Validating and Extending Semantic Knowledge Bases
using Video Games with a Purpose

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
`surname@di.uniroma1.it`

Abstract

Large-scale knowledge bases are important assets in NLP. Frequently, such resources are constructed through automatic mergers of complementary resources, such as WordNet and Wikipedia. However, manually validating these resources is prohibitively expensive, even when using methods such as crowdsourcing. We propose a cost-effective method of validating and extending knowledge bases using *video games* with a purpose. Two video games were created to validate concept-concept and concept-image relations. In experiments comparing with crowdsourcing, we show that video game-based validation consistently leads to higher-quality annotations, even when players are not compensated.

1 Introduction

Large-scale knowledge bases are an essential component of many approaches in Natural Language Processing (NLP). Semantic knowledge bases such as WordNet (Fellbaum, 1998), YAGO (Suchanek et al., 2007), and BabelNet (Navigli and Ponzetto, 2010) provide ontological structure that enables a wide range of tasks, such as measuring semantic relatedness (Budanitsky and Hirst, 2006) and similarity (Pilehvar et al., 2013), paraphrasing (Kauchak and Barzilay, 2006), and word sense disambiguation (Navigli and Ponzetto, 2012; Moro et al., 2014). Furthermore, such knowledge bases are essential for building unsupervised algorithms when training data is sparse or unavailable. However, constructing and updating semantic knowledge bases is often limited by the significant time and human resources required.

Recent approaches have attempted to build or extend these knowledge bases automatically. For example, Snow et al. (2006) and Navigli (2005)

extend WordNet using distributional or structural features to identify novel semantic connections between concepts. The recent advent of large semi-structured resources has enabled the creation of new semantic knowledge bases (Medelyan et al., 2009; Hovy et al., 2013) through automatically merging WordNet and Wikipedia (Suchanek et al., 2007; Navigli and Ponzetto, 2010; Niemann and Gurevych, 2011). While these automatic approaches offer the scale needed for open-domain applications, the automatic processes often introduce errors, which can prove detrimental to downstream applications. To overcome issues from fully-automatic construction methods, several works have proposed validating or extending knowledge bases using crowdsourcing (Biemann and Nygaard, 2010; Eom et al., 2012; Sarasua et al., 2012). However, these methods, too, are limited by the resources required for acquiring large numbers of responses.

In this paper, we propose validating and extending semantic knowledge bases using **video games with a purpose**. Here, the annotation tasks are transformed into elements of a video game where players accomplish their jobs by virtue of playing the game, rather than by performing a more traditional annotation task. While prior efforts in NLP have incorporated games for performing annotation and validation (Siorpaes and Hepp, 2008b; Herdağdelen and Baroni, 2012; Poesio et al., 2013), these games have largely been text-based, adding game-like features such as high-scores on top of an existing annotation task. In contrast, we introduce two video games with graphical 2D gameplay that is similar to what game players are familiar with. The fun nature of the games provides an intrinsic motivation for players to keep playing, which can increase the quality of their work and lower the cost per annotation.

Our work provides the following three contributions. First, we demonstrate effective video game-based methods for both validating and extending

semantic networks, using two games that operate on complementary sources of information: semantic relations and sense-image mappings. In contrast to previous work, the annotation quality is determined in a fully automatic way. Second, we demonstrate that converting games with a purpose into more traditional video games creates an increased player incentive such that players annotate for free, thereby significantly lowering annotation costs below that of crowdsourcing. Third, for both games, we show that games produce better quality annotations than crowdsourcing.

2 Related Work

Multiple works have proposed linguistic annotation-based games with a purpose for tasks such as anaphora resolution (Hladká et al., 2009; Poesio et al., 2013), paraphrasing (Chklovski and Gil, 2005), term associations (Artignan et al., 2009; Lafourcade and Joubert, 2010), query expansion (Simko et al., 2011), and word sense disambiguation (Chklovski and Mihalcea, 2002; Seemakurty et al., 2010; Venhuizen et al., 2013). Notably, all of these linguistic games focus on users interacting with text, in contrast to other highly successful games with a purpose in other domains, such as Foldit (Cooper et al., 2010), in which players fold protein sequences, and the ESP game (von Ahn and Dabbish, 2004), where players label images with words.

Most similar to our work are games that create or validate common sense knowledge. Two games with a purpose have incorporated video game-like mechanics for annotation. First, Herdağdelen and Baroni (2012) validate automatically acquired common sense relations using a slot machine game where players must identify valid relations and arguments from randomly aligned data within a time limit. Although the validation is embedded in a game-like setting, players are limited to one action (pulling the lever) unlike our games, which feature a variety of actions and rich gameplay experience to keep players interested longer. Second, Kuo et al. (2009) describe a pet-raising game where players must answer common sense questions in order to obtain pet food. While their game is among the most video game-like, the annotation task is a chore the player must perform in order to return to the game, rather than an integrated, fun part of the game's objectives, which potentially decreases motivation for answering correctly.

