Language resources & Amazon Mechanical Turk
ethical, legal and other issues

Gilles ADDA, Joseph MARIANI

LIMSI/IMMI – CNRS

May 17, 2010
Sorry, no IPR, no copyright!

- new paradigm for LR production: crowdsourcing
- most typical and most used: Amazon Mechanical Turk (AMT)
- ethical and legal problems
- need for regulation
- need for legal expertise
What Is Amazon Mechanical Turk (AMT)?

- **History**: fake automatic chess player called “Mechanical Turk” (1769) by J.W. von Kempelen, a Hungarian engineer, via *combination of machine + human competence*.

- **AMT**: online *crowdsourcing* system (birth: Nov 2nd, 2005) neologistic compound Crowd + Outsourcing concept: tasks traditionally performed by an employee or contractors are outsourced via internet to a large group of people (a crowd) willing to contribute.

- **Artificial artificial intelligence**

- **Requesters** create **Human Intelligence Tasks (HITs)**
  - describe complex tasks as a collection of simple one-time tasks;
  - fix “bonus” payment per task.

- **Turkers** log on AMT, look at the HITs, and register for the HITs they want to perform

- 2010: 400,000 Turkers; 50,000-100,000 HITs any time.
Who are MTurkers?

(Ipeirotis 2008 and 2010, Ross 2009)

- **Origin:**
  - 2008: 76% USA, 8% India, 16% Miscellaneous
  - 2010: 47% USA, **34% India**, 19% Miscellaneous

- **Profile:**
  - **India**: male (70%), very young (71% < 30 years), educated (78% Bachelors degree or higher), single no children (56%)
  - **US**: female (65%), young (70% < 40 years), educated (54% Bachelors degree or higher), single no children (35%)

- **Employment:** 30% unemployed or part-time job
Who are MTurkers?

(Ipeirotis 2008 and 2010, Ross 2009)

- **Income**: India: 55% <$10k/yr US: 67%<$60k/yr (general US 45%)
- **Turn-over**: 69% are new to AMT (< 6 months). The average Turker spends 8 h per week (earns $10 pw), *but* 18% spend > 15 h pw.
- **Motivation**: fun (40% US, 20% India); primary source of income (15% US, 28% India) 18% to “make basic ends meet” (30% among Indians)
- **Conclusion**: no uniform population: large difference between Indian turkers and US Turkers. Many MTurkers do it for fun and sparse money. AMT reward essential to a significant minority (∼ full-time job).
Level of engagement of Turkers

(Ipeirotis, 2010)

- mean 100 Turkers have completed about 30,000 HITs/week
- Turkers with the highest level of engagement: 10% most active have completed 75% of HITs;
Level of engagement of Turkers

(Ipeirotis, 2010)

- mean 100 Turkers have a total income of $1740
- Turkers with the highest level of engagement: 16% most active have earned 70% of the income
Hypothesis: strong correlation with the Turkers with high level of engagement in AMT and the fact that it represents a way to “make basic ends meet” (18%, Ross 2009) or that AMT is their primary source of income (21%, Ipeirotis, 2010). For these Turkers:

- in order to have enough HITs, they accept HITs with low wages (20% complete 84% of the HITs, and earn 72% of the income)
- when a Requester submits a HIT, > 80% of chance that it will be completed by somebody who is considering AMT as a job
- without this minority of active Turkers, the economical model of AMT is uncertain
Ethical Issues

- Crowdsourcing (anywhere outsourcing) :
  rely on competence of (m)any human volunteers.

- Webshoring (anywhere offshoring) :
  delocalization out of guaranteed work contract landscape to increase gains.

