science foundation ireland
fondúireacht eolaíochta éireann

Using Crowdsourcing for Labelling Emotional Speech Assets

Alexey Tarasov, Charlie Cullen, Sarah Jane Delany

Digital Media Centre
Dublin Institute of Technology

W3C Workshop on Emotion Language Markup - Oct 2010

Project Introduction

- Science Foundation Ireland funded project
- Objective:
  - prediction of levels of emotion in natural speech
- 2 strands:
  - acoustic analysis (Dr Charlie Cullen - DIT MmIG)
  - machine learning prediction (Dr Sarah Jane Delany - DIT AIG)
- 4 year project, started in October 2009
  - 2 PhD students

Requirements for Supervised Learning

- Performance of supervised learning techniques depends on the quality of the training data
- Requirements:
  - High quality speech assets
  - Good labels

Starting Point...

- Emotional speech corpus [Cullen et al. LREC 08]
  - natural assets
  - use of Mood Induction Procedures
  - high quality recording
    - participants recorded in separate sound isolation booths
  - contextual or meta data is recorded where available
    - based on IMDI annotation schema

Next Steps...

- Need to rate these assets...
- Challenges:
  - manual annotation can be expensive and time consuming
  - experts often disagree
  - expertise does not necessarily correlate with experience

Consider Crowdsourcing?

Crowdsourcing

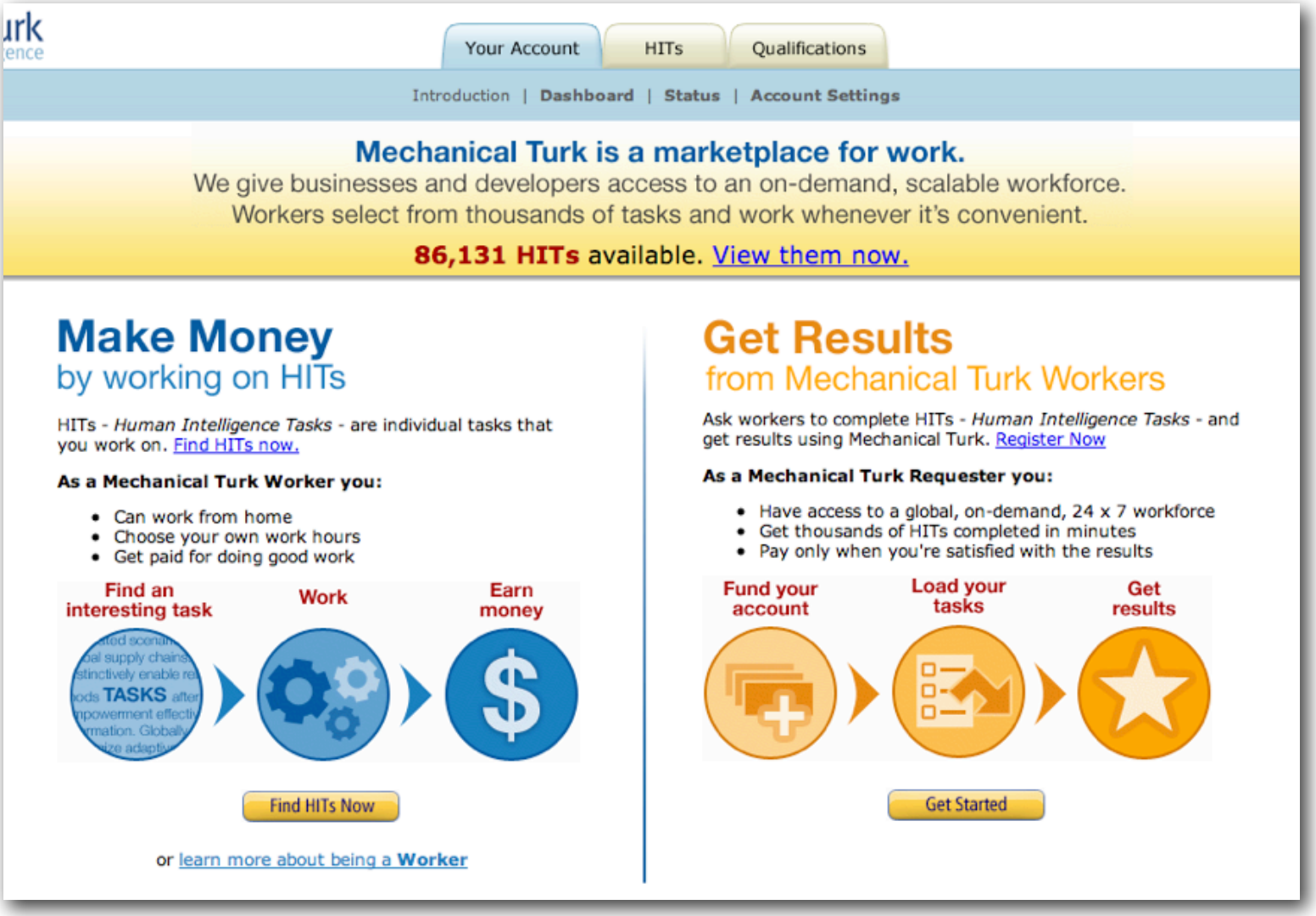
"The act of taking a task traditionally performed by a designated agent and outsourcing it to an undefined, generally large group of people in the form of an open call" [Jeff Howe]

