PackPlay: Mining semantic data in collaborative games

Nathan Green
NC State University
890 Oval Drive
Raleigh, NC 27695

Paul Breimyer
NC State University
890 Oval Drive
Raleigh, NC 27695

Vinay Kumar
NC State University
890 Oval Drive
Raleigh, NC 27695

Nagiza F. Samatova
Oak Ridge National Lab
1 Bethel Valley Rd
Oak Ridge, TN 37831

Abstract

Building training data is labor-intensive and presents a major obstacle to advancing machine learning technologies such as machine translators, named entity recognizers (NER), part-of-speech taggers, etc. Training data are often specialized for a particular language or Natural Language Processing (NLP) task. Knowledge captured by a specific set of training data is not easily transferable, even to the same NLP task in another language. Emerging technologies, such as social networks and serious games, offer a unique opportunity to change how we construct training data.

While collaborative games have been used in information retrieval, it is an open issue whether users can contribute accurate annotations in a collaborative game context for a problem that requires an exact answer, such as games that would create named entity recognition training data. We present *PackPlay*, a collaborative game framework that empirically shows players' ability to mimic annotation accuracy and thoroughness seen in gold standard annotated corpora.

1 Introduction

Annotated corpora are sets of structured text used in Natural Language Processing (NLP) that contain supplemental knowledge, such as tagged parts-of-speech, semantic concepts assigned to phrases, or semantic relationships between these concepts. Machine Learning (ML) is a subfield of Artificial Intelligence that studies how computers can obtain knowledge and create predictive models. These models require annotated corpora to learn rules and patterns. However, these annotated corpora must be manually curated for each domain or task, which is labor intensive and tedious (Scannell, 2007), thereby creating a bottleneck for advancing ML and NLP prediction tools. Furthermore, knowledge captured by a specific annotated corpus is often not transferable to another task, even to the same NLP task in another language. Domain and language specific corpora are useful for many language technology applications, including named entity recognition (NER), machine translation, spelling correction, and machine-readable dictionaries. The An Crúbadán Project, for example, has succeeded in creating corpora for more than 400 of the world's 6000+ languages by Web crawling. With a few exceptions, most of the 400+ corpora, however, lack any linguistic annotations due to the limitations of annotation tools (Rayson et al., 2006).

Despite the many documented advantages of annotated data over raw data (Granger and Rayson, 1998; Mair, 2005), there is a dearth of annotated corpora in many domains. The majority of previous corpus annotation efforts relied on manual annotation by domain experts, automated prediction tagging systems, and hybrid semi-automatic systems that used both approaches. While yielding high quality and enormously valuable corpora, manually annotating corpora can be prohibitively costly and time consuming. For example, the GENIA corpus contains 9,372 sentences, curated by five part-time annotators, one senior coordinator, and one junior coordinator over 1.5 years (Kim et al., 2008). Semi-automatic approaches decrease human effort but often introduce significant error, while still requiring human interaction.

The Web can help facilitate semi-automatic approaches by connecting distributed human users at a previously unfathomable scale and presents an opportunity to expand annotation efforts to countless users using Human Computation, the concept of outsourcing certain computational

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 227–234,
Uppsala, Sweden, 15-16 July 2010.

processes to humans, generally to solve problems that are intractable or difficult for computers. This concept is demonstrated in our previous work, WebBANC (Green et al., 2009) and BioDEAL (Breimyer et al., 2009), which allows users to annotate Web documents through a Web browser plugin for the purposes of creating linguistically and biologically tagged annotated corpora and with micro-tasking via Mechanical Turk, which allows for a low cost option for manual labor tasks (Snow et al., 2008; Kittur et al., 2008).

While the Web and Human Computation may be a powerful tandem for generating data and solving difficult problems, in order to succeed, users must be motivated to participate. Humans have been fascinated with games for centuries and play them for many reasons, including for entertainment, honing skills, and gaining knowledge (FAS Summit, 2006). Every year, a large amount of hours are spent playing online computer games. The games range form simple card and word games to more complex 3-D world games. One such site for word, puzzle, and card games is Pogo.com[1]. According to protrackr,[2] Pogo has almost 6 million unique visitors a day. Alexa.com[3] shows that the average user is on the site for 11 minutes at a time. When the average time spent on the site is propagated to each user, the combined time is equal to more than 45,000 days of human time. Arguably if, the games on Pogo were used to harvest useful data, various fields of Computer Science research could be advanced.

