It's All Fun and Games until Someone Annotates:
Video Games with a Purpose for Linguistic Annotation.

David Jurgens
Stanford University
jurgens@stanford.edu

Roberto Navigli
Sapienza University of Rome
navigli@di.uniroma1.it
Annotating is no fun.
The sprinter won the race

1) be the winner in a contest or competition
2) win something through one's efforts
3) obtain advantages, such as points
4) attain success or reach a desired goal
Students
Students
\[ \sim 40K \text{ Turkers Active Concurrently} \times 1 \text{ Week} = 6.7M \text{ hours of possible MTurk time per week} \]
~40K Turkers Active Concurrently × 1 Week = 6.7M hours of possible MTurk time per week

3,000,000,000 hours spent playing video games per week
~40K Turkers Active Concurrently × 1 Week = 6.7M hours of possible MTurk time per week

3,000,000,000,000 hours spent playing video games per week

I should make a video game
The current state of NLP games

Wordrobe (Venhuizen et al., 2013)

Jinx (Seemakurty et al., 2010)

Phrase Detectives (Poesio et al., 2013)
The current state of NLP games

Wordrobe (Venhuizen et al., 2013)

Jinx (Seemakurty et al., 2010)

More similar to gamified tasks than taskified games

Phrase Detectives (Poesio et al., 2013)
Can we take a video game and taskify it?
Contributions

Taskified Games
Design Methodologies
Expert-level Performance
Taskify a game with a popular design

Temple Run

1 Billion Downloads

Fruit Ninja

300 Million Downloads
1/3 of all iPhones
Can we adapt Fruit Ninja for disambiguation?
Key mechanic:
Click on certain kinds of things

Player needs to avoid these
She plays the **bass**

1) the lowest part of the musical range

2) an adult male singer with the lowest voice

3) a North American freshwater fish

4) a musical instrument
She plays the **bass**

1) the lowest part of the musical range

2) an adult male singer with the lowest voice

3) a North American freshwater fish

4) a musical instrument
She plays the **bass**

1) the lowest part of the musical range

2) an adult male singer with the lowest voice

3) a North American freshwater fish

4) a musical instrument
She plays the bass

1) the lowest part of the musical range

2) an adult male singer with the lowest voice

3) a North American freshwater fish

4) a musical instrument
She plays the bass
She plays the bass
She plays the bass
She plays the **bass**

**Key Problem #1:** This is boring
She plays the bass

Key Problem #1: This is boring
Key Problem #2: Game mistakes radically change results
She plays the **bass**

1) the lowest part of the musical range

2) an adult male singer with the lowest voice

3) a North American freshwater fish

4) a musical instrument
She plays the **bass**

1) the lowest part of the musical range

2) an adult male singer with the lowest voice

3) a North American freshwater fish

4) a musical instrument
She plays the **bass**
She plays the bass
She plays the bass

The most common gameplay mistake has no effect on the annotation
Where do we get the images?

1) the lowest part of the musical range

2) an adult male singer with the lowest voice

3) a North American freshwater fish

4) a musical instrument
Current image-sense libraries

Deng et al. (2009); Navigli and Ponzetto (2012)
Current image-sense libraries

No abstract nouns, no verbs

Few verbs, relatively few pictures

Deng et al. (2009); Navigli and Ponzetto (2012)
Build a game in order to create resources for another game!
Puzzle Racer

- Start
- Player Setup
- Instructions
- Game Options
- View Leaderboard
Real game features!

Unlockable Racers

Lots of Power-ups

Enemies!

Leaderboards
love

1: a strong positive emotion of affection

cat

1: a feline mammal
Task Question:
Which of these pictures best shows the following definition: “a building in which the business of banking transacted”
Task Question:
Which of these pictures best shows the following definition: “a building in which the business of banking transacted”
Taskify by making bad pictures in-game obstacles

Players must identify obstacles and dodge them
How do we get rid of the text to make it a game?

Which of these pictures best shows “a building in which the business of banking transacted”
This race's puzzle clues:

Find the idea in common and guide your racer over similar pictures to stay alive!

► Let's race!
This race's puzzle clues:

Find the idea in common and guide your racer over similar pictures to stay alive?

