Crowdsourcing for Language Resources and Evaluation

Visionary Lecture for hub.urfu.ru

DOI: 10.5281/zenodo.3960805

About Me

Dr. **Dmitry Ustalov**

Data Analysis and Research Group
Yandex, Saint Petersburg, Russia

Research Interests: Crowdsourcing, Computational Lexical Semantics

Outline

1. Introduction
2. Wisdom of the Crowds
3. Microtasks
4. Games with a Purpose
5. Miscellaneous
6. Conclusion

Section 1

Introduction

Introduction

- Natural Language Processing (NLP) heavily relies on annotated datasets
- These datasets are also known as **Language Resources** (LRs)
- **Crowdsourcing** is an efficient approach for large-scale *knowledge acquisition* and *data annotation*
- However, it requires setup effort and careful quality control
- Today, we will learn how to do it!

Core Idea: **Trust, but Verify**

We can efficiently build and evaluate datasets for NLP using Crowdsourcing.

Language Resources and NLP

Do we really need Language Resources for NLP?

+ Any supervised task requires class labels
+ Every machine learning method requires evaluation
+ Every (new) dataset requires quality assessment
− More and more methods are trying not to rely on annotated data
− Available datasets can be reused

How can we produce Language Resources for NLP?

- Semi-supervised (or unsupervised) learning
- Expert annotators (assessors)
- Crowdsourcing

Crowdsourcing for NLP

Language Resource Construction

- Knowledge Base and Thesaurus (Wikipedia and Wiktionary)
- Word Sense Inventory (Biemann, 2013)
- Question Answering (Rajpurkar et al., 2018)
- Speech Corpus Acquisition (Ardila et al., 2020)

Language Resource Annotation

- Word Sense Disambiguation (Snow et al., 2008)
- Named Entity Recognition (Finin et al., 2010)
- Part-of-Speech Tagging (Bocharov et al., 2013)

Language Resource Evaluation

- Search Relevance (Alonso et al., 2008)
- Machine Translation (Callison-Burch, 2009)
- Topic Modeling (Chang et al., 2009)

What is Crowdsourcing?

Definition by Estellés-Arolas et al. (2012)

Crowdsourcing is a type of participative *online activity* in which *the requester* proposes to *a group of individuals* … the voluntary undertaking of *a task*.

Main components (Hosseini et al., 2014):

- Crowd
- Task
- Requester
- Platform

Source: Merrill (2014)

Crowdsourcing Genres

Wisdom of the Crowds (WotC)

- Wikipedia and Wiktionary
- Genius (former Rap Genius)
- Open Source Software

Microtasks (μT)

- Mechanical Turk
- reCAPTCHA
- Common Voice

Games with a Purpose (GWAPs)

- ESP Game
- Phrase Detectives

Source: Simone_ph (2017)

Section 2

Wisdom of the Crowds

Wisdom of the Crowds (WotC)

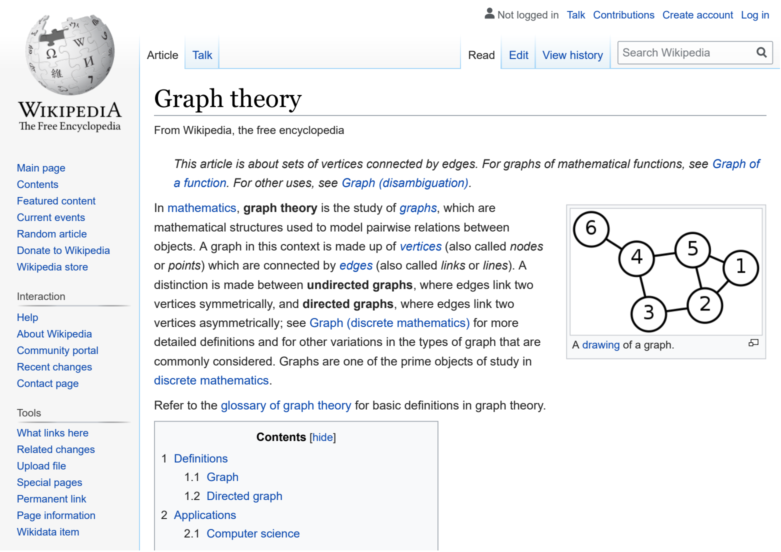
WotC deployments allow members of the general public to collaborate to build a public resource, or to predict event outcomes, or to estimate difficult to guess quantities (Wang et al., 2013a).

Crowd	Volunteers (usually driven by altruism)
Task	Content generation, etc.
Requester	A non-commercial organization (but not necessarily)
Platform	Collaborative editing tool

Examples: Wikipedia, Wiktionary, OpenStreetMap, Fandom (ex-Wikia), Open Source Software, etc.

Knowledge gathered by the **WotC** is later used in various applications: DBpedia (Auer et al., 2007), BabelNet (Navigli et al., 2012), Sense and Frame Induction (Ustalov et al., 2019), etc.

Example: Wikipedia



Graph theory

From Wikipedia, the free encyclopedia

This article is about sets of vertices connected by edges. For graphs of mathematical functions, see Graph of a function. For other uses, see Graph (disambiguation).

In mathematics, **graph theory** is the study of *graphs*, which are mathematical structures used to model pairwise relations between objects. A graph in this context is made up of *vertices* (also called *nodes* or *points*) which are connected by *edges* (also called *links* or *lines*). A distinction is made between **undirected graphs**, where edges link two vertices symmetrically, and **directed graphs**, where edges link two vertices asymmetrically; see Graph (discrete mathematics) for more detailed definitions and for other variations in the types of graph that are commonly considered. Graphs are one of the prime objects of study in discrete mathematics.

A drawing of a graph.

Refer to the glossary of graph theory for basic definitions in graph theory.

Contents [hide]

1 Definitions
  1.1 Graph
  1.2 Directed graph
2 Applications
  2.1 Computer science

Source: `https://en.wikipedia.org/wiki/Graph_theory`

Example: Wiktionary

kitten

See also: **Kitten**

Contents [show]

English [edit]

Etymology [edit]

From Middle English *kitoun*, *kiton*, *kyton* (“kitten”), diminutive of *cat* (“cat”), equivalent to *cat* + *-en*. The first element is probably from Middle English *kiteling* (“kitten, kit”), Old Norse *ketlingr* (“kitten”), or possibly Anglo-Norman **kiton* or Old French *chiton*, diminutive of *cat*, *chat* (“cat”), from Late Latin *cattus*. Compare Low German *kitten* (“kitten”). More at kitling, cat, and -en.

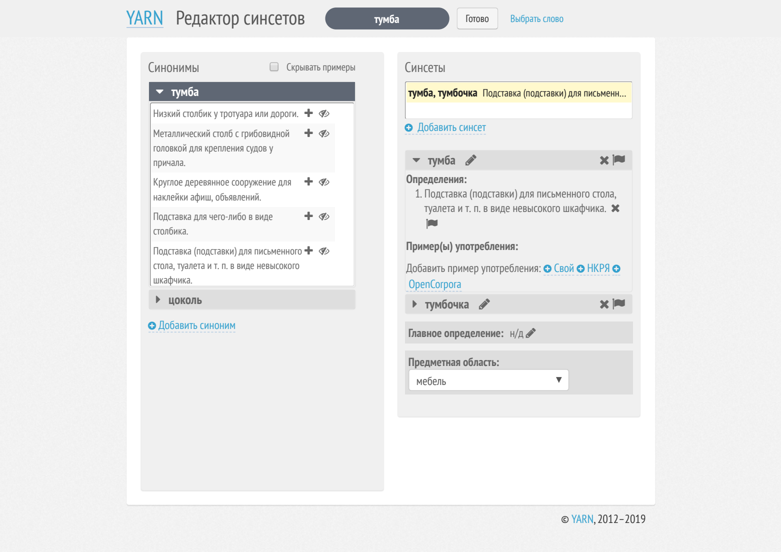
A kitten.

Pronunciation [edit]

- (*Received Pronunciation*) IPA$^{\text{(key)}}$: /ˈkɪtən/
- Audio (RP)
- (*General American*) IPA$^{\text{(key)}}$: [ˈkɪ.ʔn̩]
- Audio (GA)
- Rhymes: -ɪtən
- Hyphenation: kit·ten

Source: `https://en.wiktionary.org/wiki/kitten` and Alves Gaspar (2005)

Example: Yet Another RussNet (YARN)



YARN Редактор синсетов

тумба

Синонимы

☐ Скрывать примеры

тумба

- Низкий столбик у тротуара или дороги.
- Металлический столб с грибовидной головкой для крепления судов у причала.
- Круглое деревянное сооружение для наклейки афиш, объявлений.
- Подставка для чего-либо в виде столбика.
- Подставка (подставки) для письменного стола, туалета и т. п. в виде невысокого шкафчика.

цоколь

Синсеты

тумба, тумбочка Подставка (подставки) для письменн...

тумба

Определения:

1. Подставка (подставки) для письменного стола, туалета и т. п. в виде невысокого шкафчика.

Пример(ы) употребления:

Добавить пример употребления: Свой НКРЯ OpenCorpora

тумбочка

Главное определение: н/д

Предметная область:

мебель ▼

© YARN, 2012–2019

Source: Braslavski et al. (2014)

Quality Control in WotC

Quality Issues:

- Unwanted contributions
- Edit wars
- Misinformation

Excerpt from Halfaker et al. (2013)

Rejection of unwanted contributions is Wikipedia's primary quality control mechanism (Stvilia et al., 2008).