There has been a recent trend to leverage human's fascination in game playing to solve difficult problems through *Human Computation*. Two such games include ESP and Google's Image Labeler (Ahn and Dabbish, 2004), in which players annotate images in a cooperative environment to correctly match image tags with their partner. Semantic annotation has also been addressed in the game Phrase Detectives (Chamberlain et al., 2009), which has the goal of creating large scale training data for anaphora resolution. These types of games are part of a larger, serious games, initiative (Annetta, 2008).

This paper introduces the Web-enabled collaborative game framework, PackPlay, and investigates how collaborative online gaming can affect annotation throughput and annotation accuracy. There are two main questions for such systems: first, will overall throughput increase compared to traditional methods of annotating, such as the manual construction of the Genia Corpus? Second, how accurate are the collective annotations? A successful human computation environment, such as PackPlay, would represent a paradigm shift in the way annotated corpora are created. However, adoption of such a framework cannot be expected until these questions are answered. We address both of these questions in multiple games in our PackPlay system through evaluation of the collective players' annotations with precision and recall to judge accuracy of players' annotations and the number of games played to judge throughput. We show improvements in both areas over traditional annotation methods and show accuracy comparable to expert prediction systems that could be used for semi-supervised annotation.

2 Methodology

We empirically show casual game players' ability to accurately and throughly annotate corpora by conducting experiments following the process described in Section 2.1 with 8 players using the PackPlay System. The testers annotate the datasets described in Section 2.2 and results are analyzed using the equations in Section 2.3.

2.1 PackPlay Process Flow

Figure 1 shows the average PackPlay process flow that a player will follow for a multi-player game. Assuming the player is registered, the player will always start by logging in and selecting the game he or she wants to play. Once in the game screen, the system will try to pair the player with another player who is waiting. After a set time limit, the game will automatically pair the user with a PlayerBot. It is important to note that the player will not know that his or her partner is a PlayerBot.

Once paired, a game can start. In most games, a question will be sampled from our database. How this sampling takes place is up to the individual game. Once sampled, the question will be displayed to one player or all players, depending on whether the game is synchronous or asynchronous (see definitions in Sections 3.1.2 and 3.2.2). Once the question is displayed, two things can happen.

[1] Pogo. http://www.pogo.com/

[2] Protrackr.com site information and statistics.http://www.protrackr.com/

[3] Alexa: The Web Information Company. http://www.alexa.com/

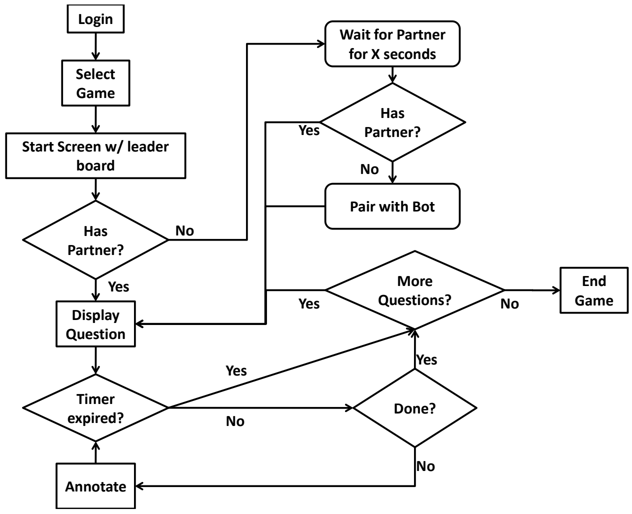


Figure 1: User process flow for PackPlay games.

First, the timer can run out; this timer is set by each game individually. Second, the player may answer the question and move on to the next question. After either one of those two options, a new question will be sampled. This cycle continues until the game session is over. This is usually determined by the game, as each game can set the number of questions in a session, or by a player quiting the game.

