The JeuxDeMots Project is 10 Years Old: what Assessments?
What are GWAPs?

- Games (meant to be funny, addictive, pleasant...)
- Designed for
  - Data acquisition
  - Problem solving
- Dubbed collective intelligence
- Core assumption

A large number of ordinary people is more efficient than a small number of specialists
Amazon Mechanical Turk?

- Online crowdsourcing, Microworking
- Legal issues
  - Piece work is not legal in many countries
- Ethical issues
  - Some people try to live from their work for AMT
- Quality issues:
  - Very poor quality (people maximize number of microtasks done)
  - Requires effort and money to check data
    - Not so economical in the end after all...

- see « Amazon Mechanical Turk: Gold Mine or Coal Mine? » by Karen Fort, Gilles Adda, K. Bretonnel Cohen
GWAPs... some properties

- A good player $\iff$ good data
- Beware of various biases
- Difficult to be funny AND efficient

- In general, short life span (many gwaps are dead before long)
- Often the expected results are overestimated
CONTEXT OF THE JEUXDEMOTS PROJECT

SEMANTIC ANALYSIS OF TEXTS WITH PROPAGATION ALGORITHMS ON/WITH A LEXICO-SEMANTIC NETWORK

= A KNOWLEDGE (DATA)BASE, A GRAPH WITH WORDS, SENSES/USAGES/RELATIONS
ACQUISITION OF LEXICAL, TERMINOLOGICAL, ONTOLOGICAL INFORMATION ...

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
Nodes
Terms, textual segment, NP
Usages, concepts
Various symbolic informations

Relations
Typed
Directed
Weighted

diagram content

free idea associations
hypernyms – hyponymes – part-of – whole– mater/substance ...
synonymes – 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 • PROVIDES NON-NEGOCIATED ANSWERS

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

**lutin**

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

Réponses données par egaillat : grelot • oreille • habit • Schtrumpfs • 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.
term + instruction

creation / strengthening of relations

Intersections

propositions

rewards

player 1

lexical network

game 1

term + instruction

confrontation

game 2

player 2
PLAYER ACTIVITY REGULATION

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
More than 1.6 M JDM plays done
Around 30,000 h playing time
Some players have more than 3000 h of playing

What we got:
> 2,000,000 terms + many word forms in the network
> 230,000,000 lexical relations
> 26,000 refined terms and > 70,000 usages
> 870,000 negative relations

never ending learning process
new words, NP, refinements... new relations
Other games

Counter games
From the most general toward the most specific

Logical and statistical blocking because of polysemy - for example:

- livre > lecture
- livre > monnaie
- livre > masse

* Bible *is-a* livre & livre *carac* convertible
=> Bible *carac* convertible
From specific to general

The 3 inference types = detector

- of error in premises (1%)
- of exceptions (< 1%)
- of missing refinements (3%)
- of irrelevant correct relations (3%)

About 93% of the inferred relations are correct and relevant
SOME REFINEMENTS

- For polysemy and word usages
- avocat --r-raff_sem--> avocat>fruit
- avocat --r-raff_sem--> avocat>justice
- grippe --r-raff_sem--> grippe>maladie
- grippe --r-raff_sem--> grippe>virus

- 26 000 termes raffinés
- et > 66 000 usages
SOME REFINEMENTS

- Decision tree, example with frégate
  --r-raff_sem--> frégate>navire
  --r-raff_sem--> frégate>navire>moderne
  --r-raff_sem--> frégate>navire>ancien
  --r-raff_sem--> frégate>oiseau

what’s specific?

WSD = selecting proper refinements
Activation algorithm through relations with other terms
Allow to represent

- exceptions
  - autruche --r-agent-1<0--> voler
- Inductive inferences potentially relevant, but wrong
  - ver de terre --r-agent-1<0--> mordre
- contrastive informations between refinements
  - avocat>fruit --r-agent-1<0--> plaider
  - avocat>justice --r-has-part<0--> noyau>fruit

Negative relations can be used as inhibition in WSD

> 870 000 negative relations in the lexical network
what's specific?

Relation Annotations

- Reification of a relation with new associated informations

- > 4 M relation nodes in the lexical network
Another reification form

lion \(\text{--r\_agent-1 --> dévorer}\)

lion [agent] dévorer

\(\text{-- r\_patient --> gazelle, zèbre}\)

(lion [agent] dévorer) [patient] gazelle

\(\text{-- r\_action lieu --> savane}\)

\(\text{-- r\_manner --> féroce}\)
Lexical resources

- If not too specialized
- Can be built with native speakers
- As many as you can, but a dozen can be enough... if motivated

Not expensive – reliable ‘coz collected data are redundant
Since sept. 2007 (~10 years)
- > 230 millions relations between over 2 000 000 terms and around 100 relation types
- annotated relations (relevant, possible, not relevant)
  → the largest network of this type
  → used for research and by some companies

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

- Relevant for general knowledge but also for specific domains (great surprise!)
- With GWAP (JDM, AskIt, LikeIt, ...) but also with direct contributions
- In general, virtuous circle is difficult to identify
  playing well ↔ producing proper data
EVEN MORE CONCLUSIONS

Some ethical aspects

- many involved players (some with more than 3000 hrs of play)
- no memory in the lexical network of who has made what
  (only temporary storage of games still to be retrieved)
- 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
THANK YOU

Jeux de mots

JEUXDEMOTS.ORG

QUESTIONS