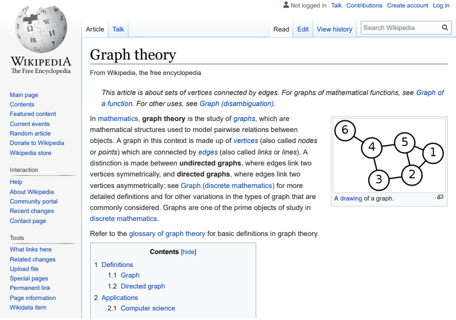
Means for Quality Control:

- Focusing community effort through content-based quality control
- Malicious contributions reversion using edit patrolling

Content-based Quality Control

- Is the Wikipedia page in the screenshot good?
- **Yes**, because it has text, images, links, sections, etc.
- They are good indicators of a high-quality, *featured*, article
- We can automate quality assessment using supervised learning!
- What if we consider every article with 2K+ words as good?

! Accuracy in the binary classification task will be 96.31% (Blumenstock, 2008)



Graph theory

From Wikipedia, the free encyclopedia

This article is about sets of vertices connected by edges. For graphs of mathematical functions, see Graph of a function. For other uses, see Graph (disambiguation).

In mathematics, **graph theory** is the study of *graphs*, which are mathematical structures used to model pairwise relations between objects. A graph in this context is made up of *vertices* (also called *nodes* or *points*) which are connected by *edges* (also called *links* or *lines*). A distinction is made between **undirected graphs**, where edges link two vertices symmetrically, and **directed graphs**, where edges link two vertices asymmetrically; see Graph (discrete mathematics) for more detailed definitions and for other variations in the types of graph that are commonly considered. Graphs are one of the prime objects of study in discrete mathematics.

A drawing of a graph.

Refer to the glossary of graph theory for basic definitions in graph theory.

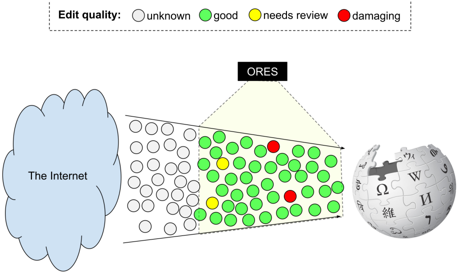
Contents [hide]

1 Definitions
  1.1 Graph
  1.2 Directed graph
2 Applications
  2.1 Computer science

Source: Wikipedia (2019)

Content-based Quality Control: Features

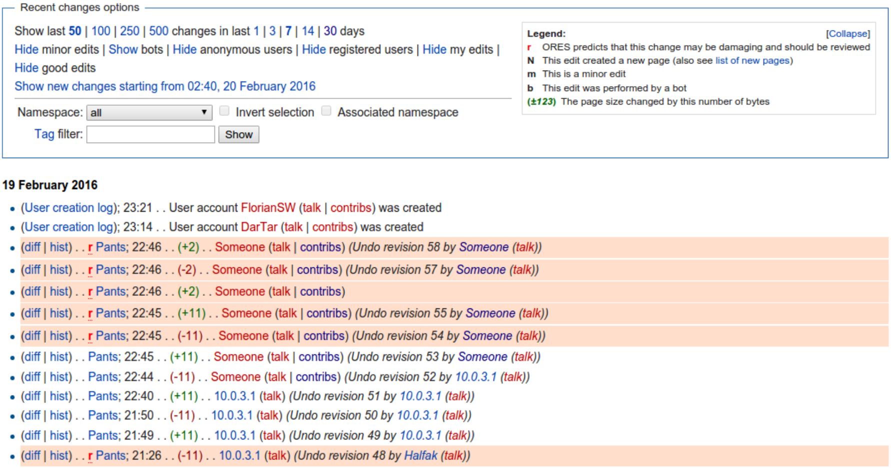
- Informative features are *page edits and discussions* (Wilkinson et al., 2007), *article length* (Blumenstock, 2008), *concentration of editors* (Kittur et al., 2008), *readability scores* (Wang et al., 2019), etc.
- Same principles apply to other collaborative projects, such as Wiktionary (Ustalov, 2014)



Source: Halfaker (2015)

ORES (Objective Revision Evaluation System) by Halfaker et al. (2019) uses gradient boosting to predict *damaging* page revisions.

Content-based Quality Control: Example



Recent changes options

Show last 50 | 100 | 250 | 500 changes in last 1 | 3 | 7 | 14 | 30 days
Hide minor edits | Show bots | Hide anonymous users | Hide registered users | Hide my edits |
Hide good edits
Show new changes starting from 02:40, 20 February 2016

Namespace: all ▾ ☐ Invert selection ☐ Associated namespace
Tag filter: Show

Legend: [Collapse]

r	ORES predicts that this change may be damaging and should be reviewed
N	This edit created a new page (also see list of new pages)
m	This is a minor edit
b	This edit was performed by a bot
(±123)	The page size changed by this number of bytes

19 February 2016

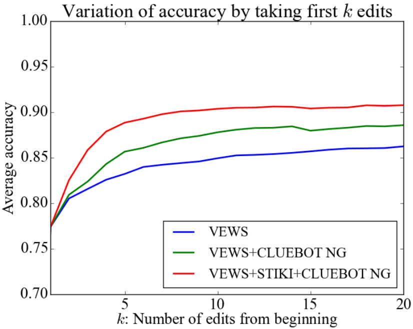
- (User creation log); 23:21 . . User account FlorianSW (talk | contribs) was created
- (User creation log); 23:14 . . User account DarTar (talk | contribs) was created
- (diff | hist) . . r Pants; 22:46 . . (+2) . . Someone (talk | contribs) *(Undo revision 58 by Someone (talk))*
- (diff | hist) . . r Pants; 22:46 . . (-2) . . Someone (talk | contribs) *(Undo revision 57 by Someone (talk))*
- (diff | hist) . . r Pants; 22:46 . . (+2) . . Someone (talk | contribs)
- (diff | hist) . . r Pants; 22:45 . . (+11) . . Someone (talk | contribs) *(Undo revision 55 by Someone (talk))*
- (diff | hist) . . r Pants; 22:45 . . (-11) . . Someone (talk | contribs) *(Undo revision 54 by Someone (talk))*
- (diff | hist) . . Pants; 22:45 . . (+11) . . Someone (talk | contribs) *(Undo revision 53 by Someone (talk))*
- (diff | hist) . . Pants; 22:44 . . (-11) . . Someone (talk | contribs) *(Undo revision 52 by 10.0.3.1 (talk))*
- (diff | hist) . . Pants; 22:40 . . (+11) . . 10.0.3.1 (talk) *(Undo revision 51 by 10.0.3.1 (talk))*
- (diff | hist) . . Pants; 21:50 . . (-11) . . 10.0.3.1 (talk) *(Undo revision 50 by 10.0.3.1 (talk))*
- (diff | hist) . . Pants; 21:49 . . (+11) . . 10.0.3.1 (talk) *(Undo revision 49 by 10.0.3.1 (talk))*
- (diff | hist) . . r Pants; 21:26 . . (-11) . . 10.0.3.1 (talk) *(Undo revision 48 by Halfak (talk))*

Source: Sarabadani (2016)

`https://www.mediawiki.org/wiki/ORES`

Edit Patrolling

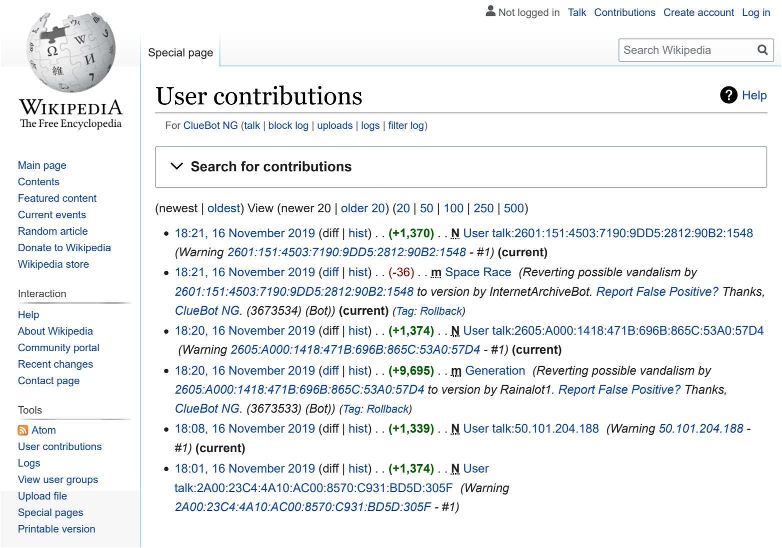
- **ClueBot NG** is an anti-vandalism bot that automatically reverts *vandal* edits based on content (Geiger et al., 2013)
- Combination of a Bayesian and a neural model allows reverting thousands of damaging edits in a few seconds
- Vandals tend to be more involved in edit wars, while benign users are more likely to participate in discussions (Kumar et al., 2015)



Source: Kumar (2015)

https://en.wikipedia.org/wiki/User:ClueBot_NG

Edit Patrolling: Example



User contributions

For ClueBot NG (talk | block log | uploads | logs | filter log)

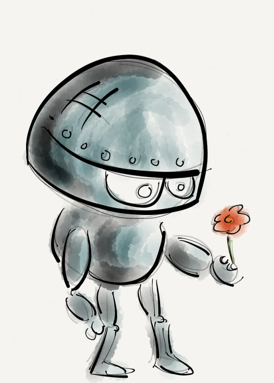
(newest | oldest) View (newer 20 | older 20) (20 | 50 | 100 | 250 | 500)

- 18:21, 16 November 2019 (diff | hist) . . **(+1,370)** . . **N** User talk:2601:151:4503:7190:9DD5:2812:90B2:1548 *(Warning 2601:151:4503:7190:9DD5:2812:90B2:1548 - #1)* **(current)**
- 18:21, 16 November 2019 (diff | hist) . . (-36) . . **m** Space Race *(Reverting possible vandalism by 2601:151:4503:7190:9DD5:2812:90B2:1548 to version by InternetArchiveBot. Report False Positive? Thanks, ClueBot NG. (3673534) (Bot))* **(current)** *(Tag: Rollback)*
- 18:20, 16 November 2019 (diff | hist) . . **(+1,374)** . . **N** User talk:2605:A000:1418:471B:696B:865C:53A0:57D4 *(Warning 2605:A000:1418:471B:696B:865C:53A0:57D4 - #1)* **(current)**
- 18:20, 16 November 2019 (diff | hist) . . **(+9,695)** . . **m** Generation *(Reverting possible vandalism by 2605:A000:1418:471B:696B:865C:53A0:57D4 to version by Rainalot1. Report False Positive? Thanks, ClueBot NG. (3673533) (Bot))* *(Tag: Rollback)*
- 18:08, 16 November 2019 (diff | hist) . . **(+1,339)** . . **N** User talk:50.101.204.188 *(Warning 50.101.204.188 - #1)* **(current)**
- 18:01, 16 November 2019 (diff | hist) . . **(+1,374)** . . **N** User talk:2A00:23C4:4A10:AC00:8570:C931:BD5D:305F *(Warning 2A00:23C4:4A10:AC00:8570:C931:BD5D:305F - #1)*

Source: Wikipedia (2019)

WotC: Wrap-Up

- WotC aims at large-scale content creation, but the data are usually *quasi-structured*
- Automated quality assurance and curation save editors' time
- Tune the false positive rate so it does not banish the (new) users
- Veracity is a serious issue, novel techniques are required (Esteves et al., 2018)



Source: bamenny (2016)

Section 3

Microtasks

Microtasks (μT)

In **μT**, requesters create and list batches of small jobs termed **Human Intelligence Tasks** (HITs), which may be done by the general public (Wang et al., 2013a).

