Classification: Evaluation and Annotation

CS 585, Fall 2017

Introduction to Natural Language Processing

http://people.cs.umass.edu/~brenocon/inlp2017

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

Outline

- How to evaluate?
  - Accuracy, Precision, Recall
- How to tune hyperparameters? (e.g. the pseudocount in NB)
  - Train vs Dev/Tuning vs Test set
  - Or: cross-validation
- Where do these labels come from anyway?
  - Naturally-occurring labels
  - Human-annotated labels: very important!!
  - Agreement rates: are my labels any good?
  - Crowdsourcing

Where to get labels?

- Natural annotations
- New human annotations
  - Yourself
  - Your friends
  - Pay people -- e.g. through crowdsourcing sites
    - Mechanical Turk
    - Crowdflower
    - (For larger/more expensive tasks: Upwork/ODesk)

Interannotator agreement (IAA)

- How "real" is a task? Replicable? Reliability of annotations?
- How much do two humans *agree* on labels?
  - Difficulty of task. Human training? Human motivation/effort?
  - Goal: get the human performance upper bound
- If some classes predominate, raw agreement rate may be misleading
  - Chance-adjusted agreement: Cohen's kappa for a pair of human annotators

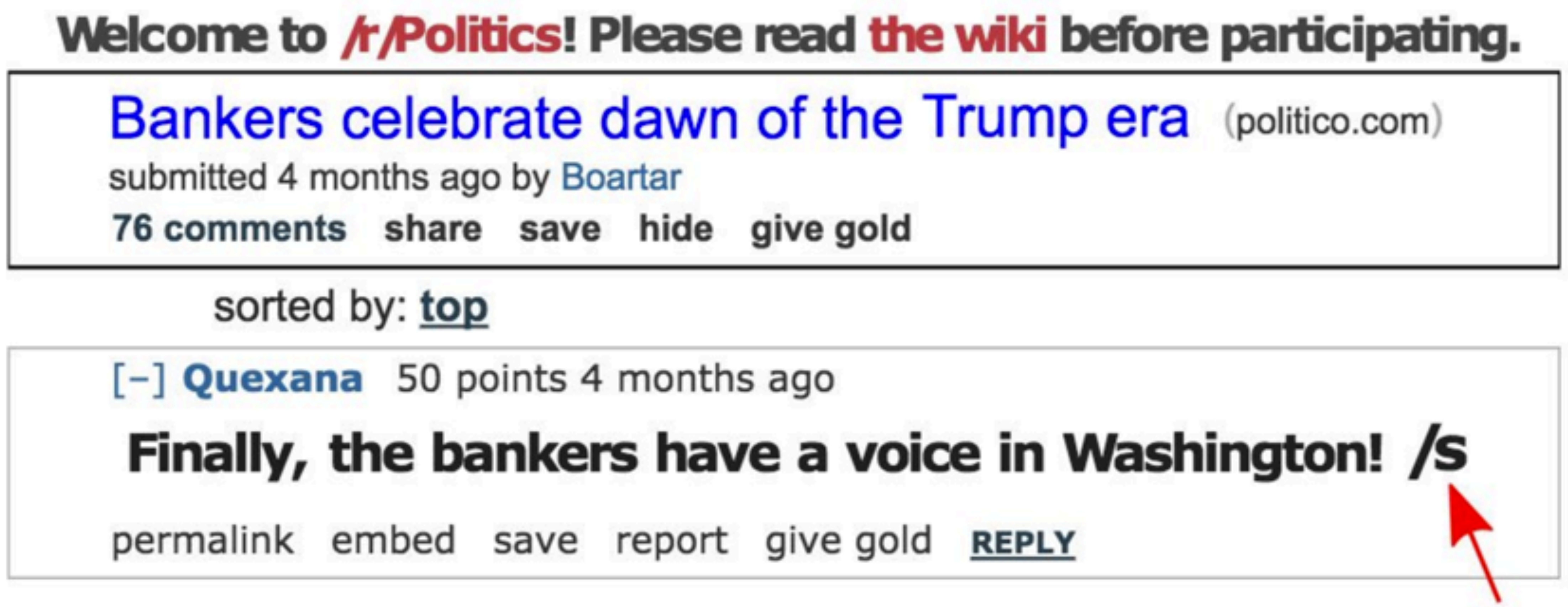
Cohen's kappa

p_o: observed agreement rate

p_e: agreement rate by chance

$$\frac{p_o - p_e}{1 - p_e}$$

- Reliability analysis: from the social sciences, especially psychology, content analysis, communications, etc.