2.2 Data Sources

To compare named entity results, PackPlay uses sentences and annotations from CoNLL 2003, a "gold" standard corpus (Tjong et al., 2003). We use the CoNLL 2003 corpus since it has been curated by experts and the PackPlay system can compare our players' annotations vs those of 16 submitted predictive models, also refered to as the CoNLL average, in the 2003 conference on natural language learning. This paper will refer to the training corpus as the CoNLL corpus, and we selected it for our evaluation due to its widespread adoption as a benchmark corpus.

2.3 Metrics

To measure how thoroughly and accurately our players annotate the data, we calculate both recall (Equation 1) and precision (Equation 2), in which α is the set of words annotated in PackPlay and β is the set of words in the base CoNLL corpus.

$$Recall = \frac{|\alpha \cap \beta|}{|\beta|} \tag{1}$$

$$Precision = \frac{|\alpha \cap \beta|}{|\alpha|} \tag{2}$$

Each game module in the PackPlay system has its own scoring module, which is intended to improve the players' precision. For this reason, scoring is handled on a per game level. Each game has its own leader board as well. The leader board is used to motivate the players to continue playing the PackPlay games. This is intended to improve recall for annotations in the system.

3 Games

3.1 Entity Discovery

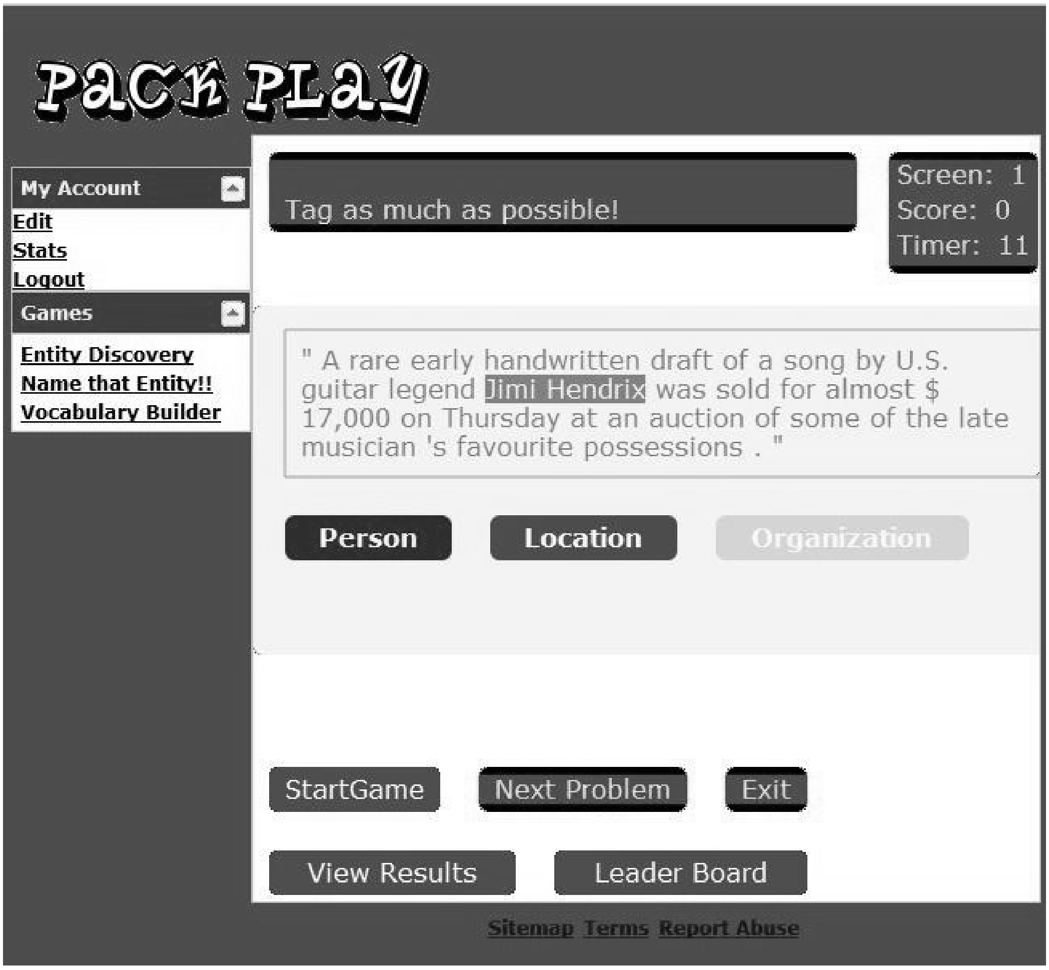
3.1.1 Game description

Named entities are a foundational part of many NLP systems from information extraction systems to machine translation systems. The ability to detect an entity is an application area called Named Entity Recognition (NER). The most common named entity categories are Person (Per), Location (Loc), and Organization (Org). The ability to extract these entities may be used in everyday work, such as extracting defendants, cities, and companies from court briefings, or it may be used for critical systems in national defense, such as monitoring communications for people and locations of interest.

To help with the creation of more NER systems, *Entity Discovery* (see Figure 2), a game for annotating sentences with supplied entities was created. The goal of the game is to pair players with each other and allow them to annotate sentences together. While this annotation task could be done by one person, it is a very time consuming activity. By creating a game, we hope that players will be more likely to annotate for fun and will annotate correctly and completely in order to receive a higher score in the PackPlay system.

3.1.2 Implementation

Entity Discovery is implemented as a *synchronous* two-player game. A *synchronous* game is one in which both players have the same task in the game, in this case, to annotate a sentence. To have a base comparison point, all players are asked to annotate a random set of 60 sentences to start, for which we have the correct answers. This way we will be able to assess the trustworthiness score in future iterations. After the pretest, the players will be shown sentences randomly sampled with replacement.



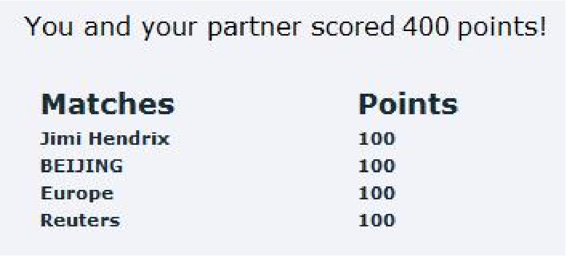