Crowd	Paid contributors (usually)
Task	Data annotation, verification, evaluation
Requester	Paying customer (usually)
Platform	Website with HITs

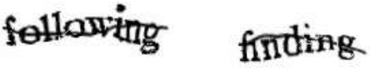
- One HIT is usually performed by *multiple* annotators
- Annotation results are then *aggregated*

Examples: reCAPTCHA, Common Voice, etc.

Example: reCAPTCHA

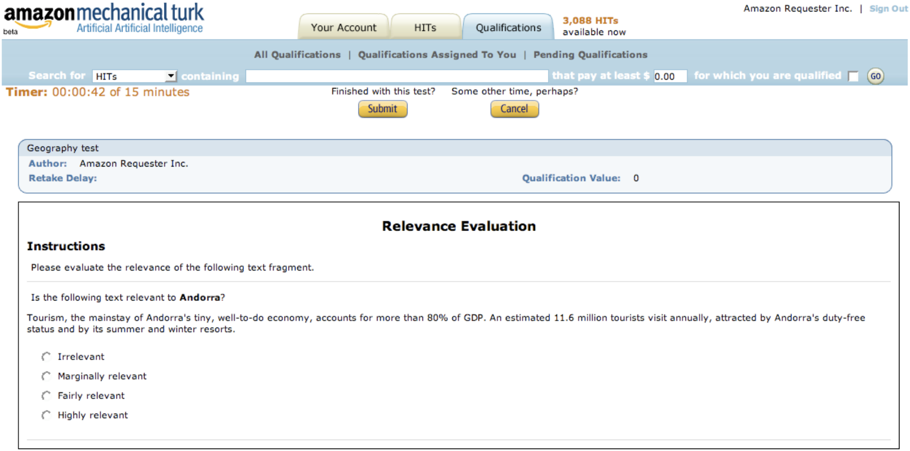
reCAPTCHA is a crowd-powered optical character recognition (OCR) system (von Ahn et al., 2008).

- Gives the user two words: the one for which *the answer is not known* and a second "control" word for which *the answer is known*
- Achieves an accuracy of 99.1% vs. 81.3% of a standard OCR as of 2008
- If the first two human guesses match each other and one of the OCR's predictions, they are considered a correct answer
- Yet novel aggregation methods exist (Shishkin et al., 2020)



Source: BMaurer (2007)

Example: Search Relevance Evaluation



Timer: 00:00:42 of 15 minutes

Finished with this test? Some other time, perhaps?

Geography test

Author: Amazon Requester Inc.

Retake Delay:

Qualification Value: 0

Relevance Evaluation

Instructions

Please evaluate the relevance of the following text fragment.

Is the following text relevant to **Andorra**?

Tourism, the mainstay of Andorra's tiny, well-to-do economy, accounts for more than 80% of GDP. An estimated 11.6 million tourists visit annually, attracted by Andorra's duty-free status and by its summer and winter resorts.

○ Irrelevant
○ Marginally relevant
○ Fairly relevant
○ Highly relevant

Source: Alonso et al. (2008)

Microtask Platforms: Managed

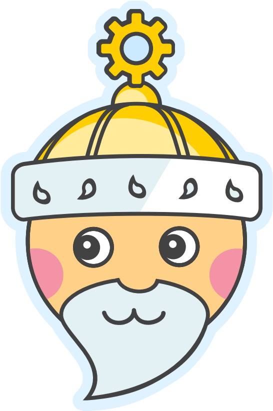
- **Amazon Mechanical Turk** (aka MTurk or AMT), `https://www.mturk.com/`
- **Appen**, `https://appen.com/`
- **Scale AI**, `https://scale.com/`
- **Yandex.Toloka**, `https://toloka.yandex.com/`

There are dozens of them!

Microtask Platforms: On-Premise

- Berkeley Open System for Skill Aggregation (**BOSSA**), https://boinc.berkeley.edu/trac/wiki/BossaIntro
- **PYBOSSA**, https://pybossa.com/
- **Mechanical Tsar** (Ustalov, 2015a), https://mtsar.nlpub.org/
- **prodigy**, https://prodi.gy/

And much more!



Source: Ustalov (2015a)

Microtask Platforms: Discussion

Managed Platforms

Pros:

- + Crowd is already “offered”
- + Reliable and battle-tested
- + Maintained by a third party

Cons:

- − Paid
- − Task adjustment is required
- − Competition between requesters
- − Maintained by a third party

On-Premise Platforms

Pros:

- + Self-hosted
- + Customizable

Cons:

- − No crowd
- − Self-hosted
- − Lack of support

Quality Control in Microtasks

Quality Issues:

- Task design
- Spam
- Reliability

Source: Finnerty et al. (2013)

Means for Quality Control:

- Task representation and decomposition
- Annotator pre-selection
- Inter-annotator agreement
- Answer aggregation

Task Representation and Decomposition

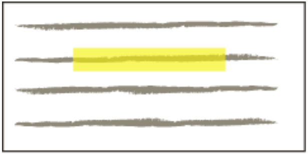
- Split a complex HIT into a sequence of several *simpler* HITs
- Use *post-acceptance* when annotators review each others' answers
- Apply computational approaches for self-evaluation

Source: rawpixel (2017)

Case Study: Soylent I

Soylent is a plugin for a popular word processor that crowdsources copy-editing of a text using the **Find-Fix-Verify** pipeline (Bernstein et al., 2010).

Find



Fix

`Soylent,`
`a prototype...`

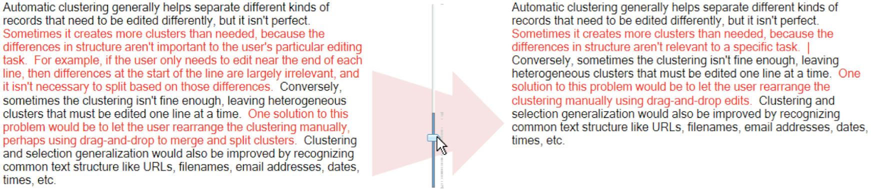
Verify

☐ `Soylent`~~`is`~~`, a prototype...`
☐ `Soylent` ~~`is a`~~ `prototypes...`
☑ `Soylent is a` ~~`prototype`~~`test...`

Source: Bernstein (2020)

Find-Fix-Verify is useful as a generic design pattern for *open-ended tasks* (not just text-centric ones).

Case Study: Soylent II



Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split based on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't relevant to a specific task. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually using drag-and-drop edits. Clustering and selection generalization would also be improved by recognizing common text structure like URLs, filenames, email addresses, dates, times, etc.

Source: Bernstein (2020)

Variations of Soylent:

- **Shortn** for text simplification
- **Crowdproof** for proof-reading
- **Human Macro** for arbitrary tasks

Case Study: Soylent III

Shortn

- Five examples of texts, each between one and seven paragraphs long
- Median processing time is 18.5 minutes
- Revisions are 78–90% of the original document length
- 104 out of 137 suggestions are accepted at the Verify step

Crowdproof

- Five input texts in need of proofreading, 49 errors in total
- Median processing time is 18 minutes
- 33 out of 49 errors are caught by the crowd
- 29 out of 33 caught errors are accepted on the Verify step

Human Macro

- Five different tasks, e.g., change tense (past → present), etc.
- 88% intention success rate
- 30% of work contained an error

Annotator Pre-Selection

- ! **Gold-based quality assurance:** pre-annotate a portion of the dataset to estimate the annotator accuracy (Oleson et al., 2011)
- **Profile the annotators:** recommend them a HIT in advance as according to their interests (Difallah et al., 2013)
- **Evaluate the behaviour traces:** pre-classify the annotators by their online activity on the platform (Gadiraju et al., 2019)

Source: McGuire (2015)

Online Behaviour Analysis

Gadiraju et al. (2019) recorded behavioural traces of annotators in two kinds of HITs:

- *content creation* (solving CAPTCHAs)
- *information finding* (question answering)

Answer accuracy and annotator motivation are checked against the recorded traces: mouse, keyboard, scrolling events, etc.

A fine-grained *typology* of crowd annotators:

- **Good Annotators** (DW, CW)
- **Cheaters** (FD, SD, RB)
- **Inexperienced Annotators** (LW, SW)

These traces are used as features for annotator type prediction using a random forest classifier.

Online Behaviour Analysis: Results

Good Annotators are accurate and usually (but not necessarily) fast; cheaters are the fastest, but they have the lowest accuracy.

