Quality Control on Crowdsourcing Image Annotation using Direct Assessment

Navaneethan Rajasekaran
School of Computing, Dublin City University
Email: navaneethan.rajasekaran2@mail.dcu.ie

***Abstract*—Image annotation has been an ongoing research topic due to its prominent significance in image retrieval and computer vision. Crowdsourcing is widely used in industry and research for manual image annotation.Crowdsourced tasks are prone to errors. Direct Assessment(DA) is a methodology for crowdsourcing human assessments of translation quality. we developed a new assessment strategy based on DA to quality control on crowdsourcing image annotation.**

Keywords:

I. Introduction

Crowdsourcing is the practice of obtaining needed services from large group of geographically dispersed people via internet. It allows us to utilize the power of human computation to complete tasks which are difficult to solve by computers alone. Crowdsourcing has been used to annotate images that are needed to train computer vision models. Amazon Mechanical Turk is the widely used crowdsourcing platform among the researchers. It is designed for crowdsourcing tasks (often referred to as "HITs,") such as annotating images or completing a survey. Task requesters come to Mechanical Turk to post their tasks, stating upfront the amount of money that they are willing to pay to have their tasks completed. Requesters can also specify certain criteria that a crowd worker must meet to be eligible for their task, such as having an approval rate of more than a particular amount (say,97%) on previous tasks or being located in a particular country [1]. Crowd workers can then browse the set of tasks available and choose the tasks they would like to work on. After a crowd worker completes a task, the requester approves his/her work and payment is made.

Workshop on Statistical Machine Translation (WMT), uses crowdsourcing for some of their tasks. WMT is about working on shared tasks which have goals to push works on the integration of computer vision and language processing, multimodal language processing towards multilingual multimodal language processing etc. On their second conference on Machine Translation (WMT 17), Multimodal Machine Translation task consists in translating English sentences that describe an image into German and/or French, given the English sentence itself and the image that it describes (or features from this image, if participants chose to). For this task, the Flickr30K Entities dataset is extended in the following way: for each image, one of the English descriptions was selected and manually translated into German and French by human translators. For English-German, translations are produced by professional translators, who are given the source segment only (training set) or the source segment and image (validation and test sets). For English-French, translations are produced via crowdsourcing where translators had access to source segment, the image and an automatic translation created with a standard phrase-based system (Moses baseline system built using the WMT'15 constrained translation task data) as a suggestion to make translation easier. The organizers stated that "that this was not a post-editing task: although translators could copy and paste the suggested translation to edit, we found that they did not do so in the vast majority of cases" for English-French, translations which are produced via crowdsourcing.

Direct Assessment(DA) is a new methodology for human evaluation of MT quality. The method relies entirely on assessments sourced from the crowd. The approach is based on actively removing sources of bias, including mechanisms to accommodate assessors with consistent individual scoring strategies. By restructuring the task as an assessment of monolingual similarity of meaning, assessing individual translations, and separating fluency and adequacy, the task was made substantially less cognitively taxing, and allowed participation by much larger pools of workers [3]. DA was employed in human evaluation campaign of WMT17 to assess translation quality and to determine the final ranking of systems taking part in the competition.

Crowdsourcing platforms are well suited to generating data, but challenges arise since the data supplied by crowd workers can be prone to errors [1]. In this paper, we discuss the possibility of using DA for quality control on crowdsourcing Image Annotation task.

II. Related Work

A. Image Annotation Techniques

1) Manual Annotation: Manual annotation is the process of utilizing human computation to annotate images. Crowdsourcing platforms are widely used for this process. Workers are asked to enter some descriptive keywords when the images are loaded/registered/browsed. Manual annotation of image content is considered "best case" in terms of accuracy since keywords are selected based on the human determination of the semantic content of images. Some of the accessible computer vision datasets are gathered using manual annotation.

2) Automatic Annotation: Automatic image annotation is the process by which computer system assigns metadata in

the form of captioning or keywords to a digital image automatically.

