Crowdsourcing: (a bit of) theory and ((quite) some) practice

Karën Fort

karen.fort@paris-sorbonne.fr

enetCollect, September 7th, 2017
Where I’m talking from

Main research interests ([http://karenfort.org/](http://karenfort.org/))

- Language resources creation for natural language processing
- Ethics in natural language processing (NLP)
Where I’m talking from

Related actions

Games (with a purpose) I participated in creating:

- Zombilingo
- BISAME
- KRIK!
- SEx

Language games portal and recurring workshop:

Lingo Boingo

Games4NLP
Crowdsourcing: back to basics

Games with a purpose (GWAPs)

GWAP-ing in practice: ZombiLingo

Conclusion
Crowdsourcing: back to basics

Definition

Beyond the myths: "Crowdsourcing is recent"
Beyond the myths: "Crowdsourcing implies a crowd"
Beyond the myths: "Crowdsourcing implies non-experts"

Games with a purpose (GWAPs)

GWAP-ing in practice: ZombiLingo

Conclusion
Crowdsourcing is "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call." [Howe, 2006]

- no \textit{a priori} identification or selection of the participants ("open call")
Crowdsourcing is "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call." [Howe, 2006]

- no *a priori* identification or selection of the participants ("open call")
- massive (in production and participation)
Crowdsourcing is "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call." [Howe, 2006]

- no *a priori* identification or selection of the participants ("open call")
- massive (in production and participation)
- (relatively) cheap
Some remarkable achievements

Wikipedia\(^1\) (September 2017):
- more than **45 million articles** in **241** languages
- more than 8 million views per hour for the English version (2014 - could not find more recent data)

Distributed Proofreaders (Gutenberg Project)\(^2\):
- nearly **40,000** digitalized and corrected books

JeuxDeMots [Lafourcade, 2007]\(^3\):
- more than **150 million** relations in the lexical network
- more than 2 million terms added by the players

\(^1\)http://stats.wikimedia.org/EN/Sitemap.htm
\(^2\)http://www.pgdp.net/c/stats/stats_central.php
\(^3\)http://www.jeuxdemots.org
A simplified taxonomy (more in [Geiger et al., 2011])
A simplified taxonomy (more in [Geiger et al., 2011])
A simplified taxonomy (more in [Geiger et al., 2011])

- Remunerated
  - Direct
  - Indirect
- Not remunerated
A simplified taxonomy (more in [Geiger et al., 2011])
Myth #1: "Crowdsourcing is recent"

*Instructions pour les voyageurs et les employés des colonies*

Citizen science:

- published by the Museum National d’Histoire Naturelle (Paris, France)
- for the travellers and the colonies’ employees:
  
  "to share the results of their own experiences, to benefit themselves and the scientific world"

- first published in 1824
Other examples

- Ligue de Protection des Oiseaux (birds protection league)⁴:
  - monitoring of the birds populations
  - created more than a century ago
  - 5,000 active participants

- Longitude Prize⁵ (1714):
  - awarded by the British government to whoever would invent a simple and practical method allowing to determine a ship’s longitude
  - still exists: in 2014 the theme was ”Global antibiotics resistance”

⁴https://www.lpo.fr
⁵https://longitudeprize.org/
Myth #2: "Crowdsourcing implies a crowd of participants"

Number of players on *Phrase Detectives* according to the number of points gained in the game (February 2011 - February 2012)
A crowd of participants? JeuxDeMots

Number of players on *JeuxDeMots* according to their ranking in the game

A crowd of participants? ZombiLingo
A crowd of workers? [Fort et al., 2011]

Number of active *Turkers* on Amazon Mechanical Turk:

- number of workers registered on the website: more than 500,000
- 80% of the tasks (HIT) are performed by the 20% most active *Turkers* [Deneme, 2009]
  ⇒ really active workers: between 15,059 and 42,912
Experts vs non-experts

Example of the annotation of named entities in a microbiology corpus:

- experts of the domain?
  - of the corpus (microbiology)?
  - of the application (NLP)?
Experts vs non-experts

Example of the annotation of named entities in a microbiology corpus:

- experts of the domain?
  - of the corpus (microbiology)?
  - of the application (NLP)?

→ experts of the task
Crowdsourcing

Using a crowd of "non-experts"?
Crowdsourcing

- Using a crowd of “non-experts”? 

→ Finding/training experts (of the task) in the crowd
Crowdsourcing: back to basics

Games with a purpose (GWAPs)
   Using the (basic) knowledge of the crowd
   Using the basic education of the crowd
   Using the learning capabilities of the crowd

GWAP-ing in practice: ZombiLingo

Conclusion
Games with a purpose (GWAPs)

Allow to benefit from:

1. the knowledge of the "world"
2. the basic education
3. the learning capabilities

of the participants (seldom a crowd)
JeuxDeMots: playing association of ideas...

