Mathieu LAFOURCADE
TEXTE team, LIRMM, U. of Montpellier, France

GWAPS AND COLLECTIVE INTELLIGENCE

THE Jeux de mots EXPERIENCE

October 2015
What are GWAPs?

- Games (supposed to be funny, addictive, pleasant...)
- Designed for
  - Data acquisition
  - Problem solving
- Dubbed as collective intelligence
- Hypothesis

A large number of ordinary people is more efficient than a small number of specialists
Some GWAPs in Biology

- **Foldit**
Some GWAPs in Biology

- Phylo
Some GWAPs in Biology

- **Nightjar**
Some GWAPs in Medicine

- Malaria Spot

![Image of Plasmodium parasite and leukocyte with text: Plasmodium parasite and Leukocyte (white blood cell).]
Some GWAPs in ...

- Artigo
Some GWAPs in NLP

- Wordrobe

The tape did not show the killing, and has not been authenticated. The diplomat was kidnapped off a Baghdad street on Saturday as he stopped to buy a newspaper. Meanwhile, Iraqi police say twin car bombs have killed at least 13 people and wounded nearly 30 others south of Baghdad in the town of Mashruh.
GWAPs... some properties

- Playing well \(\iff\) creating good data
- Beware of various bias
- Difficult to be funny AND efficient

- In general, short life span (many gwaps are dead before long)
- Often over inflated expectation
SEMANTIC ANALYSIS OF TEXTS WITH PROPAGATION ALGORITHMS ON/WITH A LEXICO-SEMANTIC NETWORK = A KNOWLEDGE BASE, A GRAPH WITH WORDS, SENSES/USAGES/RELATIONS
What for?

- Applications needing lexical, common sens and specialized field knowledge
  - Report analysis in medical imaging (Imaios)
  - Offer/demand matching in tourism (Bedycasa)
  - Debate management (SucceedTogether)
  - Class factorization in software eng. (Orange, Berger Levrault)

How?

- Automatically (extracting for corpora)?
  - Knowledge is not always explicitly present in texts not exclusively, not totally – a lot of implicit knowledge
- By hand? Long – (too) costly – normative – static data
... Altogether as a **Lexical Network**

**Nodes**
- Terms, textual segments, NP
- Usages, concepts
- Various symbolic informations

**Relations**
- Typed
- Directed
- Weighted

Free idea associations
- hypernyms – hyponymes – part-of – whole – matter/substance ...
- synonyms – antonyms – locutions – magn/antimagn ...
- agent - patients – instruments – locations – causes/consequences – telic role – temporal values...
Nodes
Terms, textual segments, NP
Usages, concepts
Various symbolic informations

Relations
Typed
Directed
Weighted

free idea associations
synonyms – antonyms – locutions – magn/antimagn …
agent - patients – instruments – locations– causes/consequences – telic role– temporal values…
LEXICAL ACQUISITION WITH GWAPS
(GAME WITH A PURPOSE)

HYPOTHESIS:

FOR THE DEVELOPER, DATA ACQUISITION THROUGH GWAP IS
FREE • FAST • EFFICIENT • PROVIDE NON-NEGOCIATED ANSWERS

AND FOR PLAYERS THE GAME WILL/MUST BE
EXCITING • REWARDING • ADDICTIVE
Donner des IDEES ASSOCIEES au terme qui suit :

lutin

Temps 49 s

Dernier terme proposé : petit

Raffinements possibles :

1. petit (taille)
2. petit (jeune)
3. petit (minuscule)
4. petit (modique)

Ce terme a plusieurs sens ou il en manque ? Demandez de l'aide à vos amis

petit >>
agaçant
taquin
pétulant
espiègle
enfant >
gobelin
gnome
farfadet
tourmenter
esprit follet
Bretagne
être fantastique
lutiner
petit >

JDM PLAY EXAMPLE
Réponses données par kaput : petit • agaçant • taquin • pétulant • espiègle • enfant • gobelin • gnome • farfadet • tourmenter • esprit follet • Bretagne • être fantastique • lutiner • petit

Réponses données par egaillat : grelot • oreille • habit • Schtroumpfs • malicieux • nez • créature légendaire • taquin • oreilles • elfe fée • forêt • nuton • Puck • Père Noël • espiègle • mythologie • légende • vert • esprit follet • bonnet • créature • conte • être imaginaire • gnome • chapeau • petit • troll • korrigan • farfadet

petit • taquin • espiègle • gnome • farfadet • esprit follet

Vous gagnez 300 crédits et 8 point(s) d'honneur

J'aime Soyez le premier de vos amis à indiquer que vous aimez ça.
**INTERACTION MODEL**

**Term + Instruction**

**Creation / Strengthening of Relations**

**Intersection**

**Propositions**

**Game 1**

**Game 2**

**Lexical Network**

**Rewards**

**Player 1**

**Player 2**

08/12/2015 LIRMM - TEXTE
Filtering - matching of player pairs

- Iterated Minimal Consensus (weighting)
- Minimizing noise, maximizing recall (long tail)

Features

- Word pseudo-randomly selected
- Other player(s) unknown during play
- Asynchronous games

Points

- **more** if relation is weak
- **less** if relation is strong
Indirect approaches:

- **Totaki** – a guessing game/AskIt – a question game/...

  - **Totaki**: \{clues\} ⇒ term
    - Player = clue giver (+ optional relation type)
    - Totaki = guesser (lexical network + learning + short term memory)

- Looking for quasi intersection in the lexical network

- **Hypothesis**: if the target term is found
  → the network is properly built/informed for this term
Can we find terms from the clue?
With the 500 riddles of the original game: AKI 494 – humans: 398

Totaki 98,8 %
Humans 79,6 %
Totaki ~ 80 %
Tests over 25 000 games where terms are chosen by players

Utilisateur ~ 43 %
Tests over 300 terms on which players made some guesses (controled env.)
**SOME CONCLUSIONS**

- Since sept. 2007 (8 years)
  - > 23 million relations between 570 000 terms and 80 relation types
  - annotated relations (relevant, possible, not relevant)
    - → the largest network of this type
    - → already used for research and by some companies

- Evaluation
  - Collation of various points of view – negotiated (diko) or not (games)
  - Implicit relations (not present in texts) are captured by instruction forcing (players are invited to be explicit)

---

**gwap** ☹️githubcrowdsourcing

**principles are globally validated for lexical networks**

- Relevant for general knowledge but also for specific domains (good surprise!)
- With GWAP (JDM, Askit, Likelt, ...) but also with contributions (Diko)
- In general, **virtuous loop** is difficult to identify
  - playing well ↔ producing proper data
Some ethical aspects

- many involved players (some with more than 2000 hrs of play)
- no memory in the lexical network of who has made what
  (only temporary storage of games waiting to be fetched)
- players are anonymous (login + pwd + email)
- less than 1% troll / vandalism – corrected as soon as discovered

- The data are made by the crowd...
  ... and should return to the crowd

→ Freely available