Several works have proposed adapting existing word-based board game designs to create or val-idate common sense knowledge. von Ahn et al. (2006) generate common sense facts by using a game similar to Taboo[TM], where one player must list facts about a computer-selected lemma and a second player must guess the original lemma having seen only the facts. Similarly, Vickrey et al. (2008) gather free associations to a target word with the constraint, similar to Taboo[TM], where players cannot enter a small set of banned words. Vickrey et al. (2008) also present two games similar to the Scattergories[TM], where players are given a category and then must list things in that category. The two variants differ in the constraints imposed on the players, such as beginning all items with a specific letter. For all three games, two players play the same game under time limits and then are rewarded if their answers match.

Last, three two-player games have focused on validating and extending knowledge bases. Rzeniewicz and Szymański (2013) extend Word-Net with common-sense knowledge using a 20 Questions-like game. In a rapid-play style game, OntoPronto attempts to classify Wikipedia pages as either categories or individuals (Siorpaes and Hepp, 2008a). SpotTheLink uses a similar rapid question format to have players align the DBpedia and PROTON ontologies by agreeing on the distinctions between classes (Thaler et al., 2011).

Unlike dynamic gaming elements common in our video games, the above games are all focused on interacting with textual items. Another major limitation is their need for always having two players, which requires them to sustain enough interest to always maintain an active pool of players. While the computer can potentially act as a second player, such a simulated player is often limited to using preexisting knowledge or responses, which makes it difficult to validate new types of entities or create novel answers. In contrast, we drop this requirement thanks to a new strategy for assigning confidence scores to the annotations based on negative associations.

3 Video Game with a Purpose Design

To create video games, our development process focused on a common design philosophy and a common data set.

3.1 Design Objectives

Three design objectives were used to develop the video games. First, the annotation task should be a central and natural action with familiar video game mechanics. That is, the annotation should

be supplied by common actions such as collecting items, puzzles, or destroying objects, rather than through extrinsic tasks that players must complete in order to return to the game. This design has the benefits of (1) growing the annotator pool with video games players, and (2) potentially increasing annotator enjoyment.

Second, the game should be playable by a single player, with reinforcement for correct game play coming from gold standard examples.[1] We note that gold standard examples may come from both true positive and true negative items.

Third, the game design should be sufficiently general to annotate a variety of linguistic phenomena, such that only the game data need be changed to accomplish a different annotation task. While some complex linguistic annotation tasks such as preposition attachment may be difficult to integrate directly into gameplay, many simpler but still necessary annotation tasks such as word and image associations can be easily modeled with traditional video game mechanics.

3.2 Annotation Setup

Tasks We focused on two annotation tasks: (1) validating associations between two concepts, and (2) validating associations between a concept and an image. For each task we developed a video game with a purpose that integrates the task within the game, as illustrated in Sections 4 and 5.

Knowledge base As the reference knowledge base, we chose BabelNet[2] (Navigli and Ponzetto, 2010), a large-scale multilingual semantic ontology created by automatically merging WordNet with other collaboratively-constructed resources such as Wikipedia and OmegaWiki. BabelNet data offers two necessary features for generating the games' datasets. First, by connecting WordNet synsets to Wikipedia pages, most synsets are associated with a set of pictures; while often noisy, these pictures sometimes illustrate the target concept and are an ideal case for validation. Second, BabelNet contains the semantic relations from both WordNet and hyperlinks in Wikipedia; these relations are again an ideal case of validation, as not all hyperlinks connect semantically-related pages in Wikipedia. Last, we stress that while our games use BabelNet data, they could easily validate or extend other knowledge bases such as YAGO (Suchanek et al., 2007) as well.

Data We created a common set of concepts, C, used in both games, containing sixty synsets selected from all BabelNet synsets with at least fifty associated images. Using the same set of synsets, separate datasets were created for the two validation tasks. In each dataset, a concept $c \in C$ is associated with two sets: a set V_c containing items to validate, and a set N_c with examples of true negative items (i.e., items where the relation to c does not hold). We use the notation V and N when referring to the to-validate and true negative sets for all concepts in a dataset, respectively.

For the concept-concept dataset, V_c is the union of V_c^B, which contains the lemmas of all synsets incident to c in BabelNet, and V_c^n, which contains novel lemmas derived from statistical associations. Specifically, novel lemmas were selected by computing the χ^2 statistic for co-occurrences between the lemmas of c and all other part of speech-tagged lemmas in Wikipedia. The 30 lemmas with the highest χ^2 are included in V_c. To enable concept-to-concept annotations, we disambiguate novel lemmas using a simple heuristic based on link co-occurrence count (Navigli and Ponzetto, 2012). Each set V_c contains 77.6 lemmas on average.