Most informative features are:

- mouse activity
- window management activity
- task completion time
- gold questions

Pre-selection allows increasing accuracy of the annotation results and helps providing additional training for inexperienced annotators.

Inter-Annotator Agreement

- How *reliable* is the annotation?
- In 51.1% cases the annotators agree with each other, is it good?
- A low value indicates issues with task design and difficulty: the crowd answers might make no sense

	w_1	w_2	w_3	w_4
t_1	cat		cat	cat
t_2	cat	dog	dog	cat
t_3	dog	dog	dog	cat
t_4	dog	cat	cat	dog

Krippendorff's α

Krippendorff's α (2018) is a versatile inter-annotator agreement measure that takes into account the *observed* disagreement D_o and the *expected* disagreement D_e:

$$\alpha = 1 - \frac{D_o}{D_e}$$

α is chance-corrected, handles missing values, and allows for arbitrary distance functions (binary, nominal, interval, etc.)

In the *nominal* case of C classes α is computed using a coincidence matrix $O \in \mathbb{R}^{|C|\times|C|}$:

$$_{\mathrm{nominal}}\alpha = 1 - (n-1)\frac{n - \sum_{c \in C} O_{cc}}{n^2 - \sum_{c \in C} n_c^2},$$

where $n_c = \sum_{k \in C} O_{ck}$ and $n = \sum_{c \in C} n_c$.

Krippendorff's α: Algorithm

```
Input: $m$ annotators, $N$ tasks, $C$ classes,                ▷ Missing values are $(-)$
       data matrix $U \in (\{-\} \cup C)^{m \times |N|}$
Output: $0 \leq {}_{\text{nominal}}\alpha \leq 1$
$O_{ck} \leftarrow 0$ for all $c \in C, k \in C$
for all $u \in N$ do                                          ▷ Each task
    for all $c, k \in P(U_u^\intercal, 2)$ do                 ▷ Each possible non-missing $(c, k)$ pair
        $O_{ck} \leftarrow O_{ck} + \frac{1}{m_u - 1}$        ▷ $m_u$ is the number of annotators in task $u$
$n_c \leftarrow \sum_{k \in C} O_{ck}$ for all $c \in C$
$n \leftarrow \sum_{c \in C} n_c$
return $1 - (n - 1)\frac{n - \sum_{c \in C} O_{cc}}{n^2 - \sum_{c \in C} n_c^2}$
```

Krippendorff's α: Example

$$O = \begin{pmatrix} 4.33 & 3.67 \\ 3.67 & 3.33 \end{pmatrix}$$

$$n_c = \begin{pmatrix} 8 & 7 \end{pmatrix}$$

$$n = 15$$

U^{T}

	w_1	w_2	w_3	w_4
t_1	cat		cat	cat
t_2	cat	dog	dog	cat
t_3	dog	dog	dog	cat
t_4	dog	cat	cat	dog

$$_{\text{nominal}}\alpha = 1 - (n-1)\frac{n - \sum_{c \in C} O_{cc}}{n^2 - \sum_{c \in C} n_c^2} = 1 - 14\frac{15 - (4.33 + 3.33)}{15^2 - (8^2 + 7^2)}$$

$$= 1 - \frac{102.76}{112} \approx 0.083$$

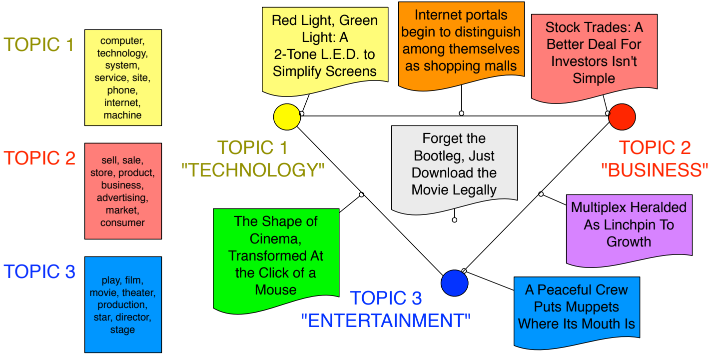
Inter-Annotator Agreement: Discussion

- α provides a convenient *single* number indicating the extent of how the annotators agree with each other
- **Interpretation** by Krippendorff (2018):
  - $\alpha > 0.800$: reliable annotation (reliability $\neq$ correctness!)
  - $0.667 \leq \alpha \leq 0.800$: tentative conclusions only
- **Implementations:** DKPro for Java (Meyer et al., 2014), NLTK for Python (Bird et al., 2017), irr for R, etc.
- A good discussion on this topic is available in Artstein et al. (2008)

Source: rawpixel (2018)

Case Study: Topic Models and Intruders I

Topic modeling algorithms are statistical methods that discover the *themes* that run through the *words* in *texts*.



Source: Boyd-Graber (2014)

- How to evaluate such an unsupervised model?
- Chang et al. (2009) proposed *intruders*

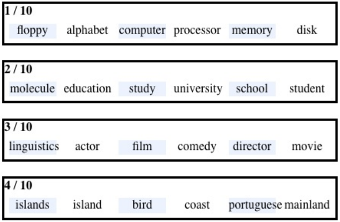
Case Study: Topic Models and Intruders II

How topics match *human concepts*?

- Present a set of words and ask to select the *intruder* word which does not belong to the others
- Compute the *model precision* as the fraction of annotators agreeing with the model

How topics match to *documents*?

- Present title and excerpt, select the *intruder* topic that does not belong
- Compute the *topic log odds* as the agreement between the model and human judgements



1 / 10

floppy alphabet computer processor memory disk

2 / 10

molecule education study university school student

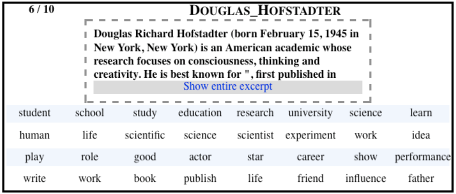
3 / 10

linguistics actor film comedy director movie

4 / 10

islands island bird coast portuguese mainland

Source: Boyd-Graber (2014)



6 / 10 DOUGLAS_HOFSTADTER

Douglas Richard Hofstadter (born February 15, 1945 in New York, New York) is an American academic whose research focuses on consciousness, thinking and creativity. He is best known for ", first published in

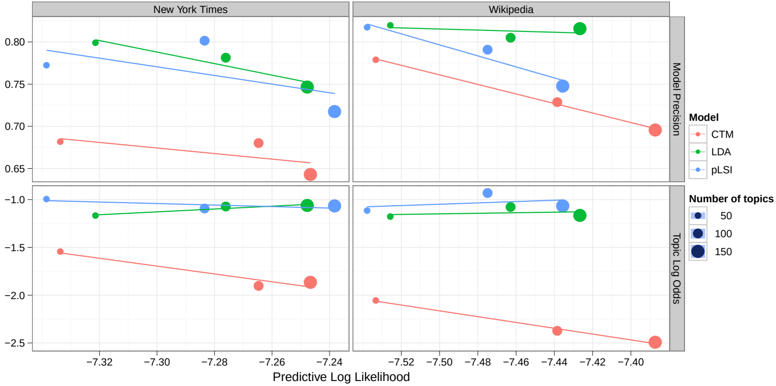
Show entire excerpt

student	school	study	education	research	university	science	learn
human	life	scientific	science	scientist	experiment	work	idea
play	role	good	actor	star	career	show	performance
write	work	book	publish	life	friend	influence	father

Source: Boyd-Graber (2014)

Case Study: Topic Models and Intruders III

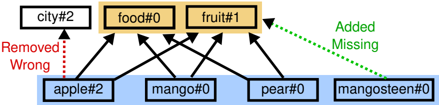
- Intruders reveal discrepancy in data using *controlled distortion*
- Human judgements do not have to correlate with the machine ones
- Can be used in different setups



Source: Boyd-Graber (2014)

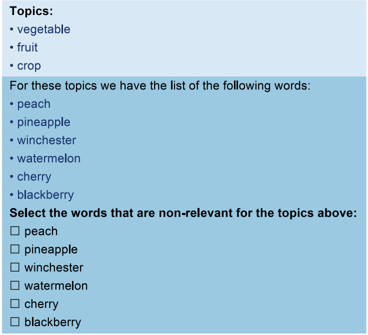
Case Study: Topic Models and Intruders IV

- A similar microtask-based evaluation scheme for distributional semantic classes by Panchenko et al. (2018a)



- **Accuracy** is the fraction of tasks where annotators correctly identified the intruder
- **Badness** is the fraction of tasks for which non-intruder words were selected

	Accuracy	Badness	Kripp. α
Clusters	86%	25%	0.588
Hypernyms	92%	21%	0.655



Topics:
- vegetable
- fruit
- crop

For these topics we have the list of the following words:
- peach
- pineapple
- winchester
- watermelon
- cherry
- blackberry

Select the words that are non-relevant for the topics above:

☐ peach
☐ pineapple
☐ winchester
☐ watermelon
☐ cherry
☐ blackberry

Source: Panchenko et al. (2018a)

Answer Aggregation

- Every HIT is usually performed by multiple annotators
- How to select the right answer?
- **Majority Vote** (MV) is an obvious, but robust strategy (Sheshadri et al., 2013)
- Ties must be broken *randomly*
- Can we have a better solution?

	w_1	w_2	w_3	w_4
t_1	cat		cat	cat
t_2	cat	dog	dog	cat
t_3	dog	dog	dog	cat
t_4	dog	cat	cat	dog

	w_1	w_2	w_3	w_4	A
t_1	cat		cat	cat	cat
t_2	cat	dog	dog	cat	cat
t_3	dog	dog	dog	cat	dog
t_4	dog	cat	cat	dog	dog

G	cat	dog	dog	cat

Dawid-Skene Model

- **Dawid-Skene** (1979) model infers confusion matrices for annotators and priors for labels
- A probabilistic graphical model for medical examinations available in an *analytical form*
- Can be implemented using an EM algorithm that *converges* to a local optimum

	w_1	w_2	w_3	w_4	A
t_1	cat		cat	cat	cat
t_2	cat	dog	dog	cat	dog
t_3	dog	dog	dog	cat	dog
t_4	dog	cat	cat	dog	cat