...to create a lexical network [Lafourcade and Joubert, 2008]

More than 154 million relations (created by approx. 6,000 players), that are constantly updated

- play by pairs
- more and more complex, typed relations
- challenges
- lawsuits
- etc.
Phrase Detectives: playing detective...
...to annotate co-reference [Chamberlain et al., 2008]

3.5M decisions from 45k players
▶ pre-annotated corpus
▶ detailed instructions
▶ training
▶ 2 different playing modes
  ▶ annotation
  ▶ validation (correction of annotations)
FoldIt: playing proteins folding...
...to solve scientific issues [Khatib et al., 2011]

Solution to the crystal structure of a monomeric retroviral protease (simian AIDS-causing monkey virus)

Solution to an issue unsolved for over a decade

- found in a couple of weeks
- by a team of players
- that will allow for the creation of antiretroviral drugs
FoldIt: playing proteins folding...

...without any prior knowledge in biochemistry [Cooper et al., 2010]

Step-by-step training

▶ tutorial decomposed by concepts
▶ puzzles for each concept
▶ access to the following puzzles is given only if your level is sufficient
Crowdsourcing: back to basics

Games with a purpose (GWAPs)

GWAP-ing in practice: ZombiLingo
  - Overview of the game
  - Motivating players
  - Behind the curtain
  - Ensuring quality
  - Results

Conclusion
A complex annotation task

- annotation guidelines
  - 29 relation types
  - approx. 50 pages
- counter-intuitive decisions (not school grammar, linguistics):
  \[ \text{aobj} = \text{au} \]
  
  \[ \text{[...]} \text{ avoir recours au type de mesures [...]} \]

  i.e. head of the PP is the preposition

  \[ \rightarrow \text{decompose} \text{ the complexity of the task [Fort et al., 2012], not simplify it!} \]
Zombi Lingo

Bienvenue à toi, jeune zombie!

Le monde est condamné, ta transformation en zombie a commencé.
Pour survivre, suis mes règles, identifie les têtes et mangle-les.
Attention aux pièges, ils sont nombreux!

Jouer
Pas de limite pour toi! Tu accèdes à toutes les options, bonus cachés!

Karen
Retrouve ici tes statistiques, et compare ton score avec celui de tes amis!

http://zombilingo.org/
Mes ennemis
630
218 934

Mes duels
Duels gagnés : 4
Matchs nuls : 8
Duels perdus : 5

Challenge
Total
1. Nicozombie : 323 393
2. Chouchou : 307 339
3. Methos31 : 290 235
4. Lyco : 143 137
5. Marielbo : 45 390
6. Firey : 18 932
7. Lulu66 : 16 300
8. Teonw00 : 16 152
9. Bos : 14 930
10. Karen : 14 634
11. Neuvk : 14 406

News
Les Pokémon se cachent aussi chez les Zombis ! Depuis ce matin, nous v... lire la suite...

Mon compte
Mot de passe :
Supprimer mon compte.
Parties gagnées : 171
Parties parfaites : 118
Nombre d'objets trouvés : 159
Trouve le complément (objet indirect introduit par "à") du verbe indiqué !

Très jeune, il a fait preuve d'initiative et de courage pour **PARTICIPER** à un sauvetage lors d'inondations.
General features

Bring the fun through:

- zombie design
- use of (crazy) objects
- regular challenges (specific corpus and design) on a trendy topic:
  - Star Wars (when the movie was playing)
  - soccer (during the Euro)
  - Pokemon (well...)
LeaderboardS (for achievers)

Criteria:

- number of annotations or points
- in total, during the month, during the challenge
Hidden features (for explorers)

- appearing randomly
- with different effects: objects, other game, etc.
Duels (for socializers (and killers?))

- select an enemy
- challenge them on a specific type of relation
Badges (?) (for collectors)

- play all the sentences for a relation type, for a corpus
- play all the sentences from a corpus
Organizing quality assurance
Preprocessing data (freely available corpora)

- Unannotated corpus (Wikipedia)
- Ref corpus (Sequoia)
- Play phase
  - Training phase
    - Training (feedback)
  - Evaluation (feedback)
  - Control (feedback)
  - Eval (no feedback)
- Pre annotation with 2 parsers
- Player’s confidence
- EXP Game
- EXP Eval

EXP Training (feedback)
EXP Control (feedback)
EXP Eval (no feedback)
Preprocessing data (freely available corpora)

Pre-annotation with two parsers

1. a statistical parser: Talismane [Urieli, 2013]
2. a symbolic parser, based on graph rewriting: FrDep-Parse [Guillaume and Perrier, 2015]

→ play the items for which the two parsers give different annotations
Training, control and evaluation