Crowdsourcing

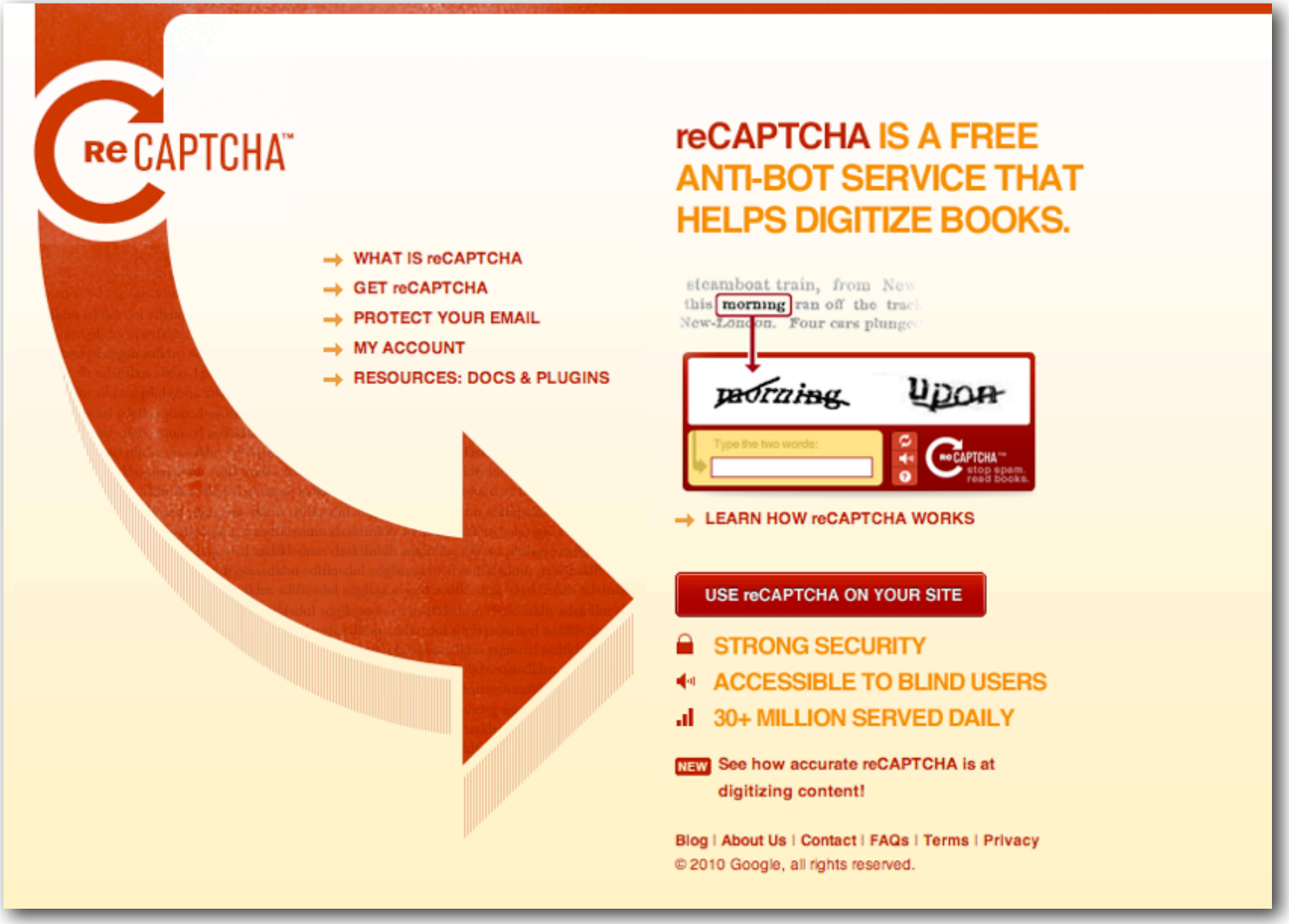
- June 2006 Wired magazine article by Jeff Howe