Welcome to /r/Politics! Please read the wiki before participating.

Bankers celebrate dawn of the Trump era (politico.com)
submitted 4 months ago by Boartar

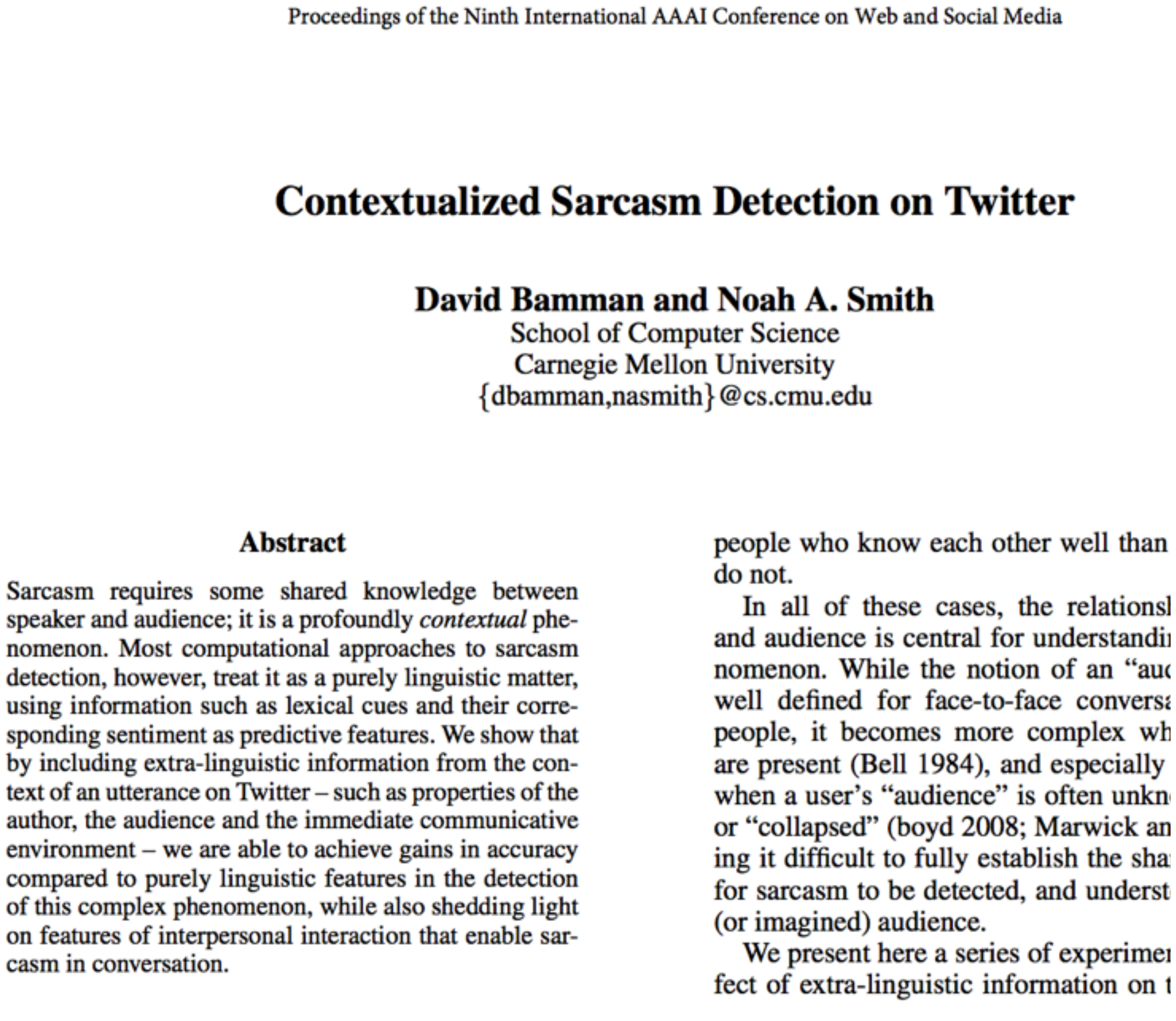
76 comments share save hide give gold

sorted by: top

[–] Quexana 50 points 4 months ago

Finally, the bankers have a voice in Washington! /s

permalink embed save report give gold REPLY



Proceedings of the Ninth International AAAI Conference on Web and Social Media

Contextualized Sarcasm Detection on Twitter

David Bamman and Noah A. Smith
School of Computer Science
Carnegie Mellon University
{dbamman,nasmith}@cs.cmu.edu