Reference: 3,099 sentences of the Sequoia corpus [Candito and Seddah, 2012]

- **REF\textsubscript{Train & Control}** is used to train the players
- **REF\textsubscript{Eval}** is used like a raw corpus, to evaluate the produced annotations

<table>
<thead>
<tr>
<th>REF\textsubscript{Train &amp; Control}</th>
<th>REF\textsubscript{Eval}</th>
<th>Unused</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% 1,549 sentences</td>
<td>25% 776 sentences</td>
<td>25% 774 sentences</td>
</tr>
</tbody>
</table>
Training the players

Compulsory for each dependency relation

- sentences are taken from the REF $^{Train&Control}$ corpus
- a feedback is given in case of error
Dealing with cognitive fatigue and long-term players

Control mechanism

Sentences from the REF $^{Train&Control}$ corpus are proposed regularly

1. if the player fails to find the right answer, a feedback with the solution is given

Ils ont été reçus à la boulangerie Leroy pour visiter le fournil et surtout pétrir la pâte afin de confectionner de délicieux pains au chocolat qu’ils ont dégustés à l’heure du goûter avec un verre de jus de fruit.

Tu as répondu surtout et il fallait répondre visiter
Il te reste 2 essais avant de devoir refaire le tutoriel de ce phénomène
Passer à la phrase suivante
Dealing with cognitive fatigue and long-term players

Control mechanism

Sentences from the REF_{Train&Control} corpus are proposed regularly

1. if the player fails to find the right answer, a feedback with the solution is given
2. after a given number of failures on the same relation, the player cannot play anymore and has to redo the corresponding training

- 1er FÉVRIER 1995 : Jean-Paul Schimpf, un ami intime de Didier Schuller, est arrêté sur un parking, alors que la dirigeante d'une entreprise d'assainissement disait vouloir lui remettre une somme d'argent en liquide.

Tu as répondu une et il fallait répondre arrêté

Tu as un peu oublié comment jouer ce phénomène. Pour continuer à jouer sur celui-ci, tu vas devoir refaire le tutoriel correspondant.
Dealing with cognitive fatigue and long-term players

Control mechanism

Sentences from the REF$_{Train\&Control}$ corpus are proposed regularly

1. if the player fails to find the right answer, a feedback with the solution is given
2. after a given number of failures on the same relation, the player cannot play anymore and has to redo the corresponding training

→ we deduce a level of confidence for the player on this relation
Production: game corpus size
compared to other existing French dependency syntax corpora

As of July 10, 2016

▶ 647 players (1,150 as of Sept. 5th, 2017)
▶ who produced 107,719 annotations (more than 390,000 as of Sept. 5th, 2017)

<table>
<thead>
<tr>
<th></th>
<th>Sequoia 7.0 free validated</th>
<th>UD-French 1.3 free after ZL$^6$ + errors</th>
<th>FTB-UC not &quot;free&quot; validated</th>
<th>FTB-SPMRL not &quot;free&quot; validated</th>
<th>Game free validated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sent.</td>
<td>3,099</td>
<td>16,448</td>
<td>12,351</td>
<td>18,535</td>
<td>5,221</td>
</tr>
<tr>
<td>Tok.</td>
<td>67,038</td>
<td>401,960</td>
<td>350,947</td>
<td>557,149</td>
<td>128,046</td>
</tr>
</tbody>
</table>

+ (ever)growing resource!

$^6$ZL 1.0, July 2014 vs UD 1.0 January 2015.
Evaluating quality on the REF$_{Eval}$ corpus

![Bar chart showing F-measure comparison between Talismane, FrDep-Parse, and Game across different parts of speech and sentence structures.]
Annotation density
on the $\text{REF}_{\text{Eval}}$ corpus

→ need **more** annotations on some relations
Next steps
Expand to new languages and new annotation types

New languages
- English
- less-resourced languages

New annotation types
- part-of-speech (POS),
- corpus building,
- etc.

Alice Millour (PhD student)
Crowdsourcing: back to basics

Games with a purpose (GWAPs)

GWAP-ing in practice: ZombiLingo

Conclusion
Crowdsourcing for language resources creation

Achievements

▶ surprisingly good results in terms of quantity and quality
▶ we demonstrated that we can train people on a complex task

Difficulties

▶ motivating people → language learning could work (7)
▶ communication / advertisement
▶ gamification

7There is no (yet) scientific evidence that it does.
THANK YOU

https://github.com/zombilingo

http://zombilingo.org/export
ZombiLingo’s team and fundings

Bruno Guillaume (researcher)

Nicolas Lefèbvre (engineer)
Appendix

Amazon Mechanical Turk: a platform of legends
Ethical issues
Varying costs
Complexity of the tasks

Bibliographie
History: von Kempelen’s Mechanical Turk

A mechanical chess player created by J. W. von Kempelen in 1770
History: von Kempelen’s Mechanical Turk