…the power of many…

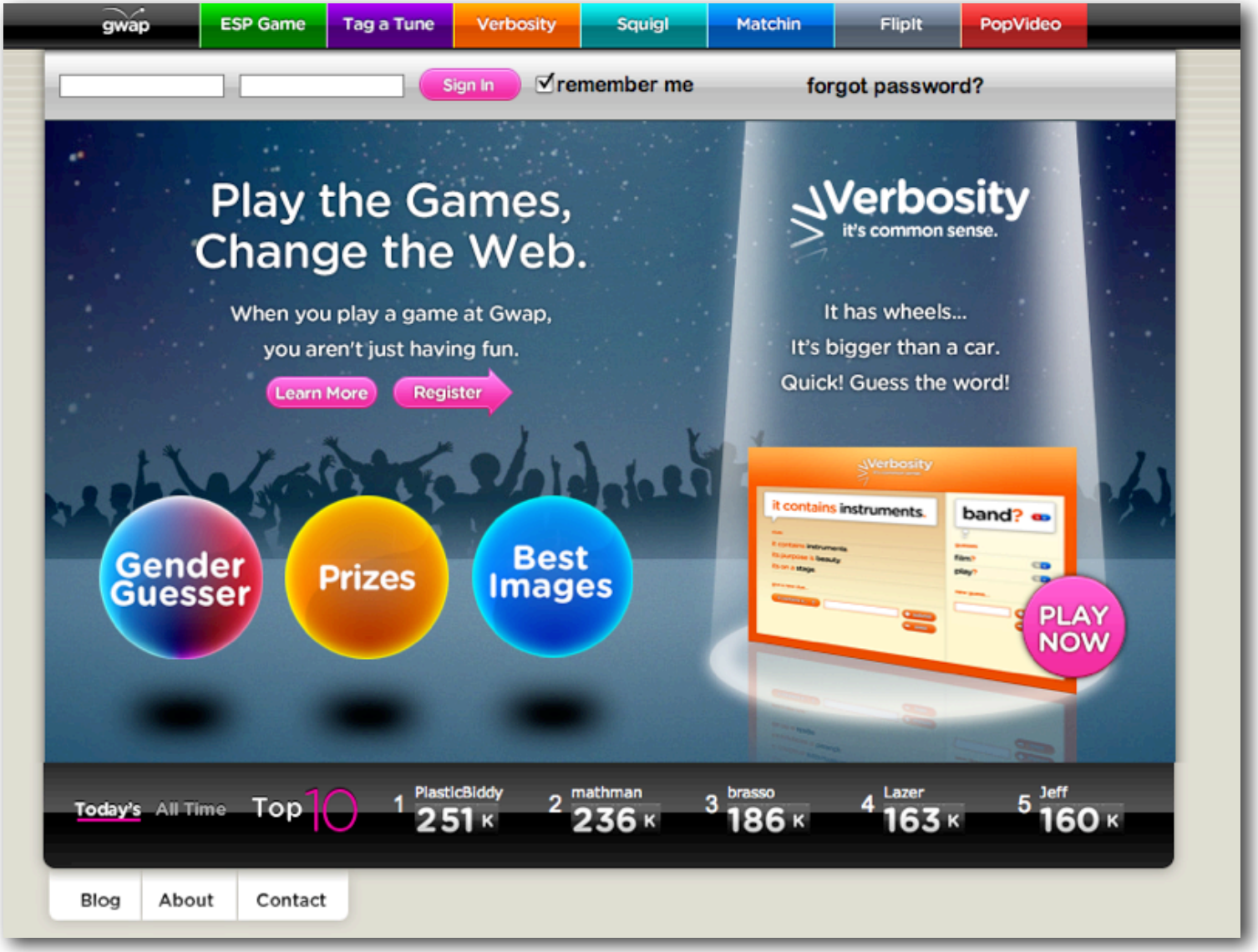
www.wired.com/wired/archive/14.06/crowds.html



