Give a penny for their thoughts

Annie Zaenen *
Palo Alto Research Center

1 Introduction

RTE intends to capture the way a normal person makes textual inferences.[1] But how do we capture the intuitions of this "man in the street"? Using the intuitions of graduate students and their supervisors is a bit suspect and using traditional polling techniques to isolate a representative sample of the English speaking population would be prohibitively expensive apart from being methodological overkill. In this squib we describe and illustrate the use of a relative light-weight method that, without capture the "man in the street" outright, might be used to give the RTE data a broader basis. The proposal is to use a service such as Mechanical Turk to insure that the inferences are shared by several people or to obtain an idea of the variability of a particular inference. Below we describe a first experiment using Mechanical Turk with a subset of the RTE3 test data. The aim of this squib is not so much to present results than well to investigate what kind of information we can get out of the turker data.

2 Mechanical Turk

The Amazon Mechanical Turk (MTurk)® is part of the Amazon Web Services.[2] The service uses human intelligence in general to perform tasks which computers are unable to do. Here we will use it to perform a task that computers don't yet do well enough. "Requesters" ask people to perform tasks known as HITs (Human Intelligence Tasks), such as choosing paraphrases, labeling pictures, or transcribing voice-input. "Workers" (also called Providers, or less officially, "Turkers") can browse existing tasks and choose which ones they want to perform for the monetary payment set by the Requester. Payments per HIT at this point are typically very small, for 1 cent to a couple of dollars. To place HITs, the requesting programs use an open Application Programming Interface, or uses the MTurk Requester site.

Requesters can ask that Workers fulfill qualifications before engaging a task, and they can set up a test in order to verify the qualification. They can also accept or reject the result sent by the Worker, which reflects on the Worker's reputation. Currently, a Requester has to have a U.S. address, but Workers can be anywhere in the world. Payments for completing tasks can be redeemed on Amazon.com via a gift certificate or be later transferred to a Worker's U.S. bank account. Requesters, which are typically corporations, pay 10 percent of the price of successfully competed HITs to Amazon.

*Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94062, USA. E-mail: zaenen@parc.com.

[1]Thanks to Ron Kaplan for the original suggestion and to Raluca Budiu, Ron Kaplan and Lauri Karttunen for discussion.

[2]The name Mechanical Turk comes from "The Turk", a chess-playing automaton that toured Europe in the 18th century. It was later revealed that this 'machine' was not an automaton at all but was in fact a chess master hidden in a special compartment controlling its operations.

The results come back with, for each HIT, information about when the task was performed, how long it took and an ID for each of the Workers that responded to it.

3 A MTurk experiment with RTE data

We selected the first RTE pair and each subsequent 10th one of the latest RTE test suite, in all 80 of the 800 pairs[3] and submitted them in the required format to the Mechanical Turk service, offering a payment of 2 cents per HIT and asking for 50 answers for each RTE pair. We did not set up any qualification tests. The task was introduced by the following description, which we adapted from a task description used by the Stanford Aquaint group:

> The task is to determine whether the second text is true, given that the first text is true.
>
> You have to imagine that you do not know anything about the situation described in the first text except what the text itself tells you. You then answer "yes" if the information in the first text allows you to conclude that the second text is true, you answer "no" otherwise. The crucial information in the second text must be explicitly supported by the first text. Assume that a person asked you not only for a "yes" or "no" but also for evidence supporting your answer. If pointing him or her to the first text would satisfy the person, then the answer should be "yes", otherwise, it should be "no".
>
> Of course there are certain things you know about the world in general, information about time, space, and some geography, etc. that you can assume. For example, that a CEO is a person, that animals die, etc.
>
> When something is reported and you think nearly everyone would assume that it was true based on the report, the answer is "yes". For example, from "the New York Times reported that 5 people died in a subway fire in London yesterday", it seems highly likely that 5 people died in a subway fire in London yesterday. On the other hand if the source seems very untrustworthy, the answer should be "no".
>
> Examples:
>
> Text (1) The anti-terrorist court found two men guilty of murdering Shapour Bakhtiar and his secretary Sorush Katibeh, who were found with their throats cut in August 1991. Text (2) Shapour Bakhtiar died in 1991. Answer: yes
>
> Text (1) Many experts think that there is likely to be another terrorist attack on American soil within the next five years. Text (2) There will be another terrorist attack on American soil within the next five years. Answer: no

MTurk presents the HITs in random order. The 50 times 80 answers were obtained in a bit over 7 hours but it is not the case that every Worker gave a response for every RTE pair. Twelve Workers did all eighty HITs, others just a few or even just one.[4] 140 Workers contributed the 4000 HITS. Sixty-six of them contributed twenty or more answers each, accounting for 3422 HITS among them. Most individual answers took between a couple of seconds and around one minute and a half.

