards, while group 3 had the control, LDA cards.

A first question that we wanted to answer was whether **the players understood the rules**. For the first stage, we found the agreement between the cards marked *correct* by players in games 1 and 2, which had the same cards, was 83.5%. For the second stage, in 66.7% of all the rounds played, the storyteller received points, meaning that he was successful in finding a critique of the necessary specificity. From the two observations we concluded that the players understood the rules and were consistent in their answers.

The card selection phase of the game is in essence an **evaluation** of the system. The more pairs are correct, the better the system performs. It is an alternative to methods based on predicting the review numeric rating, such as RMSE or MAE. This is important, because we notice that the numeric rating is *often not correlated* with the content of the reviews. Figure 1c shows the proportion of correct cards, distributed by the **rating difference** between the two reviews that contained the phrases within the pair. We can see that even an opinion pair extracted from two reviews rated 1 and 5 has *a 50% chance* of being actually correct.

The aspect based opinion mining problem requires defining the aspects and retrieving the opinions about them. The overall accuracy is dependent on the outcomes of both subprocesses. However, the overall precision and recall cannot pinpoint which subsystem is to blame for the misclassifications. We use the storyteller's definitions to determine this.

Intuition*: The more errors a subsystem makes, the more critique it will draw.* The white bars in Fig. 1d show the number of times the storyteller chose a definition related to the extraction accuracy, for instance *"wrong polarity in context"*. The grey bars show the number of times the definition was related to a topic problem, such as those shown in Fig. $1e$ or $1f$. The horizontal axis plots the game number. These results show that players considered the LDA topics in game 3 a much bigger problem than the polarity extraction. The opposite is true for games 1 and 2, where the polarity extraction problems were higher.

Intuition*: The more errors a subsystem makes, the more detailed its critique will be.* In our case, the descriptions of topic errors in game 3 (Fig. 1f) are significantly longer than the ones in games 1 and 2 (Fig. 1e).

The players are able to extrapolate from the mistakes in the cards they see. The definitions they create, such as *missed modifier*, can be used in **two** ways: to locate the problem (in this case the polarity labeling) and to provide improvements. By *aggregating* these improvements from multiple players, we can provide both a valuable critique and improvement solutions for automated systems.

Conclusions and Future Work

We created a novel human computation game that changes the focus from the traditional label gathering to critique aggregation. The game simultaneously evaluates and gathers critique that can lead to immediate improvements of the analyzed system. A distinctive characteristic of the proposed game is that it is neither *adversarial*, nor *cooperative*, but hybrid. This solves the problem of repeated trivial answers that plague many cooperative games.

The game helped us find that the LDA topics were the main cause of the poor performance of our opinion extraction system and, in addition, helped us locate a negation identification problem. We are extending the presented game in an online environment.

References

Al-Subaihin, A.; Al-Khalifa, H.; and Al-Salman, A. 2011. A proposed sentiment analysis tool for modern arabic using human-based computing. In *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*, 543–546. ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Chamberlain, J.; Poesio, M.; and Kruschwitz, U. 2008. Phrase detectives: A web-based collaborative annotation game. *Proceedings of I-Semantics, Graz*.

Musat, C. C.; Ghasemi, A.; and Faltings, B. 2012. Sentiment Analysis Using a Novel Human Computation Game. In *Proceedings of the 3rd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP*, 1–9. Jeju, Republic of Korea: Association for Computational Linguistics.

Von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326. ACM.

Von Ahn, L.; 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*, 75–78. ACM.

Weichselbraun, A.; Gindl, S.; and Scharl, A. 2011. Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. In *CIKM*, 1053–1060.

Wilson, T.; Hoffmann, P.; Somasundaran, S.; Kessler, J.; Wiebe, J.; Choi, Y.; Cardie, C.; Riloff, E.; and Patwardhan, S. 2005. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, 34–35. Stroudsburg, PA, USA: Association for Computational Linguistics.