Abstract

Sarcasm requires some shared knowledge between speaker and audience; it is a profoundly *contextual* phenomenon. Most computational approaches to sarcasm detection, however, treat it as a purely linguistic matter, using information such as lexical cues and their corresponding sentiment as predictive features. We show that by including extra-linguistic information from the context of an utterance on Twitter – such as properties of the author, the audience and the immediate communicative environment – we are able to achieve gains in accuracy compared to purely linguistic features in the detection of this complex phenomenon, while also shedding light on features of interpersonal interaction that enable sarcasm in conversation.

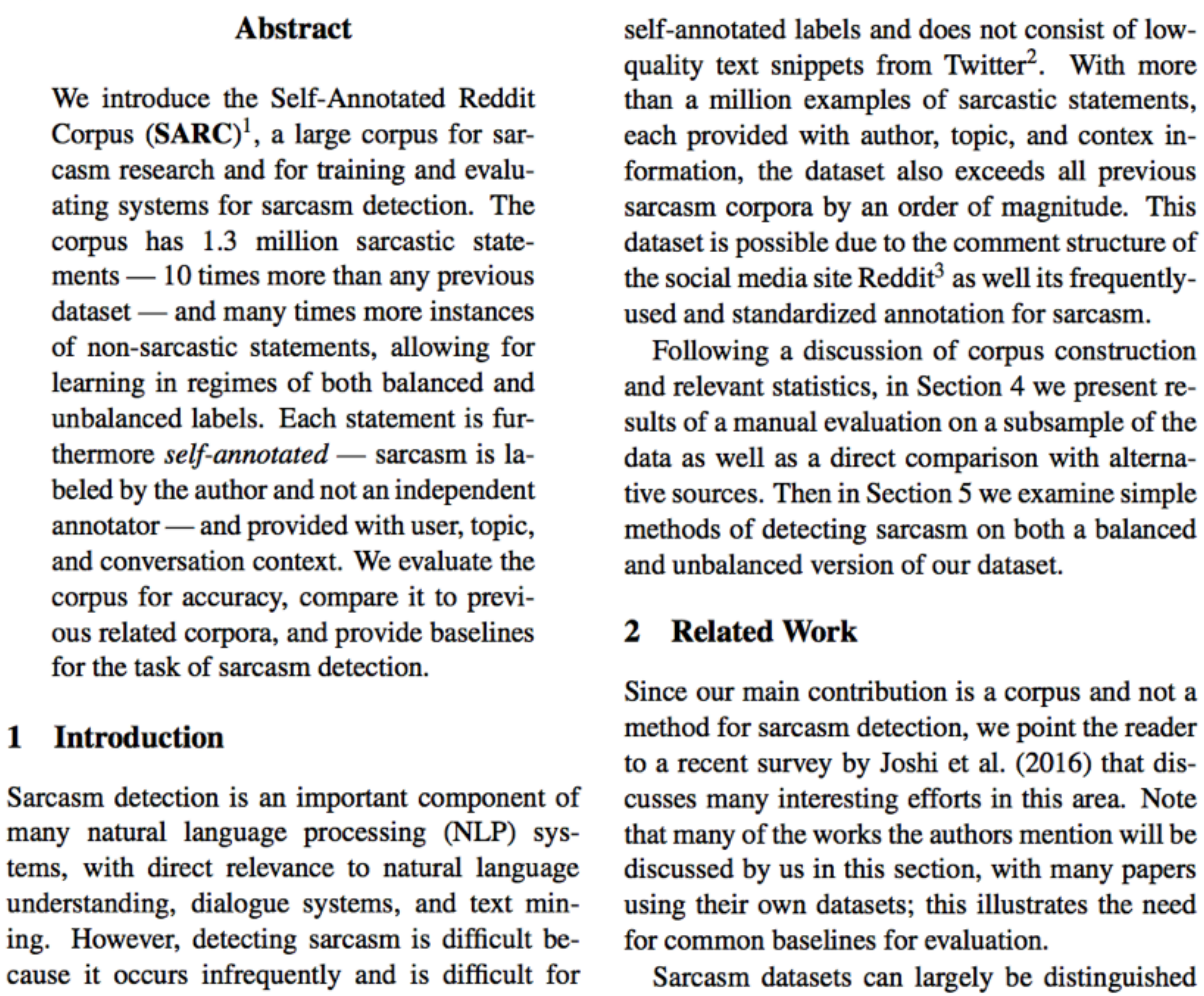
people who know each other well than do not.

