Analysis of the MTurk Annotation Task Batches 1-3

Introduction

In these three bathces, we launched 1868 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	709	39%
1	236	13%
2	388	21%
3	472	26%

In total, 61% 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 – therefore, workers tend to reach low agreement on those labels – we expanded the criteria of qualified assignments to be either:

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 **1744** 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 **3488** unlabelled videos in these **1744** HITs, **3111** of them (89.2%) of them receive unanimous voting for this question.

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

Similarly, among the **3488** unlabelled videos in these **1744** HITs, **1842** of them (**52.8%**) 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 **52.8%** is decent, and the assignments could be considered as reasonably reliable.

In addition, we also include another version, where a relaxed condition is used to find qualified sub-tag labeling: for a given video, if no two workers agree with each other on the labeling, but some workers selected a label from the pre-defined label list, i.e., not manual-input label, we also include those labels as qualified labelling. Using this relaxed condition, we collected **2068 (59.3%)** samples.

We'll present the distribution using the strict condition and the relaxed condition separately.

4. *Final Distribution of the Training Data*

a) Training content vs. non-content description classification

Out of the 3488 unlabelled samples in this batch, we collected **3111** 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	3099 (99.6%)
N	12 (0.4%)

b) Training sub-tag classification

- Strict condition

Out of the 3488 unlabelled samples in this batch, using the strict condition and putting all non-predefined tags into the tag "Others", we have a total of 1747 labelled videos with the following distribution:

Tag	# of videos
Social Networks	366
Ecommerce	160
Wearable Tech	157
BioTech	151
Smartphones	148
Drone	139
Artificial Intelligence	132
NanoTech	123
Driverless Cars	97
Virtual Reality and Augmented Reality	83
Internet of Things	63
Battery	57
Manufacturing	43
Others	26
Computer Science	4

- Relaxed condition

Out of the 3488 unlabelled samples in this batch, using the relaxed condition and putting all non-predefined tags into the tag "Others", we collected **1842** reliable samples with the following distribution:

Tag	# of videos
Social Networks	577
Manufacturing	433
Smartphones	386
Internet of Things	373
Wearable Tech	366
Artificial Intelligence	327
Ecommerce	286
Virtual Reality and Augmented Reality	286
Drone	250
BioTech	240
NanoTech	225
Driverless Cars	167
Computer Science	133
Battery	100
Big Data	44

Discussion

We now have a decent number of available training samples. We can try modeling using the samples obtained with the relaxed condition first and evaluate results.