For the concept-image data, V_c is the union of V_c^B, which contains all images associated with c in BabelNet, and V_c^n, which contains web-gathered images using a lemma of c as the query. Web-gathered images were retrieved using Yahoo! Boss image search and the first result set (35 images) was added to V_c. Each set V_c contains 77.0 images on average.

For both datasets, each negative set N_c is constructed as $\cup_{c' \in C \setminus \{c\}} V_{c'}^B$, i.e., from the items related in BabelNet to all other concepts in C. By constructing N_c directly from the knowledge base, play actions may be validated based on recognition of true negatives, removing the heavy burden for ever manually creating a gold standard test set.

Annotation Aggregation In each game, an item is annotated when players make a binary choice as to whether the item's relation is true (e.g., whether an image is related to a concept). To produce a final annotation, a rating of $p - n$ is computed, where p and n denote the number of times players have marked the item's relation as true or false, respectively. Items with a positive rating after aggregating are marked as true examples of the relation and false otherwise.
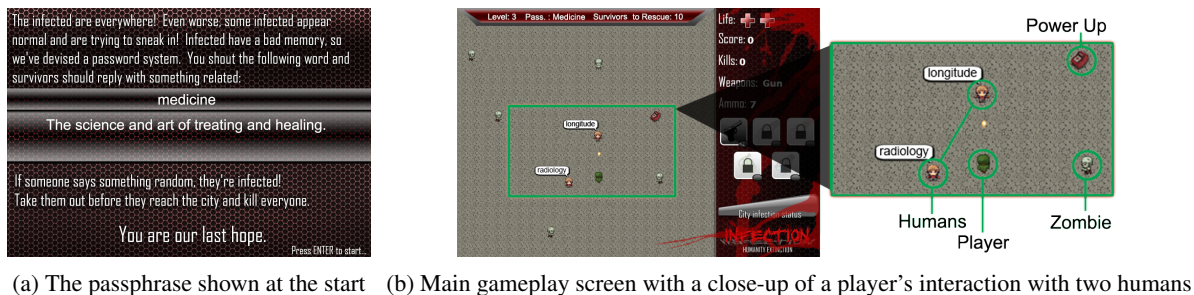
---

[1] This design is in contrast to two-player games where mutual agreement reinforces correct behavior.

[2] http://babelnet.org

(a) The passphrase shown at the start   (b) Main gameplay screen with a close-up of a player's interaction with two humans

Figure 1: Screenshots of the key elements of Infection

4   Game 1: Infection

The first game, *Infection*, validates the concept-concept relation dataset.

Design   Infection is designed as a top-down shooter game in the style of Commando. Infection features the classic game premise that a virus has partially infected humanity, turning people into zombies. The player's responsibility is to stop zombies from reaching the city and rescue humans that are fleeing to the city. Both zombies and humans appear at the top of the screen, advance to the bottom and, upon reaching it, enter the city.

In the game, some humans are infected, but have not yet become zombies; these infected humans must be stopped before reaching the city. Because infected and uninfected humans look identical, the player uses a passphrase call-and-response mechanism to distinguish between the two. Each level features a randomly-chosen passphrase that the player's character shouts. Uninfected humans are expected to respond with a word or phrase related to the passphrase; in contrast, infected humans have become confused due to the infection and will say something completely unrelated in an attempt to sneak past. When an infected human reaches the city, the city's total infection level increases; should the infection level increase beyond a certain threshold, the player fails the stage and must replay it to advance the game. Furthermore, if any time after ten humans have been seen, the player has killed more than 80% of the uninfected humans, the player's gun is taken by the survivors and she loses the stage.

Figure 1a shows instructions for the passphrase "medicine." In the corresponding gameplay, shown in the close up of Figure 1b, a human shouts a valid response, "radiology" for the level's passphrase, while the nearby infected human shouts an incorrect response "longitude."

Gameplay is divided into eight stages, each with increasing difficulty. Each stage has a goal of saving a specific number of uninfected humans. Infection incorporates common game mechanics, such as unlockable weapons, power-ups that restore health, and achievements. Scoring is based on both the number of zombies killed and the percentage of uninfected humans saved, motivating players to kill infected humans in order to increase their score. Importantly, Infection also includes a leaderboard where players compete for top positions based on their total scores.

Annotation   Each human is assigned a response selected uniformly from V or N. Humans with responses from N are treated as infected. Players annotate by selecting which humans are infected: Allowing a human with a response from V to enter the city is treated as a positive annotation; killing that human is treated as a negative annotation.