" A rare early handwritten draft of a song by U.S. guitar legend Jimi Hendrix was sold for almost $ 17,000 on Thursday at an auction of some of the late musician 's favourite possessions . "

Figure 2: Screenshot of a player annotating the Person entity Jimi Hendrix

In *Entity Discovery*, we made a design decision to keep a player's partner anonymous. This should help reduce cheating, such as agreeing to select the same word over and over, and it should reduce the ability for a player to only play with his or her friends, which might enhance their ability to cheat by using other communication systems such as instant messaging or a cell phone. Since PackPlay is still in the experimental stages, players may not always be available. For this reason, we have implemented a PlayerBot system. The PlayerBot will mimic another player by selecting previously annotated phrases for a given sentence from the database. From the human players' point of view, nothing seems different.

Players are asked to annotate, or tag, as many entities as they can find in a sentence. Players are also told at the beginning of the game that they are paired with another user. Their goal is to annotate the same things as their partner. Our assumption is that if the game is a single player game then the players may just annotate the most obvious entities for gaining more points. By having the player to try to guess at what their partner may annotate we hope to get better overall coverage of entities. We try to minimize the errors, which guessing might produce, in a second game, *Name That Entity* (Section 3.2).

To annotate a sentence, the player simply highlights a word or phrase and clicks on a relevant entity. For instance in *Entity Discovery*, a player can annotate the phrase "Jimi Hendrix" as a Person entity. From this point on, the player is free to annotate more phrases in the sentence. When the player completes annotating a sentence, the player hits "Next Problem." The system then waits for the player's partner to hit "Next Problem" as well. When both players finish annotating, the game points will be calculated and a new question will be sampled for the players.



You and your partner scored 400 points!

Matches	Points
Jimi Hendrix	100
BEIJING	100
Europe	100
Reuters	100

Figure 3: Screenshot of what the player sees at the end of the *Entity Discovery* game

3.1.3 Scoring

Scoring can be done in a variety of ways, each having an impact on players' performance and enjoyment. For *Entity Discovery*, we decided to give each user a flat score of 100 points for every answer that matched their partner. At the end of each game session, the player will see what answers matched with their partner. For instance, if both players tagged "Jimi Hendrix" as a Person, they will both receive 100 points. We do not show the players their matched scores after each sentence, since this might bias the user to tag more or less depending on what their partner does. Figure 3 shows a typical scoring screen at the end of a game; in Figure 3, the players matched 4 phrases, totaling 400 points. It is important to note that at this stage we do not distinguish between correct and incorrect annotations, just whether the two players agree.

