Analysis of the MTurk Annotation Task Batch 1 (2016-06-20) V2

Intoduction

In this batch, we launched 50 HITs in MTurk to label Technology videos. In each HIT, there are 2 unlablled videos, accompanied with one manual-labelled video as the “qualification” question. Only MTurk workers with Masters Qualification were allowed to participate.

Two questions were asked for each video:

1) For the given description of the video, is there anything related to the content of the video, or simply is a marketing strategy.
2) For the given video, from a predefined list of tags within the domain of Technology, choose up to 3 most relevant ones to describe the content of this video.

Here, we present a brief analysis of the quality of the collected results.

1. The list of invalid manual-labelled video IDs

The supplied label(s) in MongoDB does not belong to the designed sub-categories. **Please revise these labels. Otherwise, HITs using one of these samples as the "qualification" question have to be excluded from the remaining analysis.**

a) 5758f41751ac842181db73b3
b) 5758f41751ac842181db73b6
c) 5758f41751ac842181db73b9
d) 5758f41851ac842181db73bf

Now these invalid videos are fixed. We can include these HITs include further analysis.

2. Number of valid HITs

We now have **50** valid HITs for further analysis. Note that same video might be used in multiple HITs as the "qualification" question.

For each of these **50** valid HITs, each of which received 3 assignments from the Master MTurk workers. In order to measure the quality of these workers’ work, we break down the HITs by the number of “qualified” assignments received from the workers. We define “qualified” assignments as those which agree with the manual label. We therefore assume those workers’s work on the remaining two questions is also of high quality.

# of HITs	# of qualified assignments	% of total # of HITs
0	4	**8.0%**
1	9	**18.0%**
2	**14**	**28.0%**
3	**23**	**46.0%**

In total, over **92%** of the HITs receive at least 1 qualified assignment. We may need to revisit the following three manual-labelled videos, which are used in four HITs, because none of the 3 MTurk workers agreed with the manual label:

a) 5758daaa51ac842068292c7c
b) 5758f93e51ac8421ccf54c19
c) 5758dadb51ac842068292e7d

3. *Inter-annotator agreement for the DescriptionType question*

Now we look at the **37** HITs which received at least two qualified assignment. We use these assignments to compute inter-annotator agreement for the DescriptionType question.

In this question, the worker is asked whether the video description contains any information about the content of the video, or is a pure marketing description.

Among the **74** unlabelled videos in these **37** HITs, **69** of them (93%) of them receive unanimous voting for this question.

4. *Inter-annotator agreement for the SubTags question*

Similarly, among the **74** unlabelled videos in these **37** HITs, **45** of them (**61%**) of them receive a common tag from at least two workers. Note that for the two workers who chose at least one common tag for the same video, they may have chosen different other tags. However, even in this case, we still consider the two workers as reaching an agreement for this question.

Consider the complexity of this multi-label task, the agreement of **61%** is decent, and the assignments could be considered as reasonably reliable.

However, among these **45** videos, 5 of them achieve agreements on a worker-input tag, i.e., not from the pre-defined list. These worker-input tags include:

a) Music
b) Politics
c) Gaming
d) Space

The first two sub-tags do not seem relevant to the Technology category, and we will exclude them from the final training set as well.

5. *Final Distribution of the Training Data*

a) Training content vs. non-content description classification

Out of the 100 unlabelled samples in this batch, we collected **69** reliable samples. The distribution of the two classes: content vs. non-content is shown below. The majority of the videos contain content-related description, which would make the next classification relatively easy, because there would be some information contained in the textual content of the video, rather than the video itself.

Content?	# of samples
Y	63 (92%)
N	6 (8%)

b) Training sub-tag classification

Out of the 100 unlabelled samples in this batch, we collected **44** reliable samples. The distribution of the two classes: content vs. non-content is shown below:

Tag	# of videos
Smartphones	**14**
Ecommerce	5
Artificial Intelligence	**5**
Internet of Things	**4**
Drone	3
Wearable Tech	3
NanoTech	3
Battery	2
Gaming	1
Manufacturing	1
Social Networks	1
Space	1
Virtual Reality and Augmented Reality	1

Together with manual-labelled samples in this batch, we have 83 samples. The distribution of the two classes: content vs. non-content is shown below:

Tag	# of videos
Smartphones	14
Wearable Tech	14
Driverless Cars	12
Artificial Intelligence	8
Internet of Things	8
Social Networks	6
Ecommerce	5
Drone	4
Battery	3
Virtual Reality and Augmented Reality	3
Manufacturing	2
NanoTech	2
Gaming	1
Space	1

Discussion

Based on the results of this first pilot study, we may want to:

1. Revise the predefined list of tags to include the two additional ones as pointed out by the workers:
   a) Gaming
   b) Space

2. **It seems the distribution of the sub-categories among the collected samples is very skewed. We may need to investigate if there is a more sensible way of scrape videos for further labeling.**

3. Launch a larger-scale MTurk task to collect more data.
   Based on the distribution of different sub-categories among all 83 (manual- and MTurk-labelled) samples, for the scarcest classes (excluding the two worker-input ones), we would need roughly another 1,000 samples to make sure we have at least 20 samples per class.