Metrics of games-with-a-purpose for NLP applications

Jon Chamberlain, Richard Bartle, Udo Kruschwitz, Chris Madge, Massimo Poesio
School of Computer Science and Electronic Engineering
University of Essex
{jchamb,rabartle,udo,cjmadg,poesio}@essex.ac.uk

Abstract

A Game-With-A-Purpose (GWAP) approach is a promising way to collect data on a large scale but its adoption for NLP annotation projects has been slow. In this paper, we propose evaluation metrics adapted from free-to-play (F2P) games to help understand what makes games successful for NLP applications.

1 Games for Natural Language Processing (NLP)

A GWAP approach is seen as one of the most promising ways to collect annotations on a large scale for Artificial Intelligence (AI) applications in areas such as computer vision and NLP, but their adoption for the latter has been sluggish. We believe that one of the reasons for this slow progress is the lack of evaluation metrics, as this makes it difficult for NLP researchers to understand which types of games are working and which are not. In this paper, we point out a number of metrics from free-to-play (F2P) games that could be adapted and propose a set of redefined metrics based on our experience developing games for NLP.

Many papers presenting GWAPs for NLP focus exclusively on the design of the game; however, it is the long-term performance of systems that should be evaluated. Evaluation metrics tend to concentrate on measuring either the quantity of data produced or the quality of such data, but the success of games in terms of player engagement is rarely measured or only measured informally. For example, the coreference game *Phrase Detectives* provides quantitative figures on the number of players, the quantity of data annotated and measures of agreement between the game's most frequent label and a gold standard (Poesio et al., 2013). Similarly, *Puzzle Racer* and *Ka-Boom!*, two games for collecting word sense information, are primarily evaluated in terms of the accuracy of the judgements (Jurgens and Navigli, 2014). **Throughput** (the time it takes to completely annotate an item once all player decisions have been aggregated) and **Average Lifetime Play (ALP)** have been introduced in GWAP evaluation (von Ahn and Dabbish, 2008), but are not widely reported.

2 Free-to-play (F2P) games

GWAPs for NLP are designed to be as game-like as the language task will allow so it makes sense to learn from the gaming industry. In particular, free-to-play (F2P) games have a number of widely adopted performance indicators (Xicota, 2017; Unity, 2014), that could be compatible with GWAPs. In F2P games, many players won't pay anything to play, but the remaining players will spend amounts that range from a few cents to thousands of dollars to enhance their game experience, from power ups and extra time to complete tasks, to customised game items (Dziedzic, 2016). The industry has therefore developed a range of metrics that tell companies not just how many players play and for how long, but what engages these players to ultimately become payers:

- **Cost per Acquisition (CpA)**, the cost to get a game player through advertising;
- **Lifetime Value (LTV)**, the total amount of money a player will pay to play the game;
- **Average revenue per user (ARPU)**;
- **K-Factor**, an indication of growth rate;
- Number of active players over a time scale, e.g. **Monthly Active Users (MAU)**;
- **Retention**, the percentage of players retained over a time period;

Metric	Description in relation to GWAP
Cost per Judgement (CpJ)	Average cost to get a player to provide a useful judgement.
Judgements Required (JR)	Average judgements required to complete an item.
Cost per Item (CpI)	Cost to acquire a completely annotated item.
Cost per Acquisition (CpA)	Cost to have someone start to play a game.
Lifetime Judgements (LTJ)	Total judgements made in the game per player.
Average Judgements per Player (AJpP)	Judgements per player.
Average Lifetime Play (ALP)	How long players play a game.
Monthly Active Users (MAU)	Total players who have submitted a judgement in a month.
Retention and Churn	Percentage of players retained/lost over a time period.
Throughput	Number of completely annotated items produced per hour.

Table 1: Summary of proposed metrics for games-with-a-purpose for NLP.

- **Churn**, the percentage of players who stop playing over a time period.

3 Metrics for GWAPs

F2P performance indicators can be adapted as metrics for annotation work and, when combined with existing quality metrics, could provide a more illuminating evaluation of the success of a GWAP for NLP than is currently available.

Our adaptation is based on the intuition that for GWAPs, judgements (i.e., an action from a player that provides useful data) rather than revenue are the measure of success. Thus, one measure would be **Cost per Item (CpI)**, the cost to have an item (section of text, markable, image, video, etc) completely annotated, calculated from the **Cost per Judgement (CpJ)** (the sum of all costs including setup, testing, maintenance, advertising, and indirect financial incentives divided by the total judgements) and the number of **Judgements Required (JR)** to complete an item (an empirically measured estimation based on the aggregation method used). **Cost per Acquisition (CpA)** is also a useful measure of effectiveness of promotions.

Basic measures of player engagement could be **Lifetime Judgements (LTJ), Average Judgements per Player (AJpP)** and **Average Lifetime Play (ALP)**. Other useful engagement metrics include **Monthly Active Users (MAU), Retention** and **Churn**, the definitions of which remain unvaried from the definitions used in F2P.

Throughput (the number of completely annotated items produced per hour) is perhaps the most useful headline metric to compare overall GWAP performance, not only with other games but also other annotation approaches in NLP, such as traditional expert annotation (Poesio et al., 2013) or microwork crowdsourcing using Mechanical Turk.[1] See Table 1 for a summary of proposed metrics for GWAPs for NLP.

Acknowledgments

This research was supported by the DALI project, funded by the European Research Council (ERC), Grant agreement ID: 695662.

References

D. Dziedzic. 2016. Use of the Free to Play model in games with a purpose: the RoboCorp game case study. *Bio-Algorithms and Med-Systems*, 12(4):187–197.

D. Jurgens and R. Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464.

M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. 2013. Phrase Detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):1–44, April.

Unity. 2014. Glossary of Metrics. `https://unity3d.com/learn/tutorials/topics/analytics/glossary-metrics`. [Last accessed 21 Feb 2017].

L. von Ahn and L. Dabbish. 2008. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67.

D. Xicota. 2017. Free to play and its key performance indicators. `http://www.gamasutra.com/blogs/DavidXicota/20140527/218550/Free_to_play_and_its_Key_Performance_Indicators.php`. [Last accessed 21 Feb 2017].

[1] https://www.mturk.com