Analysis of the MTurk Annotation Task Batch 2 (2016-06-30) V2

Intoduction

In this batch, we launched 730 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. *Number of valid HITs*

We now have **730** valid HITs for further analysis. Note that same video might be used in multiple HITs as the "qualification" question.

For each of these **730** 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	182	25%
1	175	24%
2	307	42%
3	66	9%

In total, 75% of the HITs receive at least 1 qualified assignment. Since the coverage is a bit too low, and considering the fact that some labels are not mutually exclusive, we expanded the criteria of qualified assignments to be:

a) At least two assignments agree with the manual label.
b) At least two assignments agree with each other for the qualification question.

By applying the above criteria, we have 516 qualified HITs.

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

Now we look at the **516** qualified HITs gather by the criteria above. 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 **1032** unlabelled videos in these **516** HITs, **1004** of them (97%) of them receive unanimous voting for this question.

3. Inter-annotator agreement for the SubTags question

Similarly, among the **1032** unlabelled videos in these **516** HITs, **426** of them (**41.3%**) 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 **426** videos, 24 of them achieve agreements on a worker-input tag, i.e., not from the pre-defined list.

4. Final Distribution of the Training Data

a) Training content vs. non-content description classification

Out of the 1460 unlabelled samples in this batch, we collected **1004** reliable samples. The distribution of the two classes: content vs. non-content is shown below. All 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	1004 (100%)

b) Training sub-tag classification

Out of the 1460 unlabelled samples in this batch, we collected **426** reliable samples. The distribution of all sub-categories is shown below:

Tag	# of videos
Drone	135
Virtual Reality and Augmented Reality	79
Battery	54
NanoTech	47
Social Networks	24
Smartphones	16
Manufacturing	14

Wearable Tech	11
Ecommerce	6
Artificial Intelligence	4
Computer Science	4
Food	4
Internet of Things	4
Driverless Cars	3
Exercise	2
Health	2
3D Printing	1
Airlines	1
baseball	1
BioTech	1
Business	1
coffee	1
Doctor Who	1
education	1
fashion	1
Golf	1
Military	1
Movies	1
NBA	1
netflix	1
nfl	1
shooting	1
travel	1

The pre-defined tags are highlighted in yellow.

Combined with the results gather in the first batch, and putting all non-predefined tags into the tag “Others”, we have a total of 470 labelled videos with the following distribution:

Tag	# of videos
Drone	138
Virtual Reality and Augmented Reality	80
Battery	57
NanoTech	49
Smartphones	30
Others	26
Social Networks	25
Wearable Tech	16
Manufacturing	15
Ecommerce	11
Internet of Things	8

Artificial Intelligence	7
Computer Science	4
Driverless Cars	3
BioTech	1

Discussion

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

1. The distribution of the sub-categories among the collected samples is still very skewed. We may want to consolidate some rare tags into “Others” as well.