The main question is of course whether there are ways to insure that the Workers took the task seriously. The only external information about this in our experiment is the timing information. I only analyzed this information for the twelve participants that had done all eighty pairs. It is, of course, possible to do an analysis for all participants and for each question but for this

[3] Thanks to the RTE team for allowing us to use the data.

[4] The consequences of this aspect of the procedure will be discussed later.

pilot project I judged this unnecessary. Inspection of the complete runs reveals that they took between four minutes and two hours thirty five to complete. This immediately raises a couple of questions. The first is, do the slowest Workers really spend all their time on the task? The answer can be found by inspecting the timing data a bit more carefully for interruptions and indeed both of the Workers that took more than an hour had substantial periods where they were not checking out or entering any HITs. I subtracted the most important of these periods but I did not try to subtract all the short periods between the commitment of one HIT and the checkout of the next one. Again, in a more careful study this information can be calculated. A more troublesome question is the very low times that certain Workers had. Can the task really be done in four or nine minutes? I asked a well-organized and well-coordinated person in our lab to go through the eighty pairs as quickly as possible while making sure that she read and understood all the sentences. This took eighteen minutes. This leads to the suspicion that anybody who did the work in less time was not really performing the task. Another measure is to look at the answers and see whether the answers of these quick Workers (and others) were better than random and whether they differed substantially from those of other Workers. We discuss this in the next section.

4 Worker reliability

As said above some Workers did the HITs in a time span that realistically is impossible if the task is performed as required. To judge the reliability of these Workers we will assume that the RTE judgments are correct: although it is our aim to see whether the "man in the street" has the same judgments as the RTE constructors, we can assume that, if both are doing the same task, their judgments should not be completely divergent and that, if the results for most of the Turkers are close to the RTE judgments, the deviations are due to the fact that the deviant Workers were not doing the same task. The results for the twelve Workers who did the whole run of eighty questions are given in table 1.

WorkerID	Times (minutes)	Deviance from RTE
1	39	7
2	37	10
3	41	41
4	8:30	24
5	14	10
6	21	4
7	50	9
8	55	6
9	24	9
10	73	9
11	4	30
12	73	7

Table 1: Worker times and deviance score

One can see that the Workers with the two shortest times also had a high number of answers deviant from those given by the RTE standard. But not the highest! In fact, if one does a simple binomial test, assuming a random probability for each answer, only Worker 3 has given reliably random answers. Even for the person that did the work in four minutes 62,5 percent of

the answers were the same as the RTE annotation and the one who did it in eight and a half minutes got 70 percent of them. But if we look at the deviance scores of those that did the task within a reasonable amount of time, we see that their mean deviance of the RTE scores is 7.9 (8 if we do not count the one who did it in fourteen minutes), whereas the mean deviance of the ones that did it in an implausible amount of time is 27 (21.3 if we put the 14-minute Worker in this group). This difference is enough to conclude that the two groups were doing something different. To get valid results we should do the same type of calculations for those Workers that did not do complete runs and discard those for which we can't do this calculations reliably. We only computed the average agreement for the 54 Workers who did not do complete runs but did twenty or more HITS. Their average agreement with the RTE results is 0.87, with a spread from 1 to 0.66. Only one out of the fifty four has a binomial indicating that the answers might be due to chance (at 0.05). We conclude from these calculations that, while the methodology can be improved to insure that all the Workers do the right task, the results indicate that overall they took the assignment as intended.