G	cat	dog	dog	cat

$$n_{ij}^{(k)} = \begin{cases} 1, & \text{if } w_k \text{ answered } j \text{ for } t_i, \\ 0, & \text{otherwise} \end{cases}$$

Dawid-Skene Model: Algorithm I

```
Input:  $K$ annotators, $I$ tasks, $J$ classes
        matrix $n^{(k)} \in \{0,1\}^{|I| \times |J|}$ for $k \in K$
Output: class probabilities $T \in [0,1]^{|I| \times |J|}$
$T_{ij} \leftarrow \frac{\sum_{k \in K} n_{ij}^{(k)}}{\sum_{k \in K} \sum_{l \in J} n_{il}^{(k)}}$ for all $i \in I, j \in J$    ▷ Initialize with MV
while $T$ changes do
    for all $j \in J$ do    ▷ Each true label
        for all $k \in K$ do    ▷ Each annotator
            for all $l \in J$ do    ▷ Each observed label
                $\hat{\pi}_{jl}^{(k)} \leftarrow \frac{\sum_{i \in I} T_{ij} n_{il}^{(k)}}{\sum_{l \in J} \sum_{i \in I} T_{ij} n_{il}^{(k)}}$    ▷ Confusion matrix for annotator $k$
                $\hat{p}_j \leftarrow \sum_{i \in I} \frac{T_{ij}}{\sum_{l \in J} T_{il}}$    ▷ Prior for true label $j$
    $\pi, p \leftarrow \hat{\pi}, \hat{p}$    ▷ Use estimates for $\pi$ and $p$
```

Continued on the next slide

Dawid-Skene Model: Algorithm II

```
 2: while $T$ changes do                          ▷ Continue the while loop from line 2
 9:     for all $j \in J$ do                                        ▷ Each true label
10:         for all $i \in I$ do                                         ▷ Each task
11:             $\hat{T}_{ij} \leftarrow \frac{p_j \prod_{k \in K} \prod_{l \in J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}}}{\sum_{q \in J} p_q \prod_{k \in K} \prod_{l \in J} (\pi_{ql}^{(k)})^{n_{il}^{(k)}}}$    ▷ Answer for task $i$
12:     $T \leftarrow \hat{T}$                                ▷ Use new estimate for $T$
13: return $T$                                       ▷ $\pi$ and $p$ are also useful
```

Aggregation: $A_i = \arg\max_{j \in J} T_{ij}$ **for all** $i \in I$

Dawid-Skene: Example

	w_1	w_2	w_3	w_4
t_1	cat		cat	cat
t_2	cat	dog	dog	cat
t_3	dog	dog	dog	cat
t_4	dog	cat	cat	dog

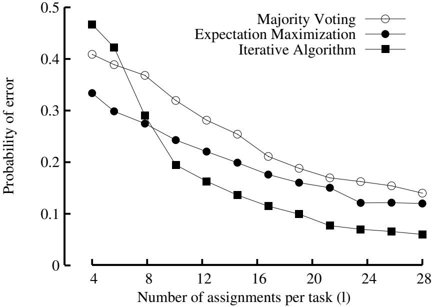
	cat	dog
t_1	0.82	0.18
t_2	0.23	0.77
t_3	0.12	0.88
t_4	0.96	0.04

G	cat	dog	dog	cat

- We aggregate the answers using the Dawid-Skene (1979) model

Answer Aggregation: Summary

- **Alternatives:** GLAD (Whitehill et al., 2009), KOS (Karger et al., 2014), etc.
- Can be used for repeated labeling (Sheng et al., 2008), dynamic pricing (Wang et al., 2013b), etc.
- **Implementations:** SQUARE (Sheshadri et al., 2013), `spark-crowd` (Rodrigo et al., 2019), etc.
- No algorithm can do better than Majority Vote if the response quality is low



Source: Karger et al. (2011)

Case Study: RUSSE’2018 `bts-rnc` Dataset I

RUSSE’2018 was the second instance of the Russian Semantic Evaluation shared task (Panchenko et al., 2018b).

- Word Sense Disambiguation: given a word in a context, predict its meaning
- No gold standard was available, so we created it *almost* from scratch
- `bts-rnc`: sense inventory of the *Large Explanatory Dictionary* of Russian
- A subset of this dataset was annotated using crowdsourcing

Source: Buissinne (2016)

`https://russe.nlpub.org/2018/wsi/`

Case Study: RUSSE'2018 `bts-rnc` Dataset II

Setup:

- 9 annotators per context
- Limited annotation speed and number of answers per annotator
- Dawid-Skene aggregation algorithm

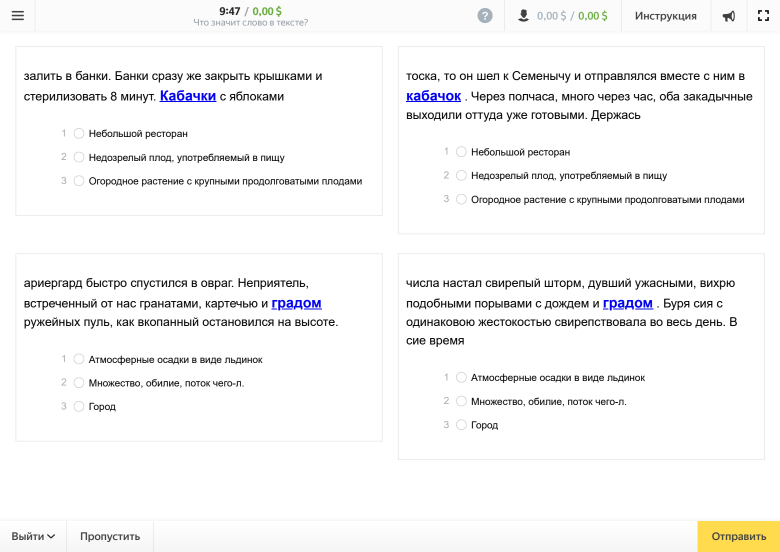
Pipeline:

1. Choose 20 words and sample their contexts
2. Prepare 80 manually-annotated contexts for training
3. Annotate 2 562 contexts on the platform
4. Tip for feedback and error reports
5. Manually curate the aggregated results

Results:

- Annotation took 35 minutes
- Krippendorff's $\alpha = 0.83$

Case Study: RUSSE’2018 `bts-rnc` Dataset III



залить в банки. Банки сразу же закрыть крышками и стерилизовать 8 минут. **Кабачки** с яблоками

1 ○ Небольшой ресторан
2 ○ Недозрелый плод, употребляемый в пищу
3 ○ Огородное растение с крупными продолговатыми плодами

тоска, то он шел к Семенычу и отправлялся вместе с ним в **кабачок** . Через полчаса, много через час, оба закадычные выходили оттуда уже готовыми. Держась

1 ○ Небольшой ресторан
2 ○ Недозрелый плод, употребляемый в пищу
3 ○ Огородное растение с крупными продолговатыми плодами

ариергард быстро спустился в овраг. Неприятель, встреченный от нас гранатами, картечью и **градом** ружейных пуль, как вкопанный остановился на высоте.

1 ○ Атмосферные осадки в виде льдинок
2 ○ Множество, обилие, поток чего-л.
3 ○ Город

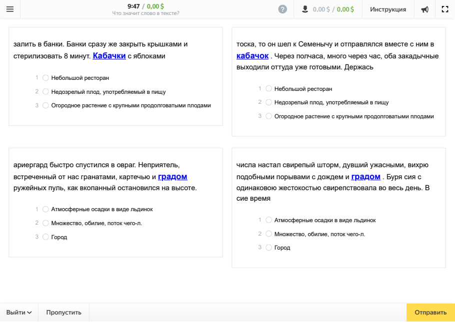
числа настал свирепый шторм, дувший ужасными, вихрю подобными порывами с дождем и **градом** . Буря сия с одинаковою жестокостью свирепствовала во весь день. В сие время

1 ○ Атмосферные осадки в виде льдинок
2 ○ Множество, обилие, поток чего-л.
3 ○ Город

Source: Panchenko et al. (2018b)

Case Study: RUSSE'2018 `bts-rnc` Dataset IV

- Curation was required due to the importance of the dataset
- Other datasets and the rest of `bts-rnc` were annotated by a team of linguists
- Quality comparable to the experts obtained in a fraction of time
- The dataset is available for download: `10.5281/zenodo.1117228` (CC BY-SA)



залить в банки. Банки сразу же закрыть крышками и стерилизовать 8 минут. **Кабачки** с яблоками

1 ○ Небольшой ресторан
2 ○ Недозрелый плод, употребляемый в пищу
3 ○ Огородное растение с крупными продолговатыми плодами

тоска, то он шел к Семенычу и отправлялся вместе с ним в **кабачок** . Через полчаса, много через час, оба закадычные выходили оттуда уже готовыми. Держась

1 ○ Небольшой ресторан
2 ○ Недозрелый плод, употребляемый в пищу
3 ○ Огородное растение с крупными продолговатыми плодами

ариергард быстро спустился в овраг. Неприятель, встреченный от нас гранатами, картечью и **градом** ружейных пуль, как вкопанный остановился на высоте.