The design of Infection enables annotating multiple types of conceptual relations such as synonymy or antonymy by changing only the description of the passphrase and how uninfected humans are expected to respond.

Quality Enforcement Mechanisms   Infection includes two game mechanics to limit adversarial players from creating many low quality annotations. Specifically, the game prevents players from both (1) allowing all humans to live, via the city infection level and (2) killing all humans, via survivors taking the player's gun; these actions would both generate many false positives and false negatives, respectively. These mechanics ensure the game naturally produces better quality annotations; in contrast, common crowdsourcing platforms do not support analogous mechanics for enforcing this type of correctness at annotation time.

5   Game 2: The Knowledge Towers

The second game, *The Knowledge Towers* (TKT), validates the concept-image dataset.

Design   TKT is designed as a single-player role playing game (RPG) where the player explores a

|(a) An example tower's concept | (b) Image selection screen | (c) Gameplay |

Figure 2: Screenshots of the key elements of The Knowledge Towers.

series of towers to unlock long-forgotten knowledge. At the start of each tower, a target concept is shown, e.g., the tower of "tango," along with a description of the concept (Figure 2a). The player must then recover the knowledge of the target concept by acquiring pictures of it. Pictures are obtained through defeating monsters and opening treasure chests, such as those shown in Figure 2c. However, players must distinguish pictures of the tower's concept from unrelated pictures. When an image is picked up, the player may keep or discard it, as shown in Figure 2b. A player's inventory is limited to eight pictures to encourage them to select the most relevant pictures only.

Once the player has collected enough pictures, the door to the boss room is unlocked and the player may enter to defeat the boss and complete the tower. Pictures may also be deposited in special reward chests that grant experience bonuses if the deposited pictures are from V. Gathering unrelated pictures has adverse effects on the player. If the player finishes the level with a majority of unrelated pictures, the player's journey is unsuccessful and she must replay the tower.

TKT includes RPG game elements commonly found in game series such as Diablo and the Legend of Zelda: players begin with a specific character class that has class-specific skills, such as Warrior or Thief, but will unlock the ability to play as other classes by successfully completing the towers. Last, TKT includes a leaderboard where players can compete for positions; a player's score is based on increasing her character's abilities and her accuracy at discarding images from N.

Annotation Players annotate by deciding which images to keep in their inventory. Images receive positive rating annotations from: (1) depositing the image in a reward chest, and (2) ending the level with the image still in the inventory. Conversely, images receive a negative rating when a player (1) views the image but intentionally avoids picking it up or (2) drops the image from her inventory.

TKT is designed to assist in the validation and extension of automatically-created image libraries that link to semantic concepts, such as ImageNet (Deng et al., 2009) and that of Torralba et al. (2008). However, its general design allows for other types of annotations, such as image labeling, by changing the tower's instructions and pictures.

Quality Enforcement Mechanisms Similar to Infection, TKT includes analogous mechanisms for limiting adversarial player annotations. Players who collect no images are prevented from entering the boss room, limiting their ability to generate false negative annotations. Similarly, players who collect all images are likely to have half of their images from N and therefore fail the tower's quality-check after defeating the boss.

6 Experiments

Two experiments were performed with Infection and TKT: (1) an evaluation of players' ability to play accurately and to validate semantic relations and image associations and (2) a comprehensive cost comparison. Each experiment compared (a) free and financially-incentivized versions of each game, (b) crowdsourcing, and (c) a non-video game with a purpose.

6.1 Experimental Setup

Gold Standard Data To compare the quality of annotation from games and crowdsourcing, a gold standard annotation was produced for a 10% sample of each dataset (cf. Section 3.2). Two annotators independently rated the items and, in cases of disagreement, a third expert annotator adjudicated. Unlike in the game setting, annotators were free to consult additional resources such as Wikipedia.

To measure inter-annotator agreement (IAA) on the gold standard annotations, we calculated Krip-

pendorff's α (Krippendorff, 2004; Artstein and Poesio, 2008); α ranges between [-1,1] where 1 indicates complete agreement, -1 indicates systematic disagreement, and values near 0 indicate agreement at chance levels. Gold standard annotators had high agreement, 0.774, for concept-concept relations. However, image-concept agreement was only moderate, 0.549. A further analysis revealed differences in the annotators' thresholds for determining association, with one annotator permitting more abstract relations. However, the adjudication process resolved these disputes, resulting in substantial agreement by all annotators on the final gold annotations.

Incentives At the start of each game, players were shown brief descriptions of the game and a description of a contest where the top-ranked players would win either (1) monetary prizes in the form of gift cards, or (2) a mention and thanks in this paper. We refer to these as the paid and free versions of the game, respectively. In the paid setting, the five top-ranking players were offered gift cards valued at 25, 15, 15, 10, and 10 USD, starting from first place (a total of 75 USD per game). To increase competition among players and to perform a fairer time comparison with crowdsourcing, the contest period was limited to two weeks.