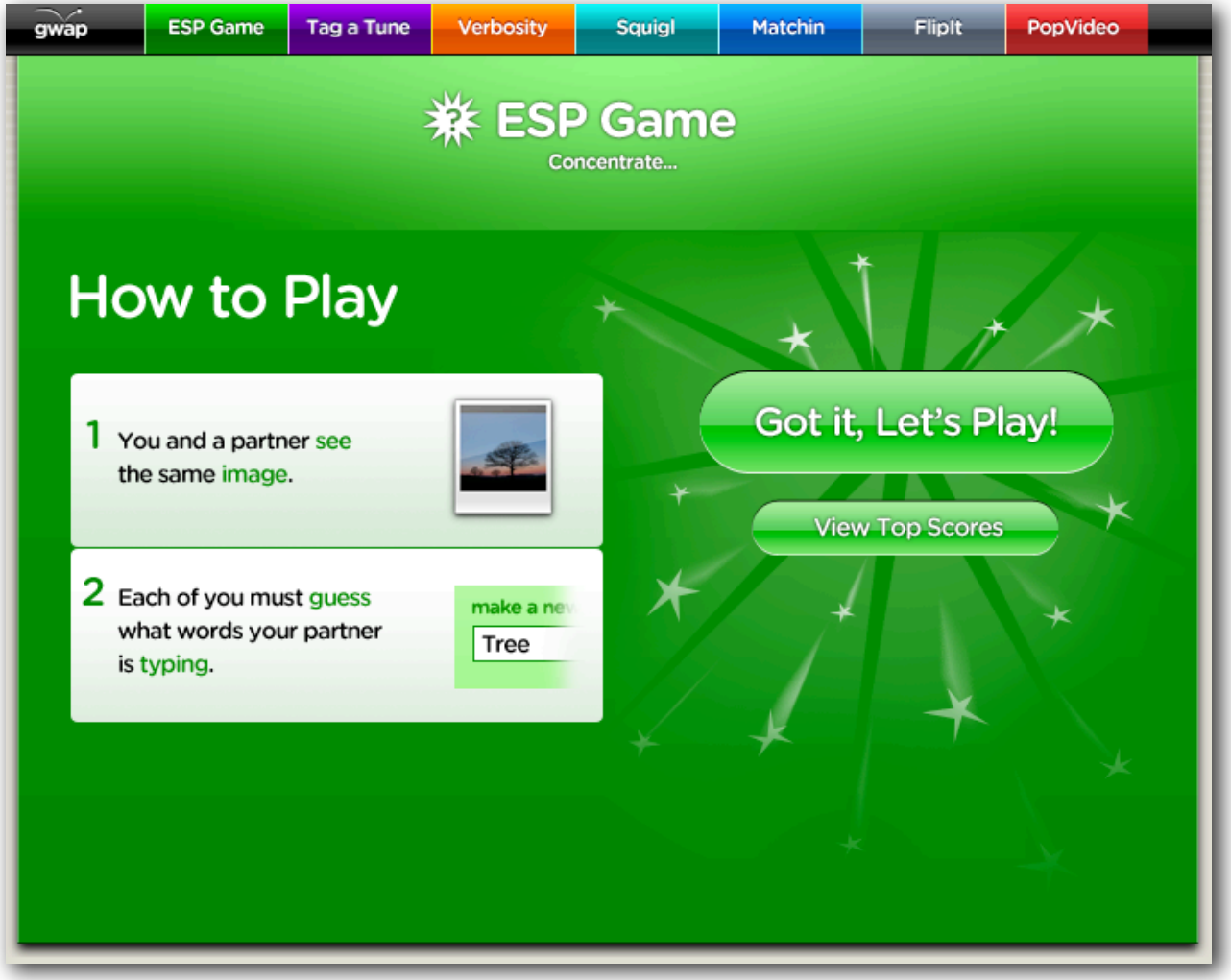
https://www.mturk.com/mturk/



www.google.com/recaptcha



www.gwap.com/gwap/


gwap
ESP Game
Tag a Tune
Verbosity
Squigl
Matchin
FlipIt
PopVideo
ESP Game
Concentrate...
How to Play
1 You and a partner see the same image.
2 Each of you must guess what words your partner is typing.
make a ne
Tree
Got it, Let's Play!
View Top Scores

Crowdsourcing

- Triggered a shift in the way labels or ratings are obtained in variety of domains:
  - natural language tasks [Snow et al. 2008]
  - computer vision [Sorokin & Forsyth 2008, vonAhn & Dabbish 2004]
  - sentiment analysis [Hsueh et al. 2008, Brew et al. 2010]
  - machine translation [Ambati et al. 2010]

Practical Experiences

- Speed
  - 300 annotations from each of 10 annotators in < 11 mins [Snow et al. 2008]
  - evidence that obtaining 'quality' annotations effects time (avg completion time 4 mins vs 1.5 mins) [Kittur et al. 2008]

Practical Experiences

- Quality
  - 875 expert-equivalent affect labels per $1
    [Snow et al. 2008]
  - by identifying 'good' annotators accurate labels can be achieved with significant reduction in effort

    [Donmez et al. 2008, Brew et al. 2010]

Challenges

- How to
  - select which assets are presented for rating?
  - estimate the reliability of the annotators?
  - ensure the reliability of the ratings?
  - select training data for the prediction systems?
  - maintain the balance between consensus and data coverage?