1 ○ Атмосферные осадки в виде льдинок
2 ○ Множество, обилие, поток чего-л.
3 ○ Город

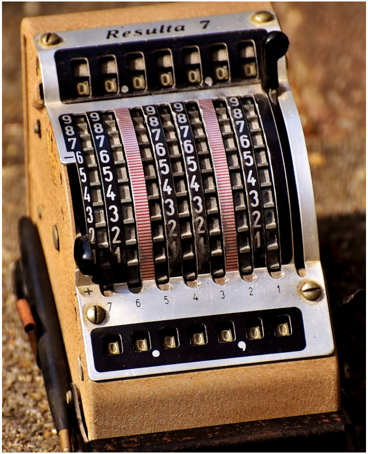
числа настал свирепый шторм, дувший ужасными, вихрю подобными порывами с дождем и **градом** . Буря сия с одинаковою жестокостью свирепствовала во весь день. В сие время

1 ○ Атмосферные осадки в виде льдинок
2 ○ Множество, обилие, поток чего-л.
3 ○ Город

Source: Panchenko et al. (2018b)

Microtasks: Wrap-Up

- A good task design should be accompanied with a proper quality control
- You are not alone: always listen to the feedback from the annotators and reward them
- There is **no excuse** for not computing (and reporting) the inter-annotator agreement
- Crowdsourcing platforms usually offer means for quality control
- Still, there is always room for improvement (Daniel et al., 2018)



Source: Alexas_Fotos (2017)

Section 4

Games with a Purpose

Games with a Purpose (GWAPs)

In **GWAPs**, annotation tasks are designed to provide entertainment to the human subject over the course of short sessions (Wang et al., 2013a).

Crowd	Volunteers
Task	Data annotation, verification, evaluation
Requester	Volunteers
Platform	Custom-made multiplayer video game

Examples: ESP Game (von Ahn et al., 2004), Phrase Detectives (Poesio et al., 2013), Ka-boom! (Jurgens et al., 2014), Infection (Vannella et al., 2014), etc.

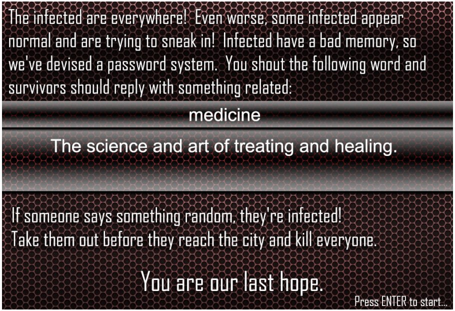
Example: ESP Game

- Image annotation is hard, so what about crowdsourcing it?
- The Extrasensory Perception (**ESP**) Game by von Ahn et al. (2004) was the historically first GWAP to go online
- Two players have to propose tags for the given image without using *taboo* words
- Mutual agreement indicates the success
- 85% of the words for each image would be useful in describing it

Source: von Ahn et al. (2004)

Case Study: Infection I

- **BabelNet** is a large-scale multilingual knowledge base (Navigli et al., 2012)
- How to improve the coverage of the language resource?
- Ask people for more words, but how to keep this useful?



The infected are everywhere! Even worse, some infected appear normal and are trying to sneak in! Infected have a bad memory, so we've devised a password system. You shout the following word and survivors should reply with something related:

medicine

The science and art of treating and healing.

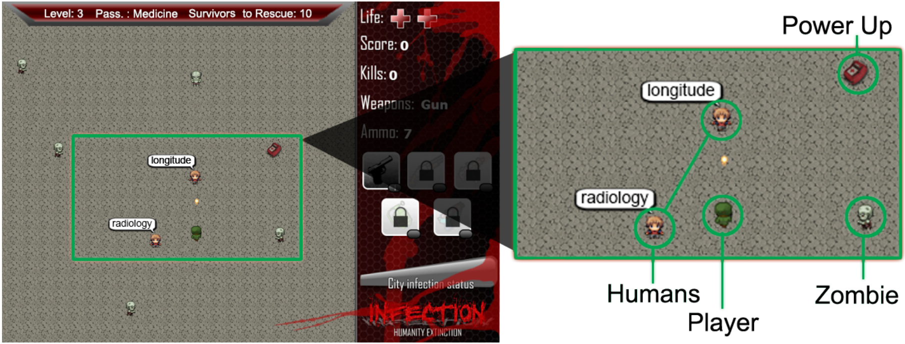
If someone says something random, they're infected!
Take them out before they reach the city and kill everyone.

You are our last hope.

Press ENTER to start...

Source: Vannella et al. (2014)

Case Study: Infection II



Source: Vannella et al. (2014)

- Shout the word related to the given word (medicine)
- If someone says something random, they are infected!

Case Study: Infection III

- Free version of the game yields high-quality annotations with no direct cost (game development is *not* a direct cost)
- Paid GWAP is slightly less cost-efficient than μT on CrowdFlower
- Volunteers make less mistakes and provide more consistent answers

	Infection (Free)	**Infection (Paid)**	**CrowdFlower**
# of Players	89	163	1097
# of Annotations	3 150	3 355	13 764
N-Accuracy	71.0	65.9	n/a
Krippendorff's α	0.445	0.330	0.167
G.S. Agreement	68.1	61.1	59.6
Cost per Annotation	–	\$0.022	\$0.008

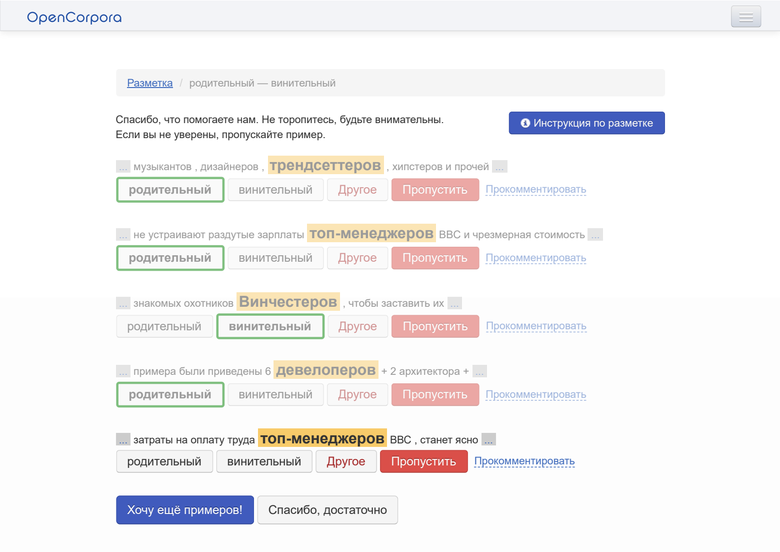
Source: Vannella et al. (2014)

Case Study: OpenCorpora Gamification I

- GWAPs are hard to develop, but we can use gamification techniques in µT
- **OpenCorpora** aims at creation of a large annotated corpus for Russian (Bocharov et al., 2013)
- Freely-available texts are processed by a morphological analyzer, dubious examples are then annotated by humans on a custom µT platform
- Volunteers are invited to perform microtasks for addressing ambiguity

Source: Finnsson (2017)

Case Study: OpenCorpora Gamification II



Спасибо, что помогаете нам. Не торопитесь, будьте внимательны.
Если вы не уверены, пропускайте пример.

... музыкантов , дизайнеров , **трендсеттеров** , хипстеров и прочей ...

... не устраивают раздутые зарплаты **топ-менеджеров** BBC и чрезмерная стоимость ...

... знакомых охотников **Винчестеров** , чтобы заставить их ...

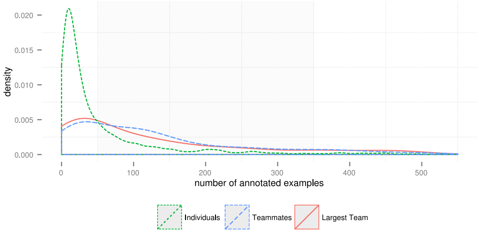
... примера были приведены 6 **девелоперов** + 2 архитектора + ...

... затраты на оплату труда **топ-менеджеров** BBC , станет ясно ...

Source: `http://opencorpora.org/tasks.php`

Case Study: OpenCorpora Gamification III

- Annotators can form a team or contribute individually
- An average *teammate* provided 26 more annotations than an average *individual contributor* (Ustalov, 2015c)
- The largest team had 170 highly-motivated annotators who contributed 76 559 answers



Source: Ustalov (2015c)

Unlike GWAPs, **gamification** is relatively easy to implement and might be a useful tool for attracting annotators.

GWAPs: Wrap-Up

- Creating GWAPs is difficult, but they attract the annotators
- Better when they are free
- Quality control is similar to μT, but also requires anti-cheating mechanisms
- Gamification is a useful option for both WotC and μT (just do not abuse this technique, please)

Source: Amos (2011)

Section 5

Miscellaneous

Which Genre to Choose?

? Can you afford paying the annotators?

! Microtasks (MTurk, etc.)

? Can you spend much time on software development?

! Games with a Purpose or Volunteer-based Microtasks

? Are you aiming at producing the content rather than annotating it?

! Wisdom of the Crowds (Wiki, etc.)

? Is there no way to represent the task for *non-experts* to solve it?

! Do not use Crowdsourcing

Other Crowdsourcing Environments

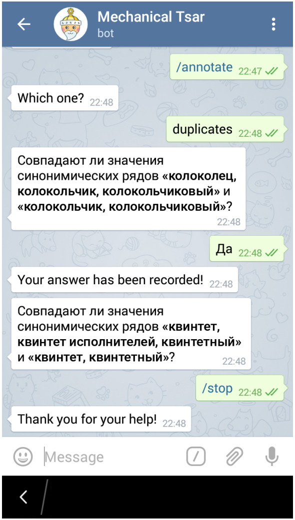
Websites are not the only available medium.

Try something else?

- Instant Messengers (IMs) are an obvious choice
- Modern IMs offer a feature-rich API
- Mobile IMs are growing fast and the crowd is already there!
- “Chatbots with a Purpose”

Teleboyarin is a chatbot proof-of-concept that offers the Mechanical Tsar annotation functionality to IM (Ustalov, 2015b).