Image segmentation algorithms are used to classify the images into a number of unevenly shaped blob regions [7]. It uses global features to work on these blobs for automated image annotation. This modeling framework is based on non-parametric density estimation, using the technique of kernel smoothing. Annotator has to choose the word for annotating image with a certain probability. This probability can be interpreted into probability density of image x and density of x conditional upon the assignment of annotation.

Image Annotation by coherent language model and active learning [8] uses the word to word correlation, as image features are inadequate in establishing the corresponding word annotation. To integrate the word-to-word correlation, it approximates the probability of annotating an image with a set of words. This approach uses the language model to provide words for image annotation. The model provides the probability of certain words to be used. The advantage of this approach is that it automatically determines the annotation length for a given image, which in turn enhances the precision of image retrieval.

Content-Based Image Annotation Refinement [9] approach improves the existing annotations of images. It refines the conditional probability so that more precise annotations will have higher probabilities. As an effect, the annotations with the highest probabilities is kept as the final annotations. For a query image, an existing image annotation method is used to obtain a set of applicant annotations. Then, the applicant annotations are re-ranked, and only the top ones with the high probabilities are considered as the ultimate annotations. While re-ranking, the fixed probability of word annotating the particular image is calculated.

B. Widely used Image dataset and the approach used to source them

1) MS COCO: Microsoft Common Object in Context (MS COCO) is the largest dataset containing 300k images with five-sentence per image and over 2.5m labelled object instances from 91 pre-defined categories. The creation of the dataset drew upon extensive crowdsource involvement via novel user interface for category detection, instance spotting and instance segmentation. The crowd work on the category labelling task was evaluated by dedicated experts.

COCO CN dataset is the recent extension of MS COCO data set with manually translated Chinese image descriptions. COCO CN argued that there is number of typos and misspelling in the MS COCO and the quality of crowdsourced data is unresolved. They came up with approach of assisting annotators in avoiding the bias. They developed a web-based image annotation system that allows users to annotate images remotely and independently. The interface had two content-based recommendation modules, one for sentences and other for tags where user can refer the recommended sentences and tags for the manual translation. The author justified that this approach has improved the quality of crowdsourcing work.

STAIR is the extension of MS COCO which has 820,310 Japanese captions for 164,062 images. The English captions were manually translated to Japanese by crowdsource with certain provided guidelines. The quality of the caption is evaluated by sampling inspection. To the best of our knowledge STAIR is the largest available Japanese image captioning dataset.

2) Flickr30K: Flickr 30k Dataset contains 31,014 images sourced from online photo-sharing websites (cite-young et al). Each image is paired with five English image descriptions, collected from crowdsourcing. The data set contains 145,000 training data, 5,070 development, and 5,000 test descriptions.

Flickr30k entities is the extension of Flickr 30k, a large-scale image description dataset that provides comprehensive ground-truth correspondence between regions in the images and phrases in the captions. Flickr 30K entities augment Flickr 30k by identifying which mentions among the captions of the same image refer to the same set of entities, resulting in 244,035 coreference chains, and which image regions depict the mentioned entities, resulting in 275,775 bounding boxes. The dataset proposes a benchmark task of phrase localization.

Multi30k dataset extends the Flickr30K dataset with I) German translations created by professional translators over a subset of the English descriptions, and ii) descriptions crowdsourced independently of the original English descriptions. The team has outlined how the data can be used for multilingual image description and multimodal machine translation. The annotators were not aware of the image associated with the sentence.

C. Existing metrics for Automatic Evaluation of Machine Translation

In the early 1990s, the U.S. government sponsored a competition among machine translation (MT) systems. One of the invaluable outcomes was a corpus of manually produced numerical judgments of MT quality, with respect to a set of reference translations (White et al., 1993). The relatively high cost of producing such judgments have inspired many researchers to seek reliable methods for estimating such measures automatically.

BLEU, NIST, F-measure, and METEOR are some of the metrics based on n-grams model developed for automatic evaluation of MT.