6.2 Comparison Methods

To compare with the video games, items were annotated using two additional methods: crowdsourcing and a non-video game with a purpose.

Crowdsourcing Setup Crowdsourcing was performed using the CrowdFlower platform. Annotation tasks were designed to closely match each game's annotation process. A task begins with a description of a target synset and its textual definition; following, ten annotation questions are shown. Separate tasks were used for validating concept-concept and concept-image relations. Each tasks' questions were shown as a binary choice of whether the item is related to the task's concept. Workers were paid 0.05 USD per task. Each question was answered by three workers.

Following common practices for guarding against adversarial workers (Mason and Suri, 2012), the tasks for concept c include quality check questions using items from N_c. Workers who rate too many relations from N_c as valid are removed by CrowdFlower and prevented from participating further. One of the ten questions in a task used an item from N_c, resulting in a task mixture of 90% annotation questions and 10% quality-

check questions. However, we note that both of our video games use data that is 50% annotation, 50% quality-check. While the crowdsourcing task could be adjusted to use an increased number of quality-check options, such a design is uncommon and artificially inflates the cost of the crowdsourcing comparison beyond what would be expected. Therefore, although the crowdsourcing and game-based annotation tasks differ slightly, we chose to use the common setup in order to create a fair cost-comparison between the two.

Non-video Game with a Purpose To measure the impact of the video game itself on the annotation process, we developed a non-video game with a purpose, referred to as *SuchGame*. Players perform a single action in SuchGame: after being shown a concept c and its textual definition, a player answers whether an item is related to the concept. Items are drawn equally from V_c and N_c, with players scoring a point each time they select that an item from N is not related. A round of gameplay contains ten questions. After the round ends, players see their score for that round and the current leaderboard. Two versions of SuchGame were released, one for each dataset. SuchGame was promoted with same free recognition incentive as Infection and TKT.

6.3 Game Release

Both video games were released to multiple online forums, social media sites, and Facebook groups. SuchGame was released to separate Facebook groups promoting free webgames and groups for indie games. For each release, we estimated an upper-bound of the audience sizes using available statistics such as Facebook group sites, website analytics, and view counts. The free and paid versions had sizes of 21,546 and 14,842 people, respectively; SuchGame had an upper bound of 569,131 people. Notices promoting the game were separated so that audiences saw promotions for one of either the paid or free incentive version. Games were also released in such a way as to preserve the anonymity of the study, which limited our ability to advertise to public venues where the anonymity might be compromised.

7 Results and Discussion

7.1 Gameplay Analysis

In this section we analyze the games in terms of participation and player's ability to correctly play. Players completed over 1388 games during the
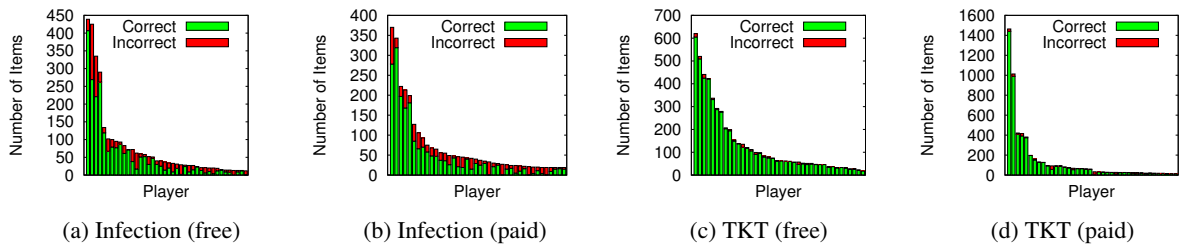
| (a) Infection (free) | (b) Infection (paid) | (c) TKT (free) | (d) TKT (paid) |

Figure 3: Accuracy of the top-40 players in rejecting true negative items during gameplay.

| | # Players | # Anno. | N-Acc. | Krip.'s α | G.S. Agreement | | | Cost per Ann. |
					True Pos.	True Neg.	All	
TKT free	100	3005	97.0	0.333	82.5	82.5	82.5	$0.000
TKT paid	97	3318	95.4	0.304	69.0	92.1	74.0	$0.023
Crowdflower	290	13854	-	0.478	59.5	93.7	66.2	$0.008
Infection free	89	3150	71.0	0.445	67.8	68.4	68.1	$0.000
Infection paid	163	3355	65.9	0.330	69.1	54.8	61.1	$0.022
Crowdflower	1097	13764	-	0.167	16.9	96.4	59.6	$0.008

Table 1: Annotation statistics from all sources. N-Accuracy denotes accuracy at rejecting items from N; G.S. Agreement denotes percentage agreement of the aggregated annotations with the gold standard.

study period. The paid and free versions of TKT had similar numbers of players, while the paid version of Infection attracted nearly twice the players compared to the free version, shown in Table 1, Column 1. However, both versions created approximately the same number of annotations, shown in Column 2. Surprisingly, SuchGame received little attention, with only a few players completing a full round of game play. We believe this emphasizes the strength of video game-based annotation; adding incentives and game-like features to an annotation task will not necessarily increase its appeal. Given SuchGame's minimal interest, we omit it from further analysis.