3.1.4 User Case Study Methodology

To examine *Entity Discovery* as a collaborative game toward the creation of an annotated corpus, we conducted a user experiment to collect sample data on a known data set. Over a short time, 8 players were asked to play both *Entity Discovery* and *Name That Entity*. In PackPlay, throughput can be estimated, since each game has a defined time limit, defined as the average number of entities annotated per question times the number of users times the average number of questions seen by a user. Unlike other systems such as Mechanical Turk (Snow et al., 2008; Kittur et al., 2008), BioDeal (Breimyer et al., 2009), or WebBANC (Green et al., 2009), in PackPlay we define the speed at which a user annotates.

Each game in *Entity Discovery* consists of 10 sentences from the CoNLL corpus. These sentences are not guaranteed to have a named entity within them. The users in the study were not informed of the entity content as to not bias the experiment and falsely raise our precision scores. With only 8 players, we obtained 291 annotations, which averaged to about 40 annotations per user. This study was not done over a long period of time, so each user only played, on average, 3.6 games.

Table 1: Statistics returned from our user study for the game *Entity Discovery*

Statistic	Total	Mean
# of games	29	3.62
# of annotations	291	40.85

Two players were asked to intentionally annotate poorly. The goal of using poor annotators was to simulate real world players, who may just click answers to ruin the game or who are clueless to what a named entity is. This information can be used in later research to help automatically detect "bad" annotators using anomaly detection techniques.

PackPlay also stores information not used in this study, such as time stamps for each question answered. This information will be incorporated into future experiment analysis to see if we can further improve our annotated corpora based on the order and time spent forming an annotation. For instance, the first annotation in a sentence may have a higher probability than the last annotation. It is possible that if a user answers too fast, the answer is likely an error.

3.1.5 Output Quality

Every player completes part of a 60 sentence pretest in which we know the answers. For each game, the questions are sampled without replacement but this does not carry over after a game. For instance, if a player finishes game 1, he or she will never see the same question twice. For game two, no question within the game will be repeated, however, the player might see a question he or she answered in game 1. Because of this, each user will not see all 60 questions, but we will have a good sample to judge whether a user is accurate or not. The ability to repeat a question in different games allows us, in future research, to test players using intra-annotator agreement statistics. This tests how well a player agrees with himself or herself. From this set of 60 questions we have calculated each player's recall and precision scores.

As Table 2 shows, the recall scores for *Entity*

Table 2: Recall and precision for *Entity Discovery* annotations of CoNLL data.

	Per	Loc	Org	Avg	CoNLL Avg
Recall (All Data)	0.94	0.95	0.85	0.9	0.82
Precision (All Data)	0.47	0.70	0.53	0.62	0.83

Discovery in this experiment were 0.94, 0.95, and 0.85 for Person, Location, and Organization, respectively. The overall average was 0.9, which beats out the CoNLL average, an average of 16 expert systems, for recall. *Entity Discovery*'s numbers are similar to the pattern seen in the CoNLL predictive systems for Person, Location and Organization, in which Organization was the lowest and Person was the highest. The precision numbers were quite lower, with an average of 0.62. When examining the data, most of the precision errors occurred because of word phrase boundary issues with the annotation and also players often are unsure whether to include titles such as President, Mr., or Dr. There were also quite a few errors where players annotated concepts as People such as "The Judge" or "The scorekeeper." While this is incorrect for named entity recognition, it might be of interest to a co-reference resolution corpus. The precision numbers are likely low because of our untrained players and because some of the players were told to intentionally annotate entities incorrectly. To improve on these numbers, we applied a coverage requirement and majority voting. The coverage requirement requires that more than one player has annotated a given phrase for the annotation to be included in the corpus. Majority voting indicates that the phrase is only included if 50% or more of the playerss who annotated a phrase, agreed on the specific entity assigned to the phrase.

