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

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani and Roberto Navigli

SAPIENZA UNIVERSITÀ DI ROMA

Introduction

Large-scale knowledge bases are an essential component of many approaches in Natural Language Processing. The recent advent of large semi-structured resources has enabled the creation of new semantic knowledge bases such as BabelNet, which automatically merges WordNet and Wikipedia (Navigli and Ponzetto, 2010).

Problem: The automatic construction of resources can introduce errors. Correcting these errors using *crowdsourcing* is constrained by the available funds and time. Furthermore, crowdsourcing workers are often not invested or interested in the annotation effort.

Observation: People spend much of their time playing video games as a pastime, which they do for *free*. Games with a Purpose (von Ahn and Dabbish, 2004) have been shown to produce good annotations when players are engaged.

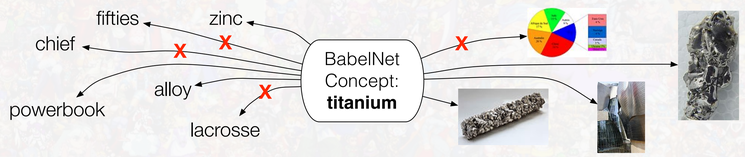
Solution: Fix the errors in automatically constructed knowledge bases by turning the annotation into a **video game**. The video games can potentially increase player motivation intrinsically and tap a new source of annotators who are willing to work for free if the game is fun.

Contributions: we demonstrate (1) effective video game-based methods for both validating and extending semantic knowledge bases, (2) converting games with a purpose into more traditional video games increases player incentive such that players annotate for free, and (3) the games produce better quality annotations than equivalent crowdsourcing

Experimental Setup

Knowledge Base: Games were created to validate and extend BabelNet, a merger of WordNet and Wikipedia. Two types of data were used for games: (1) **concept-concept** relations and (2) **concept-image** relations. Both types of data were generated for a set of sixty concepts. Novel images and concepts were gathered from web queries and corpus data.

BabelNet



Crowdsourcing setup: Workers on Crowdflower were shown the same data and descriptions as players and asked to choose whether two items were related. Separate tasks were run for each data type. Workers were paid $0.05 per task and three workers answered each question.

CrowdFlower

Game setup: Players were recruited from university computer science students. Two conditions were tested: (1) a **paid** condition that awarded small cash prizes to the best players after two weeks, and (2) a **free** condition with no reward. Students were shown only one type of advertisement and recruited equally for both conditions.

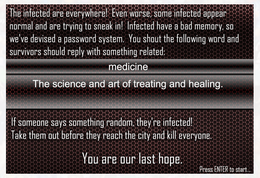
Validation Data: Players and crowdsourcing workers' accuracies were tested using *automatically-constructed* true negative data by pairing the target concept with a relation from a random concept in BabelNet.

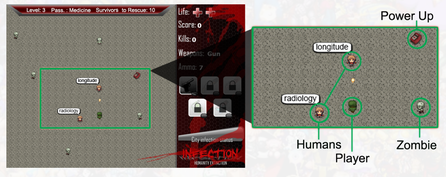
Gold Standard Data: Quality was tested using a manually-annotated 10% sample of all data. This manual effort was only needed for testing and not for game play.

Infection:

The zombie apocalypse has arrived! There are zombies everywhere but uninfected humans still need to be rescued. However, some infected humans are trying to sneak into the city. To counter them, we've setup a password system: you shout out a word and uninfected humans will shout something related back. If they shout anything random, they're infected. Stop them before they reach the camp! The future is counting on you.

Annotation Mechanism

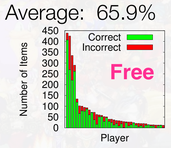




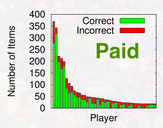
Summary: Players decide whether a human is infected based on what it says. Humans that are saved are counted as examples of a valid concept-concept relation.

Q1: How accurate are players at rejecting incorrect relations?

Average: 65.9%



71.0%



Players were highly accurate, detecting 66-71% of the incorrect relations while playing the game.

Q2: What is the quality of players compared to CrowdFlower?

	True Pos.	True Neg.	All
Infection free	67.8	68.4	68.1
Infection paid	69.1	54.8	61.1
CrowdFlower	16.9	96.4	59.6

Players were more accurate than workers at producing the same annotations as experts, even when working for free. Players were much better at recognizing valid relations.

Q3: How do Infection and CrowdFlower compare in cost?

	# Players	# Annotations	Cost per Annotation
Infection free	89	3150	$0.000
Infection paid	163	3355	$0.022
CrowdFlower	1097	13764	$0.008

Players played equally in both the free and paid conditions, showing that games can produce annotations at *no cost.*

Examples of validated and new Concept-Concept Relations

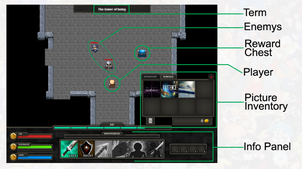
Term	Summarized Definition	Top-rated related terms
atom	The smallest possible particle	spectrum, nonparticulate radiation, molecule
chord	A combination of notes	**voicing**, **triad**, **tonality**, **strum**, **note**
color	An attribute of light	**orange**, brown, video, sadness, **RGB**
fire	The state of combusion	**sprinkler**, machine gun, chemical reduction
religion	An expression of man's beliefs	**polytheistic**, **monotheistic**, **Jainism**

Blue items are new relations; CrowdFlower workers only marked the underlined items as valid.

The Knowledge Towers:

In the distant past, powerful wizards have locked away humanity's knowledge into towers. These towers are now guarded by monsters, demons and fallen warriors. Even worse, the wizards have added false knowledge to confuse you. Now is the time for you, the hero, to reclaim the knowledge for humanity from each tower. Study the tower's secret and then go forth collecting pictures of it. With each correct picture, you grow stronger to defeat the dark wizard inside the tower!

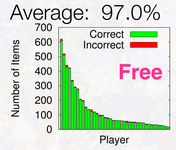
Annotation Mechanism



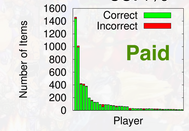
Summary: A tower's concept is shown when the player begins and then players recover images which maybe be related to the concept. Players keep related images and deposit them for rewards. Retained images are counted as valid image-concept relations.

Q1: How accurate are players at rejecting incorrect relations?

Average: 97.0%



95.4%



Players were even more accurate with TKT, detecting 95-97% of the incorrect image-based relations while playing the game. TKT benefits from a slower-paced game to improve player accuracy.

Q2: What is the quality of players compared to CrowdFlower?

	True Pos.	True Neg.	All
TKT free	82.5	82.5	82.5
TKT paid	69.0	92.1	74.0
CrowdFlower	59.5	93.7	66.2

Similar to Infection, TKT players were more accurate than workers, especially at identifying valid relations. Workers performed well on true negative only because they rejected most images!

Q3: How do TKT and CrowdFlower compare in cost?

	# Players	# Annotations	Cost per Annotation
TKT free	100	3005	$0.000
TKT paid	97	3318	$0.023
CrowdFlower	290	13854	$0.008

Just as with Infection, TKT players played equally in the free and paid conditions, showing that payment is not necessary.

Examples of Validated and New Concept-Image Relations

Term	Summarized Definition	Top-rated images
atom	The smallest possible particle	
color	An attribute of light	
religion	An expression of man's beliefs	

Images with a dashed border are new; CrowdFlower workers only marked the images with a * as valid.

References

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In Proceedings of ACL, pages 216–225.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In Proceedings of CHI, pages 319–326.

The authors gratefully acknowledge the support of ERC Starting Grant MultiJEDI No. 259234.