Second, the type of incentive did not change the percentage of items from N that players correctly reject, shown for all players as N-accuracy in Table 1 Column 3 and per-player in Figure 3. However, players were much more accurate at rejecting items from N in TKT than in Infection. We attribute this difference to the nature of the items and the format of the games. The images used by TKT provide concrete examples of a concept, which can be easily compared with the game's current concept; in addition, TKT allows players to inspect items as long as a player prefers. In contrast, concept-concept associations require more background knowledge to determine if a relation exists; furthermore, Infection gives players limited time to decide (due to board length) and also contains cognitive distractors (zombies). Neverthe-

less, player accuracy remains high for both games (Table 1, Col. 3) indicating the games represent a viable medium for making annotation decisions.

Last, the distribution of player annotation frequencies (Figure 3) suggests that the leaderboard and incentives motivated players. Especially in the paid condition, a clear group appears in the top five positions, which were advertised as receiving prizes. The close proximity of players in the paid positions is a result of continued competition as players jostled for higher-paying prizes.

7.2 Annotation Quality

This section assesses the annotation quality of both games and of CrowdFlower in terms of (1) the IAA of the participants, measured using Krippendorff's α, and (2) the percentage agreement of the resulting annotations with the gold standard. Players in both free and paid games had similar IAA, though the free version is consistently higher (Table 1, Col. 4).[3] For images, crowdsourcing workers have a higher IAA than game players; however, this increased agreement is due to adversarial workers consistently selecting the same, incorrect answer. In contrast, both video games contain mechanisms for limiting such behavior.

The strength of both crowdsourcing and games with a purpose comes from aggregating multiple annotations of a single item; i.e., while IAA may

---

[3] In conversations with players after the contest ended, several mentioned that being aware their play was contributing to research motivated them to play more accurately.
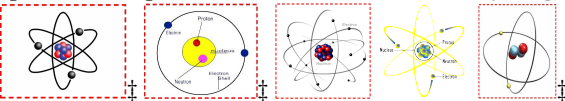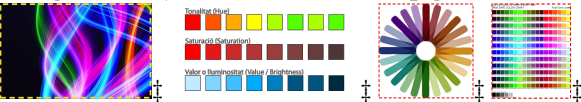
1300

Lemma	Abbreviated Definition	Most-selected Items
atom	The smallest possible particle of a chemical element	spectrum, nonparticulate radiation, **molecule**, **hydrogen**, **electron**
chord	A combination of three or more notes	**voicing**, **triad**, **tonality**,‡ **strum**, **note**, **harmony**
color	An attribute from reflected or emitted light	**orange**, brown,‡ video, sadness, **RGB**, **pigment**
fire	The state of combustion in which inflammable material burns	**sprinkler**, machine gun, chemical reduction, volcano, organic chemistry
religion	The expression of man's belief in and reverence for a super-human power	**polytheistic**,‡ **monotheistic**, **Jainism**, **Christianity**,‡ Freedom of religion

Table 2: Examples of the most-selected words and images from the free version of both games. Bolded words and images with a dashed border denote items not in BabelNet. Only the items marked with a ‡ were rated as valid in the aggregated CrowdFlower annotations.

be low, the majority annotation of an item may be correct. Therefore, in Table 1, we calculate the percentage agreement of the aggregated annotations with the gold standard annotations for approving valid relations (true positives; Col. 5), rejecting invalid relations (true negatives; Col. 6), and for both combined (Col. 7). On average, both video games in all settings produce more accurate annotations than crowdsourcing. Indeed, despite having lower IAA for images, the free version of TKT provides an absolute 16.3% improvement in gold standard agreement over crowdsourcing.

Examining the difference in annotation quality for true positives and negatives, we see a strong bias with crowdsourcing towards rejecting all items. This bias leads to annotations with few false positives, but as Column 5 shows, crowdflower workers consistently performed much worse than game players at identifying valid relations, producing many false negative annotations. Indeed, for concept-concept relations, workers identified only 16.9% of the valid relations.