5 Data and Results

The eighty pairs selected mirror the proportions of the original test suite closely but not completely: forty-one pairs were labeled 'no' and thirty-nine 'yes' by the RTE labelers in the reduced set whereas in the full suite the 'yes' and 'no' pairs are balanced. This would result in 2050 'no' answers and 1950 'yes' answers if the RTE results were completely repeated. There were twenty IE pairs, twenty IR pairs, twenty QA and twenty SUM pairs and ten long and seventy short pairs (These proportions are the same as those of the full suite). Overall the Turkers produced 1949 'no' answers and 2051 'yes'. Of the 2050 expected 'no' answers, 1714 were 'no' and 336 were 'yes'; of the 1950 expected 'yes' answers, 235 were 'no' and '1715' were yes. So for the 'no' answers they agreed in 83.6 percent of the cases with the RTE labelers and for the 'yes' in 87.9 percent. The difference between the 'yes' and the 'no' answers is significant (chi square 5.1). At first blush it seems to be due to the fact the the 'yes' answer was always in first position, a design error that would be easy to correct in a follow-up experiment, but a closer inspection of the various subcategories shows that the difference was only significant for the IE data and not for the three other categories, suggesting that the reason is not an overall design feature. As far as the correspondences of the answers to those of the RTE labelers goes, for IE, the 'no' and the 'yes' are the same at 77.8 and 89.8 percent respectively, for SUM 80.6 and 87.6, for IR 88.5 and 83 and for QA 86,9 and 90.2 percent. Interestingly, the Turkers agreed more on the long passages than on the short ones: 84.5 percent on the long'no' and 93.0 on the long 'yes' against 83.5 percent on the short 'no' and 87.0 percent on the short 'yes'.

Of course, more interesting than the overall results are the results per Text-Hypothesis pair. Here are the pairs in the order of agreement with RTE.

RTE pair deviance scores, not statistically significant

RTEpair	RTEAgreement
521	50
761	49
651	49
621	49
461	49
151	49
141	49
121	49
111	49
771	48
751	48
631	48
581	48
511	48
501	48
451	48
41	48
371	48
291	48
171	48
791	47
721	47
61	47
591	47
551	47
51	47
421	47
411	47
401	47
361	47
321	47
311	47
251	47
191	47
131	47
1	47

RTE pair with statistically significantly deviant scores but better than random agreement

RTEpair	RTEAgreement
611	46
531	46
491	46
441	46
381	46
351	46
341	46
331	46
31	45
211	45
731	44
701	44
431	44
391	44
241	44
91	43
661	43
471	43
281	43
201	43
101	43
21	42
181	42
601	41
301	41
3261	41
221	41
571	40
481	40
741	39
691	37
671	37
641	37
561	35
781	34
711	33

RTE pair deviance scores, random or negatively correlated

RTEpair	RTEAgreement
81	30
271	29
161	28
11	26
231	24
681	21
541	21
71	12

We see that 36 of the 80 pairs get a score that is not statistically significantly different from the RTE score (by a paried t-test); another 36 are significantly different but have results that go in the same direction and 8 are either random or negatiely correlated. We discuss first these eight cases. After that we take a quick look at the 36 most agreeing cases.

6 Pairs that show important divergences

In four cases the majority of the turker results went against the RTE judgment. In only one case (the first one), however, the disagreement is better than chance.

The most astonishing result is the one where the divergence between the RTE coding and the turker results is the biggest: for pair 71, the Turkers judgment was an overwhelming 'yes' (38), whereas the RTE coding was 'no'. One can only surmise that *claim* was taken in the wrong sense., or maybe a legal claim is perceived as justifying the use of *to belong to*. twelve complete runs: 3 no out of 9

> Text 71: Scott Island is part of the Ross Dependency, claimed by New Zealand.
> Hypothesis 71: Scott Island belongs to New Zealand.

Another case where the HIT answer diverge substantially from the RTE answer is set 541.

> T541: Ames then met and married Maria del Rosario Casas Ames a low level Colombian diplomat recruited to be a paid agent for the CIA and also according to some who knew the couple a strong minded woman of expensive tastes.
> H541: Aldrich Hazen Ames's wife was called Maria.

The RTE answer is no, but the turker result is: 21 'no' vs 29 'yes' 1 no out of 9 The RTE rational is that there shouldn't be an inference if there is more information in the hypothesis than in the text. In this case this is the full name of *Ames*. But there is an example that is in that respect very similar, 481.

> T481: The famous Colombian novelist known world-wide for his masterfully weaving of the magic realism genre, was born in 1928 in the small town of Aracataca, Colombia. Garcia Marquez was raised
> H481: Gabriel Garcia Marquez is a novelist.