In all of these cases, the relationsl and audience is central for understandi nomenon. While the notion of an "auc well defined for face-to-face conversa people, it becomes more complex wh are present (Bell 1984), and especially when a user's "audience" is often unkn or "collapsed" (boyd 2008; Marwick an ing it difficult to fully establish the sha for sarcasm to be detected, and underst (or imagined) audience.

We present here a series of experime fect of extra-linguistic information on t

A Large Self-Annotated Corpus for Sarcasm

Mikhail Khodak and **Nikunj Saunshi** and **Kiran Vodrahalli**
Computer Science Department, Princeton University
35 Olden St., Princeton, New Jersey 08540
{mkhodak,nsaunshi,knv}@cs.princeton.edu



Abstract

We introduce the Self-Annotated Reddit Corpus (**SARC**)[1], a large corpus for sarcasm research and for training and evaluating systems for sarcasm detection. The corpus has 1.3 million sarcastic statements — 10 times more than any previous dataset — and many times more instances of non-sarcastic statements, allowing for learning in regimes of both balanced and unbalanced labels. Each statement is furthermore *self-annotated* — sarcasm is labeled by the author and not an independent annotator — and provided with user, topic, and conversation context. We evaluate the corpus for accuracy, compare it to previous related corpora, and provide baselines for the task of sarcasm detection.

1 Introduction

Sarcasm detection is an important component of many natural language processing (NLP) systems, with direct relevance to natural language understanding, dialogue systems, and text mining. However, detecting sarcasm is difficult because it occurs infrequently and is difficult for self-annotated labels and does not consist of low-quality text snippets from Twitter[2]. With more than a million examples of sarcastic statements, each provided with author, topic, and contex information, the dataset also exceeds all previous sarcasm corpora by an order of magnitude. This dataset is possible due to the comment structure of the social media site Reddit[3] as well its frequently-used and standardized annotation for sarcasm.

Following a discussion of corpus construction and relevant statistics, in Section 4 we present results of a manual evaluation on a subsample of the data as well as a direct comparison with alternative sources. Then in Section 5 we examine simple methods of detecting sarcasm on both a balanced and unbalanced version of our dataset.

2 Related Work

Since our main contribution is a corpus and not a method for sarcasm detection, we point the reader to a recent survey by Joshi et al. (2016) that discusses many interesting efforts in this area. Note that many of the works the authors mention will be discussed by us in this section, with many papers using their own datasets; this illustrates the need for common baselines for evaluation.