In contrast to crowdsourcing, both games were effective at identifying valid relations. Table 2 shows examples of the most frequently chosen items from V for the free versions of both games. For both games, players were equally likely to select novel items, suggesting the games

can serve a useful purpose of adding these missing relations in automatically constructed knowledge bases. Highlighting one example, the five most selected concept-concept relations for *chord* were all novel; BabelNet included many relations to highly-specific concepts (e.g., "Circle of fifths") but did not include relations to more commonly-associated concepts, like *note* and *harmony*.

7.3 Cost Analysis

This section provides a cost-comparison between the video games and crowdsourcing. The free versions of both games proved highly successful, yielding high-quality annotations at no direct cost. Both free and paid conditions produced similar volumes of annotations, suggesting that players do not need financial incentives provided that the games are fun to play. It could be argued that the recognition incentive was motivating players in the free condition and thus some incentive was required. However, player behavior indicates otherwise: After the contest period ended, *no* players in the free setting registered for being acknowledged by name, which strongly suggests the incentive was not contributing to their motivation for playing. Furthermore, a minority of players continued to play even after the contest period ended, suggesting that enjoyment was a driving factor.

Last, while crowdsourcing has seen different quality and volume from workers in paid and unpaid settings (Rogstadius et al., 2011), in contrast, our games produced approximately-equivalent results from players in both settings.

Crowdsourcing was slightly more cost-effective than both games in the paid condition, as shown in Table 1, Column 8. However, three additional factors need to be considered. First, both games intentionally uniformly sample between V and N to increase player engagement,[4] which generates a larger number of annotations for items in N than are produced by crowdsourcing. When annotations on items in N are included for both games and crowdsourcing, the costs per annotation drop to comparable levels: $0.007 for CrowdFlower tasks, $0.008 for TKT, and $0.011 for Infection.

Second, for both annotation tasks, crowdsourcing produced lower quality annotations, especially for valid relations. Based on agreement with the gold standard (Table 1, Col. 5), the estimated cost for crowdsourcing a *correct* true positive annotation increases to $0.014 for a concept-image and a $0.048 for concepts-concept annotation. In contrast, the cost when using video games increases only to $0.033 for concept-image and $0.031 for concept-concept. These cost increases suggest that crowdsourcing is not always cheaper with respect to quality.

Third, we note that both video games in the paid setting incur a fixed cost (for the prizes) and therefore additional games played can only further decrease the cost per annotation. Indeed, the present study divided the audience pool into two separate groups which effectively halved the potential number of annotations per game. Assuming combining the audiences would produce the same number of annotations, both our games' costs per annotation drop to $0.012.

Last, video games can potentially come with indirect costs due to software development and maintenance. Indeed, Poesio et al. (2013) report spending 60,000£ in developing their Phrase Detectives game with a purpose over a two-year period. In contrast, both games here were developed as a part of student projects using open source software and assets and thus incurred no cost; furthermore, games were created in a few months, rather than years. Given that few online games attain significant sustained interest, we argue that

our lightweight model is preferable for producing video games with a purpose. While using students is not always possible, the development process is fast enough to sufficiently reduce costs below those reported for Phrase Detectives.

8 Conclusion

Two video games have been presented for validating and extending knowledge bases. The first game, Infection, validates concept-concept relations, and the second, The Knowledge Towers, validates image-concept relations. In experiments involving online players, we demonstrate three contributions. First, games were released in two conditions whereby players either saw financial incentives for playing or a personal satisfaction incentive where they were thanked by us. We demonstrated that both conditions produced nearly identical numbers of annotations and, moreover, that players were disinterested in the satisfaction incentive, suggesting they played out of interest in the game itself. Furthermore, we demonstrated the effectiveness of a novel design for games with a purpose which does not require two players for validation and instead reinforces behavior only using true negative items that required no manual annotation. Second, in a comparison with crowdsourcing, we demonstrate that video game-based annotations consistently generated higher-quality annotations. Last, we demonstrate that video game-based annotation can be more cost-effective than crowdsourcing or annotation tasks with game-like features: The significant number of annotations generated by the satisfaction incentive condition shows that a fun game can generate high-quality annotations at virtually no cost. All annotated resources, demos of the games, and a live version of the top-ranking items for each concept are currently available online.[5]

In the future we will apply our video games to the validation of more data, such as the new Wikipedia bitaxonomy (Flati et al., 2014).

---

[4]Earlier versions that used mostly items from V proved less engaging due to players frequently performing the same action, e.g., saving most humans or collecting most pictures.

