

Opinion Mining of Spanish Customer Comments with Non-Expert Annotations on Mechanical Turk

Rafael Banchs, Francesc Benavent, Joan Codina,
Jens Grivolla, Bart Mellebeek and Marta Ruiz

Barcelona Media Centre d'Innovacio



Barcelona, March 18, 2010

Outline

What is Mechanical Turk?

Task Outline and Goals

HIT Design

Annotation Results

Experimental Results

Conclusions

Outline

What is Mechanical Turk?

Task Outline and Goals

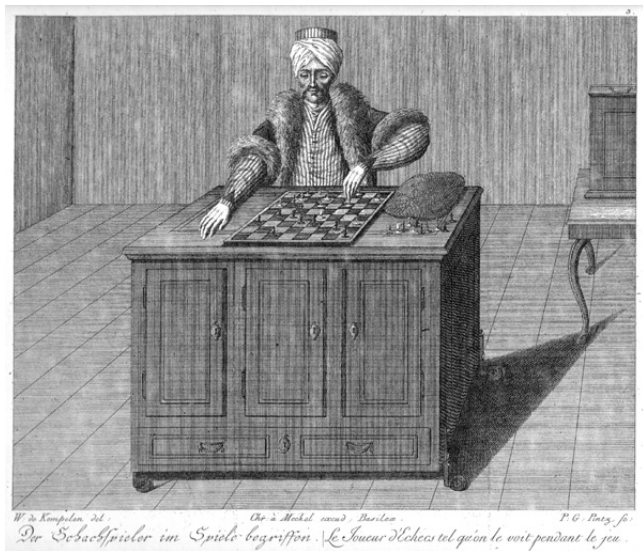
HIT Design

Annotation Results

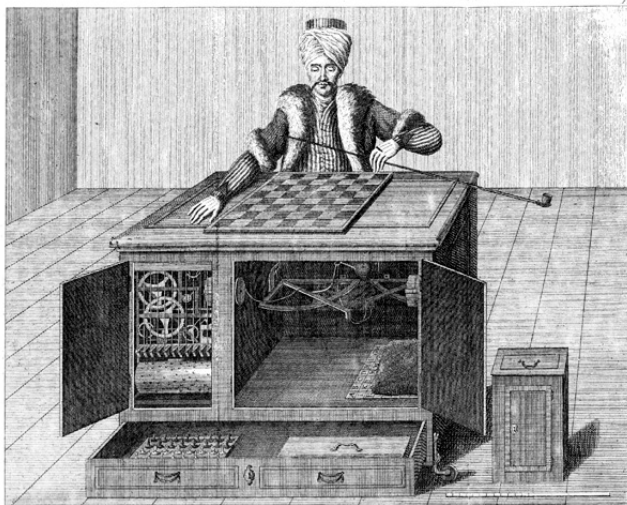
Experimental Results

Conclusions

Who was *The Turk*?



Who was *The Turk*?



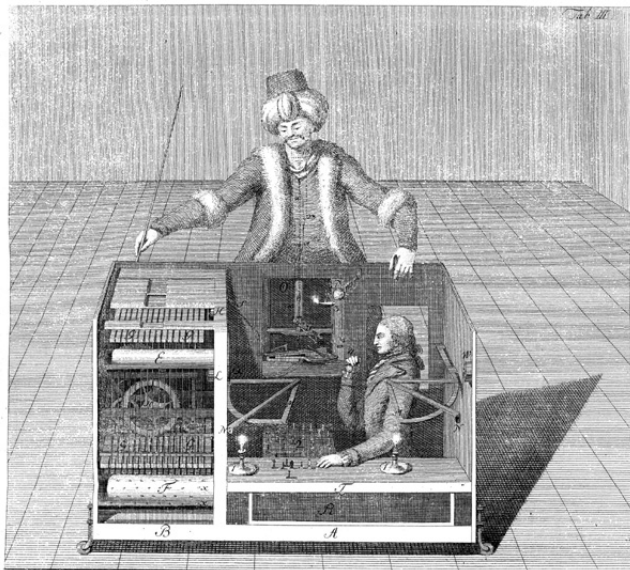
W. de Kempelen del.

Chis. a Michael oxend. Basileæ.

P. G. Ratz, fecit.

Der Schach-Spieler, wie er vor dem Spiele zu sehn wird, wenn er die Loure d'Chess, tel qu'on le montre avant le jeu, pardevant.

Who was *The Turk*?



Amazon Mechanical Turk (AMT) = *artificial* Artificial Intelligence

- ▶ Human annotations are a bottleneck for data-driven AI Systems.
- ▶ Amazon Mechanical Turk (AMT) is a marketplace for work.
- ▶ HIT = Human Intelligence Task.
- ▶ AMT workers can choose between tens of thousands of different HITs at whichever time they want.
- ▶ AMT requesters have access to a global workforce.
- ▶ Crowdsourcing → other platforms such as *Crowdfunder*

AMT worker annotations are cheap and fast. Are they also good?