Case Study: Teleboyarin



/annotate 22:47

Which one? 22:48

duplicates 22:48

Совпадают ли значения синонимических рядов **«колоколец, колокольчик, колокольчиковый»** и **«колокольчик, колокольчиковый»**? 22:48

Да 22:48

Your answer has been recorded! 22:48

Совпадают ли значения синонимических рядов **«квинтет, квинтет исполнителей, квинтетный»** и **«квинтет, квинтетный»**? 22:48

/stop 22:48

Thank you for your help! 22:48

- We consider an example from Ustalov (2015b)

Journals, Conferences, Books

Journals:

- CSCW, Journal of Collaborative Computing (Springer)

Conferences:

- HCOMP, AAAI Conference on Human Computation and Crowdsourcing
- CSCW, ACM Conference on Computer-Supported Cooperative Work
- OpenSym, International Symposium on Open Collaboration
- LREC, Conference on Language Resources and Evaluation
- HILL, ICML Workshop on Human in the Loop Learning

Books:

- The Practice of Crowdsourcing (Alonso, 2019)
- The People's Web Meets NLP (Gurevych et al., 2013)
- Crowdsourcing (Howe, 2009)

Courses and Tutorials

- **Crowdsourcing & Human Computation**, `http://crowdsourcing-class.org/`
- **Crowdsourcing for NLP**, `http://naacl.org/naacl-hlt-2015/tutorial-crowdsourcing.html`
- **Crowd-Powered Data Mining**, `http://dbgroup.cs.tsinghua.edu.cn/ligl/kdd/`
- **Efficient Data Collection via Crowdsourcing**, `https://research.yandex.com/tutorials/crowd/`

Datasets

- **Wikimedia**, https://www.wikimedia.org/
- **SQUARE** (Sheshadri et al., 2013), http://ir.ischool.utexas.edu/square/
- **TREC Crowdsourcing Track**, https://sites.google.com/site/treccrowd/
- **Appen Open Source Datasets**, https://appen.com/resources/datasets/
- **Yandex.Toloka Open Datasets**, https://toloka.ai/datasets

Section 6

Conclusion

Conclusion

- Crowdsourcing is a great tool for data collection and processing, especially in NLP tasks
- It is a two-sided process with humans on **both** sides
- A few promising research directions:
  - the role of bots (Zheng et al., 2019)
  - active learning (Yang et al., 2018)
  - task design (Bragg et al., 2018)
  - quality control (Daniel et al., 2018)

Source: Free-Photos (2016)

Thank You!

Questions?

Contacts

Dr. **Dmitry Ustalov**

Data Analysis and Research Group
Yandex, Saint Petersburg, Russia

`https://github.com/dustalov`

✉ `mailto:dustalov@yandex-team.ru`

iD `0000-0002-9979-2188`

Revision: `ba9403a`

References I

von Ahn L. and Dabbish L. (2004). Labeling Images with a Computer Game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '04. Vienna, Austria: ACM, pp. 319–326. DOI: `10.1145/985692.985733`.

von Ahn L. et al. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, vol. 321, no. 5895, pp. 1465–1468. DOI: `10.1126/science.1160379`.

Alonso O. (2019). The Practice of Crowdsourcing. Ed. by G. Marchionini. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers. DOI: `10.2200/S00904ED1V01Y201903ICR066`.

Alonso O., Rose D. E., and Stewart B. (2008). Crowdsourcing for Relevance Evaluation. *SIGIR Forum*, vol. 42, no. 2, pp. 9–15. DOI: `10.1145/1480506.1480508`.

Ardila R. et al. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of The 12th Language Resources and Evaluation Conference*. LREC 2020. Marseille, France: European Language Resources Association (ELRA), pp. 4218–4222. URL: `http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.520.pdf`.

Artstein R. and Poesio M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, vol. 34, no. 4, pp. 555–596. DOI: `10.1162/coli.07-034-R2`.

Auer S. et al. (2007). DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007. Proceedings*. Vol. 4825. Lecture Notes in Computer Science. Berlin and Heidelberg, Germany: Springer Berlin Heidelberg, pp. 722–735. DOI: `10.1007/978-3-540-76298-0_52`.

Bernstein M. S. et al. (2010). Soylent: A Word Processor with a Crowd Inside. *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*. UIST '10. New York, NY, USA: ACM, pp. 313–322. DOI: `10.1145/1866029.1866078`. HDL: `1721.1/60947`.

Biemann C. (2013). Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, vol. 47, no. 1, pp. 97–122. DOI: `10.1007/s10579-012-9180-5`.

Bird S., Klein E., and Loper E. (2017). Natural Language Processing with Python. 2nd Edition. O'Reilly Media.

Blumenstock J. E. (2008). Size Matters: Word Count As a Measure of Quality on Wikipedia. *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. Beijing, China: ACM, pp. 1095–1096. DOI: `10.1145/1367497.1367673`.

Bocharov V. et al. (2013). Crowdsourcing morphological annotation. *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*. RGGU, pp. 109–124. URL: `http://www.dialog-21.ru/media/1227/bocharovvv.pdf`.

Bragg J., Mausam, and Weld D. S. (2018). Sprout: Crowd-Powered Task Design for Crowdsourcing. *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. UIST '18. Berlin, Germany: ACM, pp. 165–176. DOI: `10.1145/3242587.3242598`.

References II

Braslavski P., Ustalov D., and Mukhin M. (2014). A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. EACL 2014. Gothenburg, Sweden: Association for Computational Linguistics, pp. 101–104. DOI: `10.3115/v1/E14-2026`.

Callison-Burch C. (2009). Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2009. Singapore: Association for Computational Linguistics and Asian Federation of Natural Language Processing, pp. 286–295. DOI: `10.3115/1699510.1699548`.

Chang J. et al. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*. NIPS 2009. Vancouver, BC, Canada: Curran Associates, Inc., pp. 288–296. URL: `https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf`.

Daniel F. et al. (2018). Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Computing Surveys*, vol. 51, no. 1, 7:1–7:40. DOI: `10.1145/3148148`.

Dawid A. P. and Skene A. M. (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28. DOI: `10.2307/2346806`.

Difallah D. E., Demartini G., and Cudré-Mauroux P. (2013). Pick-A-Crowd: Tell Me What You Like, and I'll Tell You What to Do. *Proceedings of the 22Nd International Conference on World Wide Web*. WWW '13. Rio de Janeiro, Brazil: ACM, pp. 367–374. DOI: `10.1145/2488388.2488421`.

Estellés-Arolas E. and González-Ladrón-de-Guevara F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, vol. 38, no. 2, pp. 189–200. DOI: `10.1177/0165551512437638`.

Esteves D. et al. (2018). Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis. *Journal of Data and Information Quality*, vol. 9, no. 3: *Special Issue on Improving the Veracity and Value of Big Data*, 16:1–16:26. DOI: `10.1145/3177873`.

Finin T. et al. (2010). Annotating Named Entities in Twitter Data with Crowdsourcing. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. CSLDAMT '10. Los Angeles, CA, USA: Association for Computational Linguistics, pp. 80–88. URL: `https://aclweb.org/anthology/W10-0713`.

Finnerty A. et al. (2013). Keep It Simple: Reward and Task Design in Crowdsourcing. *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*. CHItaly '10. Trento, Italy: ACM. DOI: `10.1145/2499149.2499168`.

Gadiraju U. et al. (2019). Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work (CSCW)*, vol. 28, no. 5, pp. 815–841. DOI: `10.1007/s10606-018-9336-y`.

Geiger R. S. and Halfaker A. (2013). When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes? *Proceedings of the 9th International Symposium on Open Collaboration*. WikiSym '13. Hong Kong: ACM, 6:1–6:6. DOI: `10.1145/2491055.2491061`.

References III

Gurevych I. and J. Kim, eds. (2013). The People's Web Meets NLP: Collaboratively Constructed Language Resources. Berlin and Heidelberg, Germany: Springer-Verlag Berlin Heidelberg. DOI: `10.1007/978-3-642-35085-6`.

Halfaker A. and Geiger R. S. (2019). ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. arXiv: `1909.05189 [cs.HC]`. Submitted.

Halfaker A. et al. (2013). The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist*, vol. 57, no. 5, pp. 664–688. DOI: `10.1177/0002764212469365`.

Hosseini M. et al. (2014). The Four Pillars of Crowdsourcing: a Reference: a Reference Model. *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*. Marrakech, Morocco: IEEE, pp. 1–12. DOI: `10.1109/RCIS.2014.6861072`.

Howe J. (2009). Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. New York, NY, USA: Crown Publishing Group.

Jurgens D. and Navigli R. (2014). It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 449–464. DOI: `10.1162/tacl_a_00195`.

Karger D. R., Oh S., and Shah D. (2014). Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems. *Operations Research*, vol. 62, no. 1, pp. 1–24. DOI: `10.1287/opre.2013.1235`.

Kittur A. and Kraut R. E. (2008). Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. CSCW '08. San Diego, CA, USA: ACM, pp. 37–46. DOI: `10.1145/1460563.1460572`.

Krippendorff K. (2018). Content Analysis: An Introduction to Its Methodology. Fourth Edition. Thousand Oaks, CA, USA: SAGE Publications, Inc.

Kumar S., Spezzano F., and Subrahmanian V. (2015). VEWS: A Wikipedia Vandal Early Warning System. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Sydney, NSW, Australia: ACM, pp. 607–616. DOI: `10.1145/2783258.2783367`.

Meyer C. M. et al. (2014). DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. COLING 2014. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 105–109. URL: `https://aclweb.org/anthology/C14-2023`.

Navigli R. and Ponzetto S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, vol. 193, pp. 217–250. DOI: `10.1016/j.artint.2012.07.001`.

References IV

Oleson D. et al. (2011). Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *Human Computation: Papers from the 2011 AAAI Workshop (WS-11-11)*. HCOMP 2011. San Francisco, CA, USA: Association for the Advancement of Artificial Intelligence, pp. 43–48. URL: `https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3995`.

Panchenko A. et al. (2018a). Improving Hypernymy Extraction with Distributional Semantic Classes. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. LREC 2018. Miyazaki, Japan: European Language Resources Association (ELRA), pp. 1541–1551. URL: `http://www.lrec-conf.org/proceedings/lrec2018/summaries/234.html`.