Let's race!
Players are shown two types of puzzle gates:

- **Golden Gate**
- **Mystery Gate**
First three gates are always golden

Accuracy: 66%

Accuracy: 100%

Accuracy: 100%
Show a **mystery** gate with probability $= 0.66 \times \text{Accuracy}$

First three gates are always **golden**

Accuracy: 66%

Accuracy: 100%

Accuracy: 100%
Does it work?
Game Setup

- Picked 23 nouns, verbs, and adjectives
  - 4-10 senses each; 132 senses total
- Start with ~10 gold images per sense and 16.6K unlabeled images total
- Recruited students to play, with offer of gift cards for top positions in leaderboard after two weeks ($70 total)
Gameplay Results

- 126 people played at least one game
- 7,199 races over two weeks
- 20,254 ratings across all images
  - 231 – 329 ratings per sense
- 83% accuracy at Golden Gates
How does PuzzleRacer compare in quality with Crowdsourcing?
Recreate the Puzzle Racer annotation task on CrowdFlower

Given the three example images in the instructions, which of the following images most resembles underlying idea?
Recreate the Puzzle Racer annotation task on CrowdFlower

Given the three example images in the instructions, which of the following images most resembles underlying idea?

One of these questions is from a **Golden** Gate, the others are from **Mystery** Gates
Evaluate by comparing top-ranked images

cat (n): a feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats

which is better?  left  equal  right
About equal in quality...  

7% 79% 14%  

... but Puzzle Racer was 27% the price!*
Does it work?
No statistically-significant difference in quality between Puzzle Racer-created and Expert-selected images
**argument** (n): a contentious speech act; a dispute where there is strong disagreement

**atmosphere** (n): the weather or climate at some place

**important** (a): of great significance or value

**climb** (v): go upward with gradual or continuous progress

**smell** (v): smell bad
Now back to disambiguation!
Disambiguate by clicking on pictures for the wrong senses
She plays the bass

Show one picture for each of the $n$ senses

- Musician
- Fish
- Guitar
She plays the **bass**

Show one picture for each of the $n$ senses

Include $n$ pictures from random senses
She plays the bass

Show one picture for each of the \( n \) senses

Include \( n \) pictures from random senses

Monitor player’s ability by them destroying unrelated images from random senses
She plays the bass

Show one picture for each of the \( n \) senses

Include \( n \) pictures from random senses

Each game produces a probability distribution over senses

Monitor player’s ability by them destroying unrelated images from random senses
Disambiguate by clicking on pictures for the wrong senses

Objective

The following sentence contains a clue to your survival. Look at **wins** and think of pictures that remind you of its meaning.

It is the one exercise that drastically influences the definition of the thighs at the hipline - that mark of the champion that sets him apart from all other bodybuilders - a criterion of muscle "drama" that is unforgettable to judges and audiences alike - the facet of muscular development that **wins** prizes.

When you click below, pictures will be thrown on screen. Your job is destroy every picture that **does not** remind you of **wins** in the sentence above. When in doubt, blow it up!

Let me blow stuff up.
Does it work?
Direct comparison with **Wordrobe**, a WSD game

Tested on 111 sentences total for 74 nouns and 16 verbs (3.4 senses on average)
Disambiguation Accuracy

Ka-Boom! (this work) 0.77
Most Frequent Sense (MFS) 0.67
Wordrobe (Venhuizen et al., 2013) 0.60
Random 0.32
Disambiguation Accuracy

- **All Words**
  - Ka-Boom! (this work): 0.77
  - Most Frequent Sense (MFS): 0.67
  - Wordrobe (Venhuizen et al., 2013): 0.60
  - Random: 0.32

- **Nouns**
  - Ka-Boom! (this work): 0.80
  - Most Frequent Sense (MFS): 
  - Wordrobe (Venhuizen et al., 2013): 0.60
  - Random: 0.32

- **Verbs**
  - Ka-Boom! (this work): 0.59
  - Most Frequent Sense (MFS): 
  - Wordrobe (Venhuizen et al., 2013): 
  - Random: 0.32
How long did players take to converge on the right sense?
How long did players take to converge on the right sense?

Three flights takes under a minute, which is equivalent to the annotation speed of experts (Krishnamurthy and Nicholls, 2000).
What went right?

Game elements proved fun and addicting (for us too)

Identified reusable patterns for taskifying games
What could have gone better?

- Game development is hard if you have no experience
  - 2 Months for Puzzle Racer vs. 1 week for Ka-boom!
- Still needed manual annotation to bootstrap the games
- Game were slower than crowdsourcing
  - But only because we didn’t have a ready pool of players
Don’t gamify your tasks, taskify your games!

David Jurgens
jurgens@stanford.edu

Roberto Navigli
navigli@di.uniroma1.it

http://knowledgeforge.org/