2 no out of 9 Here the RTE answer is 'yes'. The majority of the Turkers agree: ten 'no' vs 'forty 'yes'. It is interesting that the 'yes' majority for *Garcia Marquez* is substantially bigger than for *Ames* suggesting that more people took *Garcia Marquez* as being part of general knowledge or it might be that the 541 passage is just more complicated or that a bigger part of the name was given in the text in the *Garcia Marquez* case than in the *Ames* case.

681 is another case where many of the Turkers do not take a verb, here *to seek* in the legal sense that the RTE encoders take for granted. For twenty-nine Turkers the answer here is 'no', whereas the RTE coding gives a 'yes' 3 no out of 9

> Text 681: They refused to appear in the World Court 10 years ago when Washington sought the release of American hostages in Tehran.
>
> Hypothesis 681: The court heard US appeals for the release of hostages held by Iran.

For pair 231 RTE annotators seem to have reasoned that a fall in tourism at the Canadian border is a fall in tourism in the US as a whole whereas the Turkers seem to think that the fall in one area could be compensated by a gain elsewhere so that the conclusion is unwarranted. In any case only twenty-four Turkers answered 'yes', the answer given by RTE. 5 no out of 9

> Text 231: The U.S. plan to require people traveling between the United States and Canada to have passports or similar identification is already hurting cross-border tourism, and it's not even in effect yet, tourism officials say.
>
> Hypothesis 231: Tourism falls in the U.S.

In the following four examples the majority of the Turkers agreed with the RTE judgment but in a proportion that does not exclude chance.

Case 11 seems again to be a problem of lexical semantics: is something that is 'released' by X also 'produced' by X? RTE thinks 'yes'. Twenty-six Turkers agree, twenty-four disagree. 4 no out of 9

> Text 11: In the Super Nintendo Entertainment System release of the game as Final Fantasy III , Biggs' name was Vicks.
>
> Hypothesis 11: Final Fantasy III is produced by the Super Nintendo Entertainment System.

Pair 161 hinges on a lexical subtlety and twnety-eight Turkers agree with RTE that *verbal attacks* are not *attacks* tout court whereas twenty-two disagree. 2 no out of 9

> Text 161: The demonstrators, convoked by the solidarity with Latin America committee, verbally attacked Salvadoran President Alfredo Cristiani.
>
> Hypothesis 161: President Alfredo Cristiani was attacked by demonstrators.

In pair 271 we seem again to have the problem of generalizing from a partial situation (art theft in churches) to a more general one (art theft in general). Here RTE finds the generalization unwarranted, twenty-nine Turkers agree whereas twenty-one do not. 4 no out of 9

> Text 271: Since the 1970s, crime has been increasing at churches around the country as religious art became more popular among collectors.
>
> Hypothesis 271: Art crime is increasing.

For pair 81 I can only surmise that the lack of a clear established referent for the first title and the general complexity threw off twenty of the Turkers who answered 'yes', whereas thirty picked up on the *additional* and performed the right coreference resolution to get to the RTE answer. 5 no out of 9

> Text 81: The title was again created for John Holles. When he died in 1711 the title became extinct but his estates passed to his nephew Thomas Pelham, who three years later upon coming of age received the title in its third creation. In 1757 he received the additional title of "Newcastle-under-Lyne".
>
> Hypothesis 81: John Holles received the title of "Newcastle-under-Lyne".

7 Pairs where the RTE annotators and the Turkers agree

For those 36 pairs there were more 'yes' answers (22) than 'no' ones (14). IE was represented with 11 cases, IR with 6, SUM with 8 and QA with 11. There were six long pairs and 30 short ones. I have not calculated whether these numbers are statistically significant.