Panchenko A. et al. (2018b). RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*. Moscow, Russia: RSUH, pp. 547–564. URL: `http://www.dialog-21.ru/media/4539/panchenkoaplusetal.pdf`.

Poesio M. et al. (2013). Phrase Detectives: Utilizing Collective Intelligence for Internet-scale Language Resource Creation. *ACM Transactions on Interactive Intelligent Systems*, vol. 3, no. 1: *Special section on internet-scale human problem solving and regular papers*, 3:1–3:44. DOI: `10.1145/2448116.2448119`.

Rajpurkar P., Jia R., and Liang P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL 2018. Melbourne, VIC, Australia: Association for Computational Linguistics, pp. 784–789. DOI: `10.18653/v1/P18-2124`.

Rodrigo E. G., Aledo J. A., and Gámez J. A. (2019). `spark-crowd`: A Spark Package for Learning from Crowdsourced Big Data. *Journal of Machine Learning Research*, vol. 20, pp. 1–5. URL: `http://jmlr.org/papers/v20/17-743.html`.

Sheng V. S., Provost F., and Ipeirotis P. G. (2008). Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, NV, USA: ACM, pp. 614–622. DOI: `10.1145/1401890.1401965`.

Sheshadri A. and Lease M. (2013). SQUARE: A Benchmark for Research on Computing Crowd Consensus. *First AAAI Conference on Human Computation and Crowdsourcing*. HCOMP 2013. Association for the Advancement of Artificial Intelligence, pp. 156–164. URL: `https://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7550`.

Shishkin A. et al. (2020). Text Recognition Using Anonymous CAPTCHA Answers. *Proceedings of the 13th International Conference on Web Search and Data Mining*. WSDM '20. Houston, TX, USA: ACM, pp. 537–545. DOI: `10.1145/3336191.3371795`.

Snow R. et al. (2008). Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP 2008. Honolulu, HI, USA: Association for Computational Linguistics, pp. 254–263. DOI: `10.3115/1613715.1613751`.

References V

Stvilia B. et al. (2008). Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 983–1001. DOI: `10.1002/asi.20813`.

Ustalov D. (2014). Words Worth Attention: Predicting Words of the Week on the Russian Wiktionary. *Knowledge Engineering and the Semantic Web, 5th International Conference, KESW 2014, Kazan, Russia, September 29–October 1, 2014. Proceedings*. Vol. 468. Communications in Computer and Information Science. Cham, Switzerland: Springer International Publishing, pp. 196–207. DOI: `10.1007/978-3-319-11716-4_17`.

Ustalov D. (2015a). A Crowdsourcing Engine for Mechanized Labor. *Proceedings of the Institute for System Programming*, vol. 27, no. 3, pp. 351–364. DOI: `10.15514/ISPRAS-2015-27(3)-25`.

Ustalov D. (2015b). Teleboyarin—Mechanized Labor for Telegram. *Proceedings of the AINL-ISMW FRUCT 2015*, pp. 195–197. URL: `https://www.fruct.org/publications/ainl-abstract/files/Ust.pdf`.

Ustalov D. (2015c). Towards Crowdsourcing and Cooperation in Linguistic Resources. *Information Retrieval: 8th Russian Summer School, RuSSIR 2014, Nizhniy Novgorod, Russia, August 18-22, 2014, Revised Selected Papers*. Vol. 505. Communications in Computer and Information Science. Cham, Switzerland: Springer International Publishing, pp. 348–358. DOI: `10.1007/978-3-319-25485-2_14`.

Ustalov D. et al. (2019). Watset: Local-Global Graph Clustering with Applications in Sense and Frame Induction. *Computational Linguistics*, vol. 45, no. 3, pp. 423–479. DOI: `10.1162/COLI_a_00354`.

Vannella D. et al. (2014). Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, MD, USA: Association for Computational Linguistics, pp. 1294–1304. DOI: `10.3115/v1/P14-1122`.

Wang A., Hoang C. D. V., and Kan M.-Y. (2013a). Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, vol. 47, no. 1, pp. 9–31. DOI: `10.1007/s10579-012-9176-1`.

Wang J., Ipeirotis P. G., and Provost F. (2013b). Quality-Based Pricing for Crowdsourced Workers. Working paper no. 2451/31833. New York University. URL: `https://ssrn.com/abstract=2283000`.

Wang P., Li X., and Wu R. (2019). A deep learning-based quality assessment model of collaboratively edited documents: A case study of Wikipedia. *Journal of Information Science*, pp. 1–16. DOI: `10.1177/0165551519877646`.

Whitehill J. et al. (2009). Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Advances in Neural Information Processing Systems 22*. NIPS 2009. Vancouver, BC, Canada: Curran Associates, Inc., pp. 2035–2043. URL: `https://papers.nips.cc/paper/3644-whose-vote-should-count-more-optimal-integration-of-labels-from-labelers-of-unknown-expertise.pdf`.

References VI

Wilkinson D. M. and Huberman B. A. (2007). Cooperation and Quality in Wikipedia. *Proceedings of the 2007 International Symposium on Wikis*. WikiSym '07. Montréal, QC, Canada: ACM, pp. 157–164. DOI: `10.1145/1296951.1296968`.

Yang J. et al. (2018). Leveraging Crowdsourcing Data for Deep Active Learning An Application: Learning Intents in Alexa. *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, pp. 23–32. DOI: `10.1145/3178876.3186033`.

Zheng L. et al. (2019). The Roles Bots Play in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, 215:1–215:20. DOI: `10.1145/3359317`.

Supplementary Media I

Alexas_Fotos (October 7, 2017). Calculating Machine Resulta Old. Pixabay. URL: https://pixabay.com/photos/calculating-machine-resulta-old-2825179/. Licensed under Pixabay License.

Alves Gaspar J. (2005). Iris - a beautiful kitten two weeks old. Summer 2005. Wikimedia Commons. URL: https://commons.wikimedia.org/wiki/File:Iris_cat.jpg. Licensed under CC BY-SA 2.5, used with author's permission.

Amos E. (December 19, 2011). The Vectrex video game console, shown with controller. Wikimedia Commons. URL: https://commons.wikimedia.org/wiki/File:Vectrex-Console-Set.jpg. Licensed under CC BY-SA 3.0, used with author's permission.

bamenny (February 24, 2016). Robot Flower Technology. Pixabay. URL: https://pixabay.com/photos/robot-flower-technology-future-1214536/. Licensed under Pixabay License.

Bernstein M. S. (May 4–8, 2020). Crowdsourcing and Peer Production. CS 278: Social Computing. Stanford University. URL: http://cs278.stanford.edu/slides/10-crowd-production-peer-production.pdf. Licensed under CC BY-NC-SA 4.0, used with author's permission.

BMaurer (May 27, 2007). This is a modern CAPTCHA, from the reCAPTCHA project. Wikimedia Commons. URL: https://commons.wikimedia.org/wiki/File:Modern-captcha.jpg. Licensed under Public Domain.

Boyd-Graber J. (March 22, 2014). jbg-talks/reading_tea_leaves/figures. URL: https://github.com/ezubaric/jbg-talks/tree/29753e143a479c3ef88d17fb9965ca2a9b85806c/reading_tea_leaves/figures. Used with author's permission.

Buissinne S. (August 25, 2016). Dictionary Reference Book Learning. Pixabay. URL: https://pixabay.com/photos/dictionary-reference-book-learning-1619740/. Licensed under Pixabay License.

Finnsson I. (May 19, 2017). Books Covers Book Case. Pixabay. URL: https://pixabay.com/photos/books-covers-book-case-old-library-2321934/. Licensed under Pixabay License.

Free-Photos (August 9, 2016). Person Mountain Top Achieve. Pixabay. URL: https://pixabay.com/photos/person-mountain-top-achieve-1245959/. Licensed under Pixabay License.

Supplementary Media II

Halfaker A. (November 14, 2015). A descriptive diagram of edits flowing from "The Internet" to Wikipedia depicts the "unknown" quality of edits before ORES and the "good", "needs review", "damaging" labeling that is possible after ORES is made available. Wikimedia Commons. URL: `https://commons.wikimedia.org/wiki/File:ORES_edit_quality_flow.svg`. Licensed under CC BY-SA 4.0, used with author's permission.

Karger D. R., Oh S., and Shah D. (November 8, 2011). Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems. arXiv: `1110.3564v2 [cs.LG]`. Used with author's permission.

Kumar S. (August 15, 2015). Figure showing accuracy of VEWS system in combination with ClueBot NG and STiki. Wikimedia Commons. URL: `https://commons.wikimedia.org/wiki/File:Vews-accuracy-with-cluebot.png`. Licensed under CC BY-SA 4.0, used with author's permission.

McGuire R. (March 24, 2015). Suit Business Man. Pixabay. URL: `https://pixabay.com/photos/suit-business-man-business-man-673697/`. Licensed under Pixabay License.

Merrill B. (July 24, 2014). Pedestrians People Busy. Pixabay. URL: `https://pixabay.com/photos/pedestrians-people-busy-movement-400811/`. Licensed under Pixabay License.

rawpixel (April 18, 2017). Calm Freedom Location. Pixabay. URL: `https://pixabay.com/photos/calm-freedom-location-relax-2218409/`. Licensed under Pixabay License.

rawpixel (June 23, 2018). Agreement Business Businessman. Pixabay. URL: `https://pixabay.com/photos/agreement-business-businessman-3489902/`. Licensed under Pixabay License.

Sarabadani A. (February 16, 2016). Screenshot of ORES extension. Wikimedia Commons. URL: `https://commons.wikimedia.org/wiki/File:ORES_extension_screenshot.png`. Licensed under CC BY-SA 4.0, used with author's permission.

Simone_ph (March 21, 2017). Music Low Electric Bass. Pixabay. URL: `https://pixabay.com/photos/music-low-electric-bass-strings-2149880/`. Licensed under Pixabay License.