- Precise framework for Amazon and requesters vs nothing for MTurkers:
  - very low wages (< 2$/hr), no tax, no job security, no prospect to improve conditions
  - no identification: disassociation of requesters/users
    ⇒ unfair prices, even payment refusal

- researchers rely on low wages and uncertain legality?
Legal Issues

Lack of judicious legal regulation for crowdsourcing

- Job? transcription/translation is certainly a job.
- Active Turkers are considering AMT as a job (see foregoing)
- Are Turkers independent contractors or employees?
- Is Amazon a joint employer?
- No need for taxes? pure hypocrisy
- Is it possible to protect IPR/copyright within AMT?
Employee/Contractor distinction (US)

- the label attached by the parties is not conclusive
- employee/contractor distinction, ”courts must use the existing and long-standing right-of-control test, but apply it with sensitivity to the unique characteristics of Internet-enabled workplaces.” (Gabel 2003)
- attempt to use FLSA (The Fair Labor Standards Act), NLRA (National Labor Relations Act) or OSHA (Occupational Safety and Health Administration) or State laws are not clear and tricky
- “contrary to the expectations of Amazon and many of its Providers, claims to employee status are not presumptively barred or inherently invalid” (Felstiner 2010)
Amazon’s role

- AMT Participation Agreement $\Rightarrow$ Amazon is out of the Requester/Provider relation
- Requesters are difficult to locate/identify $\Rightarrow$ difficult to gain something
- Amazon as joint employer? “a single individual may stand in the relation of an employee to two or more employers at the same time under the FLSA”
Taxes

- if Providers are independent contractors, they should take care of the tax, insurance ... 
- the very low wages implicitly exclude this possibility 
- AMT and Requesters use *educated* people with no retribution to the states. 
- governments ask for payment of tax on income generated from work performed over the Internet. (www.washingtonpost.com 02/05/2010)
IPR in AMT

- the lack of direct link between requester/provider, makes any protection impossible
- if a requester has signed some non-distribution agreement or for research only, he may face legal problems, because he cannot guarantee where and by whom the data has been delivered
- the legal status of the data produced using AMT is not clear, because the legal status of the Turkers is not clear
Conclusion

- Employers and employees would benefit from clear standard and guidelines
- Researchers making use of crowdsourcing should do some pioneer work to define some regulations
- Legislature, administrative agencies, courts may act to clarify the situation
AMT paradigm: an opportunity for Language Resources

- Growing number of NLP papers related to AMT (see NAACL workshop (06/06/2010) Creating Speech and Language Data With Amazon’s Mechanical Turk

- Domains of NLP: word sense disambiguation, word similarity, textual entailment, machine translation, speech transcription . . .

- Human expert quality (via combination of answers).

- Very low cost: $10 to transcribe 1 hour of speech, $1,000 to create 4 reference translations for 2000 sentences.

⇒ Development of missing LRs in many domains, and implementation of “human-in-the-loop” models and evaluation (e.g HTER)
“Prediction is very difficult, especially if it is about the future” *Niels Bohr*, *in (Callison-Burch, 2010)*

**AMT vs. usual scheme: will we have the choice?**

1. Very low cost LR production $\Rightarrow$ less money to develop LRs
2. Less money $\Rightarrow$ no choice except AMT-like paradigm for LR

**Future of LR agencies?**

1. Very easy to develop a corpus $\Rightarrow$ no need to centralize LR production
2. No need to pay an overhead $\Rightarrow$ LR agencies will be only distribution agencies
Some practical propositions

CrowdFlower and Samasource are service providers which use the Mechanical Turk platform

- minimum wage (CrowdFlower (crowdflower.com) sets a $2/hr minimum)
- for low-resourced language see SamaSource (www.samasource.org) in the third-world
- IRB-like (institutional review board to get permission to do research with human subjects) for AMT.
references LR production

- Chris Callison-Burch & Mark Dredze (06/06/2010), *NAACL 2010 Workshop Creating Speech and Language Data With Amazon’s Mechanical Turk*
  sites.google.com/site/amtworkshop2010/
  www.cs.jhu.edu/~mdredze/publications/amt_overview.pdf

- Matthew Marge, Satanjeev Banerjee, and Alexander I. Rudnicky, 2010 *using the Amazon Mechanical Turk for transcription of spoken language*, ICASSP 2010
  www.cs.cmu.edu/~mrmarge/icassp10margem.pdf
references Demographics