- ▶ A number of recent papers argue they are: Kittur et al. (2008), Snow et al. (2008), Mason and Watts (2009), ...
- ▶ Highly dependent on type of task.
- ▶ General conclusion: the combination of several AMT worker annotations is better and cheaper than the annotations of a single expert.

Outline

What is Mechanical Turk?

Task Outline and Goals

HIT Design

Annotation Results

Experimental Results

Conclusions

Task Outline

- ▶ NAACL 2010 Workshop: Creating Speech and Language Data With Amazon's Mechanical Turk.
- ▶ Shared Task: What can you do with \$100 and Amazon Mechanical Turk?
- ▶ Participants share data & HIT templates.

We compare different HIT design strategies by evaluating the usefulness of resulting Mechanical Turk (AMT) annotations to train an Opinion Mining System on Spanish consumer data.

Research questions:

1. Annotation quality: AMT workers vs. experts.
2. Annotation applicability: training of batch of classifiers.
3. Language barriers: is there a large enough worker pool for Spanish?
4. Return on Investment: AMT workers vs. experts.

Outline

What is Mechanical Turk?

Task Outline and Goals

HIT Design

Annotation Results

Experimental Results

Conclusions

HIT = Human Intelligence Task


Color Naming Experiment View a HIT in this group			
Requester: Dolores Laba	HIT Expiration Date: Mar 24, 2010 (6 days 23 hours)	Reward: \$0.05	
	Time Allotted: 60 minutes	HITs Available: 1067	
Write a 200-300 word original article! Take Qualification test (Why?) View a HIT in this group			
Requester: ContentGalore	HIT Expiration Date: Mar 22, 2010 (5 days 7 hours)	Reward: \$1.70	
	Time Allotted: 1 hour 30 minutes	HITs Available: 20	
What Festival information has this festival? Not Qualified to work on this HIT (Why?) View a HIT in this group			
Requester: Stefan Hoevenag	HIT Expiration Date: Mar 19, 2010 (1 day 23 hours)	Reward: \$0.08	
	Time Allotted: 3 minutes	HITs Available: 10	
Edo a Transcript (The Role of Facilities in Green IT Strategies) (avg reward+bns: \$0.46) Request Qualification (Why?) View a HIT in this group			
Requester: CastingsWords	HIT Expiration Date: Mar 17, 2010 (3 hours 50 minutes)	Reward: \$0.23	
	Time Allotted: 12 hours	HITs Available: 1	
Fun 10-second quiz about favorite celebrities! Only 4 questions! Not Qualified to work on this HIT (Why?) View a HIT in this group			
Requester: VineTech	HIT Expiration Date: Mar 24, 2010 (6 days 23 hours)	Reward: \$0.10	
	Time Allotted: 60 seconds	HITs Available: 2530	
Corregir traducción automática de frases célebres Take Qualification test (Why?) View a HIT in this group			
Requester: Edo Segal	HIT Expiration Date: Mar 24, 2010 (6 days 23 hours)	Reward: \$0.01	
	Time Allotted: 24 hours	HITs Available: 1136	

a)
"¡No te lo pienses más, cómpratelo!"

b)

 [+] Es una opinión claramente POSITIVA

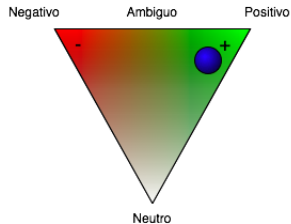
 [-] Es una opinión claramente NEGATIVA

 [?] No es ninguna de las anteriores, OTRO

c)



d)



Competence Test

- ▶ Uploaded test batch of 30 HITs.
- ▶ Performed in matter of minutes but ...
- ▶ most answers were random clicks.

Outline

What is Mechanical Turk?

Task Outline and Goals

HIT Design

Annotation Results

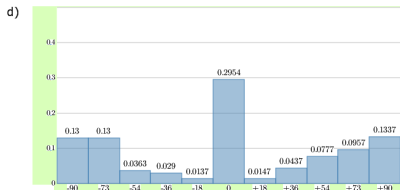
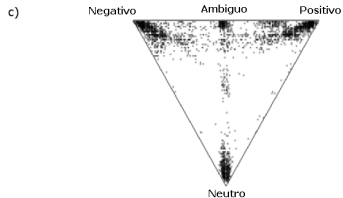
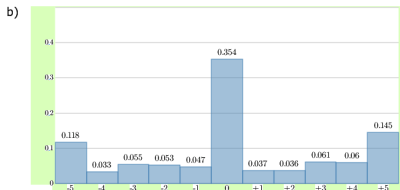
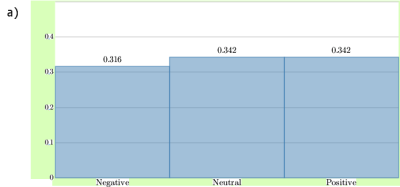
Experimental Results