---

[5]http://lcl.uniroma1.it/games/

References

Guillaume Artignan, Mountaz Hascoët, and Mathieu Lafourcade. 2009. Multiscale visual analysis of lexical networks. In *Proceedings of the International Conference on Information Visualisation*, pages 685–690.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Chris Biemann and Valerie Nygaard. 2010. Crowdsourcing wordnet. In *Proceedings of the 5th Global WordNet conference*.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Timothy Chklovski and Yolanda Gil. 2005. Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *Proceedings of the International Conference on Knowledge Capture*, pages 35–42. ACM.

Tim Chklovski and Rada Mihalcea. 2002. Building a Sense Tagged Corpus with Open Mind Word Expert. In *Proceedings of ACL 2002 Workshop on WSD: Recent Successes and Future Directions*, Philadelphia, PA, USA.

Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit players. 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Soojeong Eom, Markus Dickinson, and Graham Katz. 2012. Using semi-experts to derive judgments on word sense alignment: a pilot study. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 605–611.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2014. Two is bigger (and better) than one: the Wikipedia Bitaxonomy Project. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland.

Amaç Herdağdelen and Marco Baroni. 2012. Bootstrapping a game with a purpose for common sense collection. *ACM Transactions on Intelligent Systems and Technology*, 3(4):1–24.

Barbora Hladká, Jiří Mírovský, and Pavel Schlesinger. 2009. Play the language: Play coreference. In *Proceedings of the Joint Conference of the Association for Computational Linguistics and International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, pages 209–212. Association for Computational Linguistics.

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semistructured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 455–462.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, second edition.

Yen-ling Kuo, Jong-Chuan Lee, Kai-yang Chiang, Rex Wang, Edward Shen, Cheng-wei Chan, and Jane Yung-jen Hsu. 2009. Community-based game design: experiments on social games for commonsense data collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 15–22.

Mathieu Lafourcade and Alain Joubert. 2010. Computing trees of named word usages from a crowdsourced lexical network. In *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT)*, pages 439–446, Wisla, Poland.

Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on amazons mechanical turk. *Behavior Research Methods*, 44(1):1–23.

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Uppsala, Sweden*, pages 216–225.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Joining forces pays off: Multilingual Joint Word Sense Disambiguation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1399–1410, Jeju, Korea.

Roberto Navigli. 2005. Semi-automatic extension of large-scale linguistic knowledge bases. In *Proceedings of the 18th Internationa Florida AI Research Symposium Conference,* Clearwater Beach, Florida, 15–17 May 2005, pages 548–553.

Elisabeth Niemann and Iryna Gurevych. 2011. The people's web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, pages 205–214.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1341–1351, Sofia, Bulgaria.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):3:1–3:44, April.

Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Jacek Rzeniewicz and Julian Szymański. 2013. Bringing Common Sense to WordNet with a Word Game. In *Computational Collective Intelligence. Technologies and Applications*, volume 8083 of *Lecture Notes in Computer Science*, pages 296–305. Springer.

Cristina Sarasua, Elena Simperl, and Natalya F Noy. 2012. CrowdMap: Crowdsourcing ontology alignment with microtasks. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 525–541.

Nitin Seemakurty, Jonathan Chu, Luis Von Ahn, and Anthony Tomasic. 2010. Word sense disambiguation via human computation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 60–63. ACM.

Jakub Simko, Michal Tvarozek, and Maria Bielikova. 2011. Little search game: term network acquisition via a human computation game. In *Proceedings of the ACM conference on Hypertext and Hypermedia*, pages 57–62.

Katharina Siorpaes and Martin Hepp. 2008a. Games with a purpose for the semantic web. *IEEE Intelligent Systems*, 23(3):50–60.

Katharina Siorpaes and Martin Hepp. 2008b. Ontogame: Weaving the semantic web by online games. In Sean Bechhofer, Manfred Hauswirth, Jrg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web: Research and Applications*, volume 5021 of *Lecture Notes in Computer Science*, pages 751–766. Springer Berlin Heidelberg.

Rion Snow, Dan Jurafsky, and Andrew Ng. 2006. Semantic taxonomy induction from heterogeneous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL), Sydney, Australia*, pages 801–808.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge. unifying WordNet and Wikipedia. In *Proceedings of the 16th World Wide Web Conference,* Banff, Canada, 8–12 May 2007, pages 697–706.

Stefan Thaler, Elena Paslaru Bontas Simperl, and Katharina Siorpaes. 2011. SpotTheLink: A Game for Ontology Alignment. In *Proceedings of the 6th Conference on Professional Knowledge Management: From Knowledge to Action*, pages 246–253.

Antonio Torralba, Robert Fergus, and William T Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970.

Noortje J. Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In *Proceedings of the International Conference on Computational Semantics (IWCS)*.

David Vickrey, Aaron Bronzan, William Choi, Aman Kumar, Jason Turner-Maier, Arthur Wang, and Daphne Koller. 2008. Online word games for semantic data collection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 533–542.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 319–326.

Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 75–78.