Sarcasm datasets can largely be distinguished

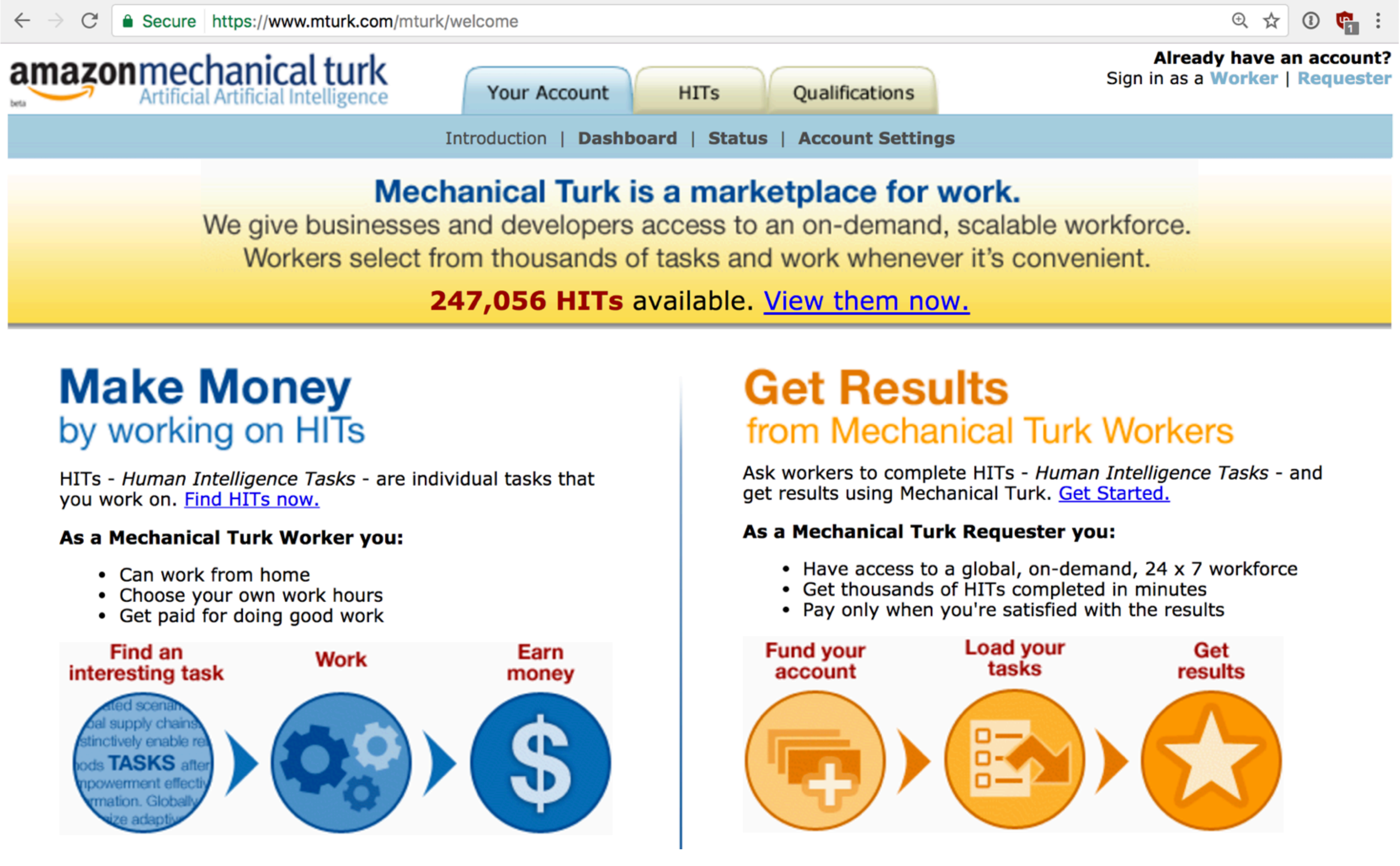
```
1       but CNN told me the leaks are all faked by Russia!
0       Is this some sort of mug shot mash up?
0       there isn't... the website says specifically there is no
way to change it once placed, you must cancel the order and place
a new one to make any modifications.
1       All men are handsome!
1       Yeah but he paid his dues so it's his turn.
0       ........what?
1       Maybe you should stop reading it
1       It's okay, it's just his opinion!
0       I know that's what she's doing now, I'm saying this isn't
the first time I've seen her.
0       honestly i'd have turned around and sold it to buy the K.
But at the end of the day, if you aren't really one to overclock,
then what difference does it make to you anyway?
```

Cohen's kappa

p_o: observed agreement rate

p_e: agreement rate by chance

$$\frac{p_o - p_e}{1 - p_e}$$


Secure
https://www.mturk.com/mturk/welcome
amazon mechanical turk
Artificial Artificial Intelligence
beta
Your Account
HITs
Qualifications
Introduction | Dashboard | Status | Account Settings

Already have an account?
Sign in as a Worker | Requester

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

247,056 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

- Millions of labeled images, collected through Mechanical Turk
- Revolutionized computer vision research (early '10s), facilitating convolutional neural networks (require large labeled data)
  - Data is key! (for certain/??all model/problem combinations)