Conclusions

HIT Statistics

Overall			HIT1		HIT2		HIT3	
ID	C	%	#	sec.	#	sec.	#	sec.
1	mx	29.9	794	11.0	967	8.6	930	11.6
2	us	27.6	980	8.3	507	7.8	994	7.4
3	nl	11.0	85	8.3	573	10.9	333	11.4
4	us	9.5	853	16.8	-	-	-	-
5	es	9.4	-	-	579	9.1	265	8.0
6	ec	4.1	151	9.4	14	16.7	200	13.0
7	us	3.6	3	15.7	139	8.5	133	11.6
8	us	2.2	77	8.2	106	7.3	11	10.5
9	us	0.6	-	-	-	-	50	11.2
10	us	0.5	43	5.3	1	5	-	-
11	us	0.4	-	-	38	25.2	-	-
12	us	0.4	-	-	10	9.5	27	10.8
13	es	0.4	-	-	-	-	35	15.1
14	es	0.3	-	-	30	13.5	-	-
15	us	0.3	8	24.7	18	21.5	-	-
16	us	0.2	-	-	-	-	22	8.9
17	us	0.2	-	-	17	16.5	-	-
18	?	0.1	6	20	-	-	-	-
19	us	0.1	-	-	1	33	-	-

Annotation Distributions



Annotation Quality

	κ_1	κ_2
Inter-batch	0.598	0.598
Batch_1 vs. Expert	0.628	0.628
Batch_2 vs. Expert	0.649	0.649
Batch_3 vs. Expert	0.626	0.626
Majority vs. Expert	0.716	0.716
Experts ¹	0.725	0.729

¹Results on a different 500-sentence random sample of the same corpus.

Annotation Costs

- ▶ A total amount of 9000 assignments were uploaded on AMT.
- ▶ Reward of .02\$ per assignment \rightarrow total sum = 225\$ (180\$ + 45\$ Amazon fees).
- ▶ In-house expert: 1000 assignments in three hours at 70% per hour.
- ▶ We saved $210 - 75 = 135$ \$, which constitutes almost 65% of the cost of an expert annotator.
- ▶ + a lower reward (.01\$) would most probably have worked equally well.

Outline

What is Mechanical Turk?

Task Outline and Goals

HIT Design

Annotation Results

Experimental Results

Conclusions

Two Experiments

- ▶ Experiment 1: practical utility. Train a single polarity classifier on different data sets. Compare system with noisy available metadata and with AMT generated annotations of HIT1.
- ▶ Experiment 2: theoretical utility. Train several polarity classifiers using the same data set, annotated by experts and AMT workers.

Description of Data Sets

All sentences were extracted from a corpus of user opinions on cars from the automotive section of www.ciao.es.

	Baseline	Annotated	Evaluation
Positive	1882	341	200
Negative	1876	323	137
Neutral	1812	336	161
Totals	5570	1000	500

Experiment 1

classifier	baseline	annotated
positive/not_positive	59.63 (3.04)	69.53 (1.70)
negative/not_negative	60.09 (2.90)	63.73 (1.60)
neutral/not_neutral	51.27 (2.49)	62.57 (2.08)

Experiment 2

System	Experts	Batch1	Batch2	Batch3	Majority	All
Winnow	44.2	43.6	40.4	47.6	46.2	50.6
SVM	57.6	53.0	55.4	54.0	57.2	52.8
C45	42.2	33.6	42.0	41.2	41.6	45.0
Maxent	59.2	55.8	57.6	54.0	57.6	58.6

Outline

What is Mechanical Turk?

Task Outline and Goals

HIT Design

Annotation Results

Experimental Results

Conclusions

We compare different HIT design strategies by evaluating the usefulness of resulting Mechanical Turk (AMT) annotations to train an Opinion Mining System on Spanish consumer data.

Research questions:

1. Annotation quality: AMT workers vs. experts.
2. Annotation applicability: training of batch of classifiers.
3. Language barriers: is there a large enough worker pool for Spanish?
4. Return on Investment: AMT workers vs. experts.

Results

We compare different HIT design strategies by evaluating the usefulness of resulting Mechanical Turk (AMT) annotations to train an Opinion Mining System on Spanish consumer data.

Research questions:

1. Annotation quality: High inter-annotator agreement scores.
2. Annotation applicability: AMT worker annotations outperform initial noisy data and on par with expert annotations.
3. Language barriers: after introducing a competence test, getting results for Spanish was easy and rapid.
4. Return on Investment: we saved 65% of an expert annotator at a relatively high reward.

Questions?



 Barcelona
Media