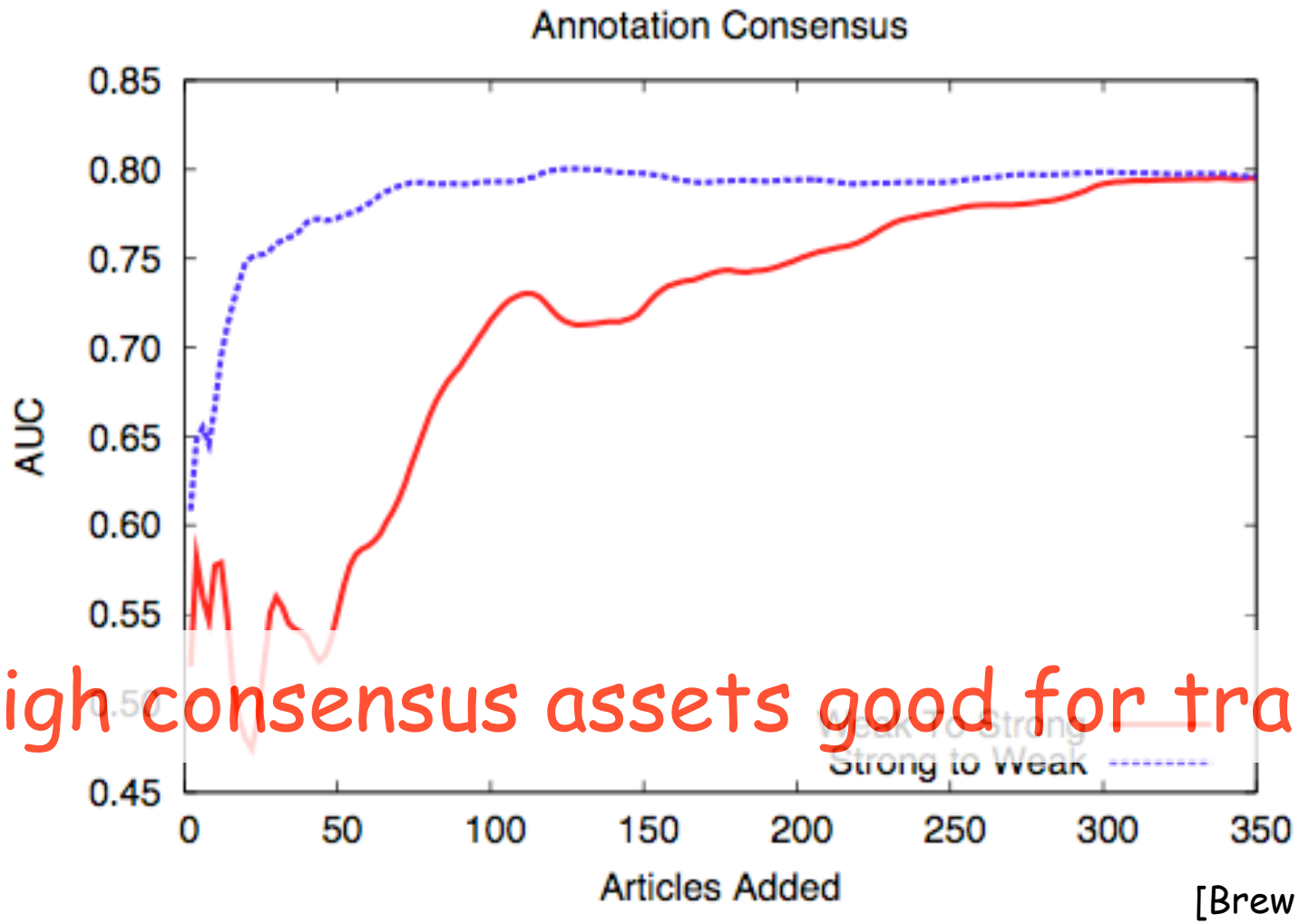
Asset Selection

- Active Learning used by [Ambati et al. 2010, Domnez et al. 2009]
  - a supervised learning technique which selects the most informative examples for annotation
- Clustering used by [Brew et al. 2010]
  - grouping examples and selecting representative examples from cluster to annotate

Annotator Reliability

- Depends on whether annotators are identifiable or not...
- Strategies for recognising strong annotators
  - 'Good' Annotators those that 'agree' with the consensus rating [Brew et al. 2010]
  - Iterative approach to filter out weaker annotators [Domnez et al. 2009]

Good Annotators are Useful...



high consensus assets good for training...

[Brew et al. 2010]

Deriving ratings

- Use consensus rating [Brew et al. 2010]
  - select the rating with highest consensus
  - thresholds can apply
- Only use good annotators to derive rating [Domnez et al. 2009]
- Using learning techniques to estimate 'ground truth' from multiple noisy labels [Smyth et al. 1995, Raykar et al. 2009/10]

Consensus vs. Coverage

- Is it better to label more assets or get more labels per asset?
  - Research suggests fewer annotations are needed in domains with high consensus [Brew et al. 2010]

Reliability of the Ratings

- Evidence of 'gaming' with crowdsourcing services
  - numbers of untrustworthy users is not large
- Techniques
  - require users to complete a test first [Ambiati et al. 2010]
  - use percentage of previously accepted submissions [Hsueh et al. 2008]
  - include explicitly verifiable questions [Kittur et al. 2008]

Use Case

"Seán has a set of speech assets extracted from recordings of experiments using mood induction procedures.

He wants to get these assets rated on a number of different scales, including activation and evaluation, by a large number of non-expert annotators.