521 First(521): Hundreds of thousands of demonstrators filled the streets of Taipei Saturday even
Second(521): Chen Shui-bian is the President of Taiwan. yes title/apposition

761 First(761): The agreement came after about 10 hours of negotiations between Mr Fradkov and hi
Second(761): Mr Fradkov and Sergei Sidorsky found an agreement after 10 hours of negotiations.
yes

651 First(651): A federal judge sentenced an apparently stunned Michael Milken to 10 years in pri
Second(651): Milken was given a 10-year sentence. yes deverbal

621 First(621): US STEEL, the largest integrated steel manufacturer in the US, is considering bui
Second(621): US Steel's mini-mill will be located in Ohio. no

461 First(461): Mike Jittlov, a master of special effects who's strutted his stuff in several sho
Second(461): The movie "Speed" was directed by Jan De Bont. no

151 First(151): Dean Lynn Hart was arrested Tuesday at his Storey County home after the attack at
Second(151): Dean Lynn Hart resides in Storey County. yes home--> reside

141 First(141): "Dozens of anti-terrorist police a swept into a neighborhood of Britain's second-
Second(141): Omar is of British origin. no

121 First(121): A band of ranchers under the leadership of Pedro Quintanar, upon hearing that Fat
Second(121): Pedro Quintanar was the leader of a band of ranchers. yes

111 First(111): Leloir was promptly given the Premio de la Sociedad Cientifica Argentina, one of
Second(111): Leloir was born in Argentina. no

771 First(771): Wal-Mart, the world's largest retailer, has reversed its opposition to the mornin
Second(771): Wal-Mart is largest retailer of the world. yes appositive

751 First(751): Bush and Clinton, who ran against each other in 1992, have worked together since
Second(751): Both Bush and Clinton helped raise funds for the recovery from Hurricane Katrina.
yes coreference

631 First(631): Today Argentina gets out the red carpet for the UK Duke of York, the first offici
Second(631): The Duke of York became the first official royal visitor to Argentina since 1982.
yes

581 First(581): With a diameter of 5,230 kilometers, Ganymede is the largest satellite in the sol
Second(581): The terrestrial moon has a diameter of 3,746 kilometers. no

511 First(511): Red Planet Consulting, Inc. is a full-scale project implementation firm that prov
Second(511): Mars is called "the red planet". no

```
501 First(501): Victor Emmanuel III (1869-1947) was king of Italy from 1900 to 1946. His cooperat
Second(501): Victor Emmanuel III was king of Italy from 1900 to 1946. litteral

451 First(451): Kourkoulos was married to the daughter of the Greek tycoon Yiannis Latsis, Marian
Second(451): Nikos Kourkoulos was a Greek actor. no

41 First(41): Robinson left his native Ireland in 1861 to take up a job at the Botanical Gardens
Second(41): Robinson was born in Ireland. yes

371 First(371): Heart disease and stroke are cardiovascular (heart and blood vessel) diseases cau
Second(371): Smoking causes diseases. yes

291 First(291): The United States is engaged in extensive international efforts on climate, both
Second(291): United states is engaged in illegal technology transfer. no

171 First(171): India's yearly pilgrimage to the Ganges river, worshiped by Hindus as the goddess
Second(171): Around 70 million people participate in pilgrimage to the Ganges river.
yes
```

8 Conclusion

The methodology of this experiment can be improved upon and Turker data do not represent the "man in the street" directly but the Turkers overall seem to understand the RTE task and they take it seriously. Thus they add information about how much variation there is in some judgments of RTE data. The RTE annotation proposes a gold standard achieved by making annotators agree on one judgment per pair. The turker data, however, show that, depending on the pair, there can be either wide disagreements or close agreements. In 10 percent of the cases, this led to a divergence where the agreement between turker judgments and RTE annotation was either negative or no better than chance. The percentage might have been a bit lower if the experiment had been better controlled but if it is representative on the RTE data as a whole, it is sufficiently high to have an impact on how the performance of various systems is judged. RTE might consider collecting similar data for coming challenges and treating them in either of two ways. It could either eliminate the most contentious examples from the development and test sets or it could weight the answers to them in a way that resembles the judgments of the broader community. Mixing both approaches is of course also a possibility.

To my mind, the most troublesome examples for either approach are the one where the answer depends on the selection of a particular lexical meaning. The examples 71 and 681 above. Here a subpart of the population selects a specialized meaning whereas another part uses a more general one, either because it is not aware of the specialized one or because the context is not rich enough to make them select it. Maybe some experiments could be done to see which of these two is the case. Another alternative is for RTE to decide that the segment of the population they are interested in, is the one that zooms in on the specialized senses. In that case some elimination questions can be added to the turker task to insure that only the desired subset of Workers is selected.