In fact, a human chess master was hiding inside to operate the machine
History: von Kempelen’s Mechanical Turk

It’s **artificial** artificial intelligence!
And Amazon created AMT

Amazon created for its own needs a microworking platform and opened it to everyone in 2005 (taking 10% of the transactions (20% now))
Cheap and Fast — But is it Good?
Evaluating Non-Expert Annotations for Natural Language Tasks

Rion Snow†  Brendan O’Connor‡  Daniel Jurafsky§  Andrew Y. Ng†

†Computer Science Dept.  ‡Dolores Labs, Inc.  §Linguistics Dept.
Stanford University  832 Capp St.  Stanford University
Stanford, CA 94305  San Francisco, CA 94110  Stanford, CA 94305
{rion,ang}@cs.stanford.edu  brendano@doloreslabs.com  jurafsky@stanford.edu

[Snow et al., 2008]
AMT: the dream come true?
for NLP

Cheap and Fast — But is it Good?
Evaluating Non-Expert Annotations for Natural Language Tasks

Rion Snow†  Brendan O’Connor‡  Daniel Jurafsky§  Andrew Y. Ng†

†Computer Science Dept.
Stanford University
Stanford, CA 94305
{rion,ang}@cs.stanford.edu

‡Dolores Labs, Inc.
832 Capp St.
San Francisco, CA 94110
brendano@doloreslabs.com

§Linguistics Dept.
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu

[Snow et al., 2008]

It’s very cheap, fast, good quality work and it’s a hobby for the Turkers!
Amazon Mechanical Turk (AMT)

MTurk

Make Money by working on HITs
HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:
- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results from Mechanical Turk Workers
Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:
- Have access to a global, on-demand, 24x7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results
Amazon Mechanical Turk (AMT)

MTurk is a **crowdsourcing** system: work *outsourced* via the Web, done by many people (the *crowd*), here, the *Turkers*.
Amazon Mechanical Turk (AMT)

MTurk is a crowdsourcing, microworking system: tasks are cut into small pieces (HITs) and their execution is paid for by the Requesters.
Amazon Mechanical Turk (AMT)

MTurk is a **crowdsourcing**, **microworking** system: tasks are cut into small pieces (HITs) and their execution is **paid for**.
Amazon Mechanical Turk (AMT)

MTurk is a **crowdsourcing**, **microworking** system: tasks are cut into small pieces (HITs) and their execution is **paid for**.
Is AMT Ethical and/or legal?

Ethics:

- **No identification:** no relation between Requesters and *Turkers* and among *Turkers*
- **No possibility to unionize,** to protest against wrongdoings or to go to court.
- **No minimal wage** (< 2$/hr in average)
- **Possibility to refuse to pay** the *Turkers*
Is AMT Ethical and/or legal?
AMT: a hobby for the *Turkers*?

[Ross et al., 2010, Ipeirotis, 2010] show that:

- *Turkers* are priorly **financially** motivated (91%):
  - 20% use AMT as their primary source of income;
  - 50% as their secondary source of income;
  - leisure is important for only a minority (30%).
- 20% of the *Turkers* spend more than 15 hour a week on AMT, and contribute to 80% of the tasks.
- observed mean hourly wages is **below US$ 2**.

[Gupta et al., 2014]: given that training the workers is impossible on AMT, an important part of the work of the *Turkers* is **hidden**.
Turkers are not anonymous [Lease et al., 2013]

The Turkers’ id corresponds to their Amazon client id
AMT allows to reduce costs

Very low wages $\Rightarrow$ low costs? Yes, but...

- UI development
- creation of protections against spammers
- validation and post-processing costs

Some tasks (for example, in less-resourced languages) generate costs that are similar to the usual costs in the domain, due to a lack of qualified Turkers [Novotney and Callison-Burch, 2010].
Depending on an external platform

Impossible to control:
- the costs
- the participants’ working conditions
- the selection of participants
- the conditions of the experiment
AMT does not allow for the training of the workers:

▶ quality is not satisfactory for complexe tasks (for example, summarizing [Gillick and Liu, 2010])

▶ for some simple tasks, NLP tools produce better results than AMT [Wais et al., 2010].

⇒ **Simplification** of the tasks:

▶ a real textual entailment annotation task (inference, neutral, contradiction) is reduced to 2 phases and 1 question: "Would most people say that if the first sentence is true, then the second sentence must be true?" [Bowman et al., 2015]


Mechanical turk is not anonymous.
Technical report.

Cheap, fast and good enough: automatic speech recognition with non-expert transcription.
In Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), HLT’10, pages 207–215, Stroudsburg, PA, USA. Association for Computational Linguistics.

Who are the crowdworkers?: shifting demographics in mechanical turk.