As Table 3 shows, both majority voting and coverage requirements improve precision by more than 10%. When combined, they improve the overall precision to 0.88, a 26% improvement. This is an improvement to the expert CoNLL systems score of 0.83. The majority voting likely removed the annotations from our purposefully "bad" annotators.

For future work, as the number of players increases, we will have to increase our coverage requirement to match. This ratio has not been determined and will need to be tested. A more successful way to detect errors in our annotations may be to create a separate game to verify given answers. To initially test this concept we have made and set up an experiment with a game, called *Name That Entity*.

Table 3: Precision for *Entity Discovery* annotations of CoNLL data with filtering

	Per	Loc	Org	Avg
Precision (Majority Voting)	0.56	0.79	0.65	0.72
Precision (Coverage Req.)	0.69	0.83	0.63	0.73
Precision (Majority Voting + Coverage Req.)	0.90	0.95	0.81	0.88

3.2 Name That Entity

3.2.1 Game Description

Name That Entity is another game with a focus on named entities. *Name That Entity* was created to show that game mechanics and the creation of further games would enhance the value of an annotated corpus. In the case of *Name That Entity*, we have created a multiple choice game in which the player will select the entity that best represents the highlighted word or phrase. Unlike *Entity Discovery*, this allows us to focus the annotation effort on particular words or phrases. Once again, this is modeled as a two-player game but the players are not playing simultaneously. The goal for the player is to select the same entity type for the highlighted word that their partner selects. In this game, speed is of the essence since each question will ask for one entity as opposed to *Entity Discovery*, which was open ended to how many entities might exist in a sentence.

3.2.2 Implementation

As described above, *Name That Entity* appears to be a two-player *synchronous* game. The player is under the assumption that he or she must once again match his or her partner's choice. What the player does not know is that the multi-player is simulated in this case. The player is replaced with a PlayerBot which chooses annotations from the *Entity Discovery* game. This, in essence, creates an *asynchronous* game, in which one player has the task of finding entities and the other player has the task of verifying entities. This gives us a further mechanism to check the validity of entities annotated by the *Entity Discovery* game.

As with *Entity Discovery*, the player's partner is anonymous. This anonymity allows us to keep the asynchronous structure hidden, as well as judge a new metric, intra-annotator agreement, not tested in the previous game. Since it is possible that a player in PackPlay may have a question sampled that was previously annotated in the *Entity Discovery* game by the same player, we can use intra-annotator agreement. While well-known inter-annotator statistics, such as Cohen's Kappa, evaluate one annotator versus the other annotator, intra-annotator statistics allow us to judge an annotator versus himself or herself to test for consistency (Artstein and Poesio, 2008). In the PackPlay framework this allows us to detect playerss who are randomly guessing and are therefore not consistent with themselves.

3.2.3 Scoring

Since entity coverage of a sentence is not an issue in the multiple choice game, we made use of a different scoring system that would reward first instincts. While the *Entity Discovery* game has a set score for every answer, *Name That Entity* has a sliding scale. For each question, the max score is 100 points, as the time ticks away the user receives fewer points. The points remaining are indicated to the user via a timing bar at the bottom of the screen.

When the player completes a game, he or she is allowed to view the results for that game. Unlike the *Entity Discovery* game, we display to the player what entity his or her partner chooses on the question in which they both did not match. This gives us a quick and simple form of annotator training, since a player with no experience may not be familiar with a particular entity. This was seen with the players' ability to detect an Organization entity. We expect that when a player sees what his or her partner annotates a phrase as, the player, is, in effect, being trained. However, displaying this at the end should not have any affects toward cheating since their partners are anonymous.

3.2.4 User Case Study Methodology

Of the 8 players who participated in the *Entity Discovery* study, 7 also played *Name That Entity* during their game sessions. We did not inform the players, but the questions asked in *Name That Entity* were the same answers that the players gave in the experiment in Section 3.1.4. The basic annotation numbers from our small user study can be seen in Table 4.

Table 4: Statistics returned from our user study for the game *Name That Entity*

Statistic	Total	Mean
# of games	20	2.85
# of annotations	195	27.85

3.2.5 Output Quality

As *Name That Entity* is not intended to be a solo mechanism to generate annotations, but instead a way to verify existing annotations, we did not assess the recall and precision of the game. Instead we are looking at the number of annotations, unique annotations, and conflicting annotations generated by our players in this game.

