Novel Incentives in Language Resource Development

Christopher Cieri
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
{ccieri} AT ldc.upenn.edu
Resource Gaps and Implications

- LR supply still far short of demand
  - in the average and for every human language
  - remains an impediment to HLT development (Choukri)

- MetaNet 2010 LRs->HLTs to prevent EU language digital extinction
  - no language, not even English, enjoys the full range
  - *21/30 European languages could become extinct in the digital world*

- chronic interpreter shortage in crisis zones
  - International Association of Conference Interpreters (2008)

- “*Effective communication in Haiti was confronted by language barriers and the limited utilization of technology*”
  - Harvard Humanitarian Initiative 2011

- growing need for greater translingual capability in counseling
  - American Psychological Association 2010
Varma et al. (2011) used NLP to filter tweets, with 80% accuracy, according to whether they provided situational awareness.

However, system required training data annotated
- situational awareness
- subjectivity
- formality
- personal versus impersonal viewpoint

Processing included POS tagger generally absent from most low resource languages as are:
- tokenizer
- list of stop words
- unigram and bigram frequencies
- perhaps even the text from which to derive them
The Dirty Secret & a Bold Prediction

- Current approaches will not come close to
  - creating full range of LRs, or even a respectable subset
  - for world’s 7097 languages, or even respectable subset
  - within the foreseeable future
- not because they are inefficient (though some are)
- but because they employ a finite resource to address a nearly infinite problem.

By implementing novel incentives we harness the renewable resources of the human drives to learn, compete, enjoy and make meaningful contributions.
Incentives Aware Model of LR Development

Figure 1: Different incentives attract different workforces that require different tasking and workflows and produce different outputs.
Nick Campbell: data collections supporting development of systems capable of producing expressive speech

experimented with multiple incentives

adjusted to the different characteristics of the output

- monetary compensation
- access to resulting data for research purposes,
- sustenance
- curiosity
- fun
- ability to keep the recording device used
- opportunities for unusual social interactions
  - apparent conversations with a robot
  - even more exotic: extended interactions with colleagues outside the lab

acquiring some product, service can lead customers to provide vast quantities of ‘data’ to HLT researchers in industry
WS: Collection

- Mitsuzawa et al. process product/company reviews
- industrial developers = reduced train-test mismatch
- Incentives
  - communicate dissatisfaction
  - points that convert into monetary value, based on volume, quality
- mixture of incentives yields variation in data
Greenfield, Chan, Campbell experiment with crowdsourcing annotation to support information extraction

Incentives
- some Turkers motivated by quality of the interface
- desire to maintain a high approval rating
- as well as the monetary incentives

By focusing on interface design they elicit higher quality data while avoiding pernicious problems common in Mturk
Poesio et al. describe Phrase Detectives
GWAP for collecting anaphora annotation
Incentives
- entertainment
- interesting source material
- variable point system
- opportunity to progress through experience levels,
- leaderboards
- social motivations of teaming with friends in FB version
- prizes awarded via a lottery, according to performance.
Tyson et al. automate link discovery among About.com texts

Incentives
- corporate mission of recirculating users
- content creators different motivations => fewer links than desired

addressed through
- automated techniques
- additional human annotation
WS: Exploitation

- Eskenazi et al. describes dialog system R&D
- Incentives
  - automated access to information
  - improved customer experience in real world interactions
- challenging levels of noise, variation in speech
- extended notion of novel incentives to research community
- free access to data, system
- outreach activities
- attract researcher cycles to problems of interest to them
- “optimization for lab test subjects may not reflect the outcome with real users”.

Great Language Game (GLG)
- contributors hear short audio clips randomly, from 80 languages
- indicate what language is spoken
- released corpus of 16 million judgments w/l 1 year
- incentives
- information
- entertainment
- competition
- status

However, not directly useful for LRE
- relies on ability to identify correct answers
- language know
- each new judgment adds little information about confusability

Developer moved on
What language is this?

Lives: 3

Score: 0

Danish  Marathi  Telugu  Vietnamese
LibriVox creates “free public domain audiobooks” by recruiting, training and organizing volunteers who record themselves reading literary works out of copyright in US.

LibriVox catalog (3/25/2016)
- 10,185 books comprising at least, 57,369 hours of read speech
- estimated cost to reproduce with monetary incentive: $28 million

Incentives
- LibriVox mission, open source, free culture movements
- enjoy reading aloud, expanding a family activity
- maintain the art of storytelling
- collaborating with others of similar interests
- ability to control the size of their own contributions
- develop or maintain skill
- opportunity for paid work
Figure 1: LibriVox Hours Recorded per Language on a log(10) scale
Zooniverse citizen science portal

Tasks include
- identifying signs of movement in star fields
- classifying animal species based on photographs
- transcribing museum records for insect specimen collections

Incentives
- contribute to research most of which is in the hard sciences
- beautiful interfaces attract participants
- fine grained tasking, complete meaningful tasks in minutes

800,000 volunteers
- contributed data to peer-reviewed publications
- serendipitous discoveries of astronomical objects.
Welcome to Snapshot Serengeti

Hundreds of camera traps in Serengeti National Park, Tanzania, are providing a powerful new window into the dynamics of Africa’s most elusive wildlife species. We need your help to classify all the different animals caught in millions of camera trap images.

Season 1
Season 2
Season 3
Season 4
Season 5
Season 6
Season 7
Season 8
Lost Season
Season 9
Conclusion

- Novel Incentives improve our toolkit for developing LRs
- Today’s papers identify the vanguard among our colleagues
- Other fields provide model, methods to consider
- Joint the discussion