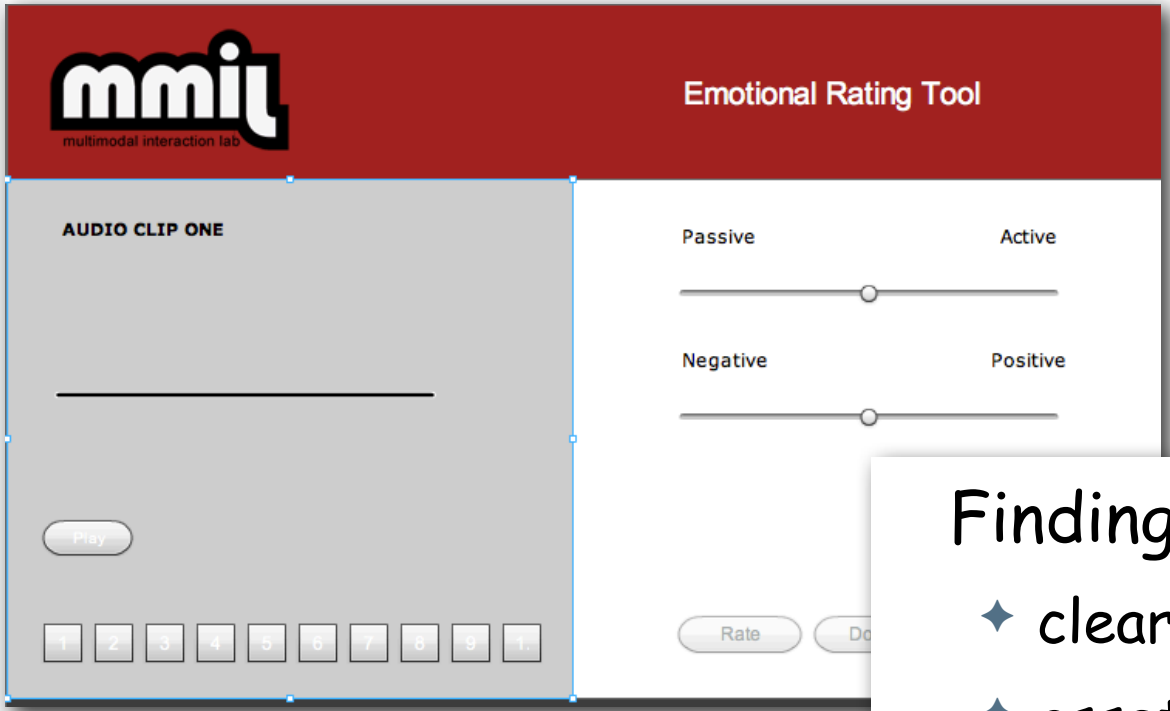
He wants to use a micro-task system such as Mechanical Turk to get these ratings.

Active learning will be used to select the most appropriate assets to present for labels from the annotators.

He will then analyse and evaluate different techniques for identifying good annotators and determining consensus ratings for the assets which will be used as training data for developing prediction systems for emotion recognition."

Experience in our group

- Preliminary rating using crowdsourcing [Brian Vaughan]



Findings

- clear instructions
- asset selection strategy
- payment amounts

References

- V. Ambati, S. Vogel, and J. Carbonell. Active Learning and Crowd-Sourcing for Machine Translation. In Proc. of LREC'10, pages 2169–2174, 2010.
- A. Brew, D. Greene, and P. Cunningham. Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In Proc. of PAIS 2010, pages 1–11. IOS Press, 2010.
- J. Howe. Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. Crown Business, 2008.
- A. Kittur, E. Chi, and B. Suh. Crowdsourcing for Usability: Using Micro-Task Markets for Rapid, Remote, and Low-cost User Measurements. Proceedings of CHI 2008.
- V.C. Raykar, S. Yu, L.H. Zhao, A. Jerebko, C. Florin, G.H. Valadez, L. Bogoni, and L. Moy. Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit. In Proc. of ICML-2009, pages 889–896, 2009.
- V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from Crowds. Journal of Machine Learning Research, 11:1297–1322, 2010.
- P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring Ground Truth from Subjective Labelling of Venus Images. Advances in neural information processing systems, 7:1085–1092, 1995.
- R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 254–263. ACL, 2008.
- A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In Proc. of CVPR 2008, pages 1–8, 2008
- L. von Ahn and L. Dabbish. Labelling Images with a Computer Game, In Procs of CHI 2004