Table 5: Types of annotations generated by *Name That Entity*

Error	Count
Annotations	195
Unique Annotations	141
Conflicts	38
Unique Conflicts	35

In Table 5, unique annotations refer to annotations verified by only one user. Of the 195 total verified annotation, 38 had conflicting answers. In the majority of the cases the players marked these conflicts as "None of the Above," indicating that the annotated phrase from *Entity Discovery* was incorrect. For instance, many players made the mistake in *Entity Discovery* of marking phrases such as "German," "English," and "French" as Location entities when they are, in fact, just adjectives. In *Name That Entity*, the majority of players corrected each other and marked these as "None of the Above."

The main use of this game will be to incorporate it as an accuracy check for players based on these conflicting annotation. This accuracy check will be used in future work to deal with user confidence scores and conflict resolution.

4 Conclusion

Annotated corpora generation presents a major obstacle to advancing modern Natural Language Processing technologies. In this paper we introduced the PackPlay framework, which aims to leverage a distributed web user community in a collaborative game to build semantically-rich annotated corpora from players annotations. PackPlay is shown to have high precision and recall numbers when compared to expert systems in the area of named entity recognition. These annotated corpora were generated from two collaborative games in PackPlay, *Entity Discovery* and *Name That Entity*. The two games combined let us exploit the benefits of both *synchronous* and *asynchronous* gameplay as mechanisms to verify the quality of our annotated corpora. Future work should combine the players output with a player confidence score based on conflict resolution algorithms, using both inter- and intra-annotator metrics.

References

Luis von Ahn and Laura Dabbish. 2004 Labeling images with a computer game. ACM, pages 319-326, Vienna, Austria.

Leonard A. Annetta. 2008 Serious Educational Games: From Theory to Practice. Sense Publishers.

Ron Artstein and Massimo Poesio. 2008 Inter-coder agreement for computational linguistics. Computational Linguistics, Vol. 34, Issue 4, pages 555-596.

Maged N. Kamel Boulos and Steve Wheeler. 2007. The emerging web 2.0 social software: an enabling suite of sociable technologies in health and health care education. Health information and libraries journal, Vol. 24, pages 223.

Paul Breimyer, Nathan Green, Vinay Kumar, and Nagiza F. Samatova. 2009. BioDEAL: community generation of biological annotations. BMC Medical Informatics and Decision Making, Vol. 9, pages Suppl+1.

Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2009. Constructing an anaphorically annotated corpus with non-experts: assessing the quality of collaborative annotations. People's Web '09: Proceedings of the 2009 Workshop on The People's Web Meets NLP, pages 57-62.

FAS Summit on educational games: Harnessing the power of video games for learning (report), 2006.

Sylviane Granger and Paul Rayson. 1998. Learner English on Computer. Longman, London, and New Yorks pp. 119-131.

Nathan Green, Paul Breimyer, Vinay Kumar, and Nagiza F. Samatova. 2009. WebBANC: Building Semantically-Rich Annotated Corpora from Web User Annotations of Minority Languages. Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA), Vol. 4, pages 48-56, Odense, Denmark.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics, 9:10.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pages 453-456, Florence, Italy.

Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2006 Structure and evolution of online social networks. KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 611-617, New York, NY.

C. Mair. 2005. The corpus-based study of language change in progress: The extra value of tagged corpora. The AAACL/ICAME Conference, Ann Arbor, MI.

Paul Rayson, James Walkerdine,William H. Fletcher, and Adam Kilgarriff. 2006. Annotated web as corpus The 2nd International Workshop on Web as Corpus (EACL06), Trento, Italy.

Kevin P. Scannell. 2007. The Crbadn Project: Corpus building for under-resourced languages. Proceedings of the 3rd Web as Corpus Workshop Louvain-la-Neuve, Belgium.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast – but is it good?: evaluating non-expert annotations for natural language tasks EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 254–263, Honolulu, Hawaii.

Erik F. Tjong, Kim Sang and Fien De Meulder 2003 Introduction to the conll-2003 shared task: language-independent named entity recognition. Association for Computational Linguistics, pages 142-147, Edmonton, Canada.