1) BLEU: BLEU (bilingual evaluation understudy) is a metric for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the agreement between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU. BLEU was one of the first metrics to claim a high correlation with human judgments of quality and remains one of the most popular automated and economic parameters.

Scores are computed for single translated segments—generally sentences—by matching them with a set of good quality reference translations. Those scores are

then averaged over the whole corpus to reach an estimate of the translation's overall quality. Intelligibility or grammatical exactness are not taken into account

2) NIST: Doddington (2002) proposed a version called NIST (National Institute of Standards and Technology). NIST precision measure is a metric used to evaluate the MT variants. The more of these N-grams that a translation shares with the reference translations, the better the translation is judged to be. NIST is intended as an enhanced version of BLEU. In this case, the arithmetic mean of n-grams is calculated. An important variation from the BLEU metric is the fact that NIST also relies on the frequency component (precision and recall). If BLEU calculates the n-gram precision by adding an equal weight for each exact match, NIST also calculates how informative each matching n-gram is. For instance, even if the bigram 'on the' coincides with the same phrase in the reference text, the translation still receives a lower score than the correct matching of the bigram 'size distribution,' because the latter phrase is less likely to occur.

3) F-Measure: The F-measure is a metric which measures the harmonic mean of precision and recall [4]. The metric is based on the search for the most suitable match between the candidate and reference translations (the ratio between the total number of matching words to the length of the translation and the reference text). Sometimes it is valuable to combine the precision and recall of the same averaged value.

4) METEOR: The Meteor automatic evaluation metric scores machine translation hypotheses by aligning them to one or more reference translations. Meteor is an improved version of F-Measure Alignments are based on exact, stem, synonym, and paraphrase match between words and phrases. Segment and system-level metric scores are calculated based on the alignments between hypothesis-reference pairs. The metric includes several free parameters that are tuned to emulate various human judgment tasks, including WMT ranking and NIST adequacy. The current version also includes a tuning configuration for use with MERT and MIRA. Meteor has extended support (paraphrase matching and tuned parameters) for the following languages: English, Czech, German, French, Spanish, and Arabic. Meteor is implemented in pure Java and requires no installation or dependencies to score MT output. On average, hypotheses are scored at a rate of 500 segments per second per CPU core. Meteor consistently demonstrates a high correlation with human judgments in independent evaluations such as EMNLP WMT 2011 and NIST Metrics MATR 2010.

Meteor X-ray uses XeTeX and Gnuplot to create visualizations of alignment matrices and score distributions from the output of Meteor. These visualizations allow easy comparison of MT systems or system configurations and facilitate in-depth performance analysis by examination of underlying Meteor alignments.The final output is in PDF form with intermediate TeX and Gnuplot files preserved for inclusion in reports or presentations.

The drawback of these automatic evaluation metrics is "What works on one corpus might not work on another" [5]. There is no best MT automatic evaluation metric because different tasks have different interpretations of what constitutes a good, fit for purpose translation. Although BLEU was one of the first metrics to achieve a high correlation with human judgements of quality and is still one of the most widely used metrics, it is not be suited for all the evaluation tasks.

III. Dataset

Extended Flickr30K Entities dataset which was part of test data used in WMT17 for Multimodal Machine Translation task is used. The dataset contains 1000 images.

IV. Direct Assessment

Direct Assessment is a direct human MT evaluation methodology. Direct Assessment was demonstrated to show that it was possible to measure MT systems reliably based on crowd-sourced judgments alone. Key features of this methodology are:

- It can be used to evaluate both adequacy and fluency;
- The ratings are captured via direct estimates on a 100-point Likert scale, enabling fine-grained statistical analysis (Graham et al. 2013);
- It incorporates mechanisms for quality control, based on internal consistency over pairings of original and 'degraded' translations (Graham et al. 2014);
- It is backwards-compatible with the style of system preference judgment used for WMT evaluations, and provides a mechanism for enabling longitudinal evaluation of MT systems;
- It is cheap;
- It requires only monolingual annotators conversant in the target language, thus allowing the use of a larger pool of lower-skilled annotators than is possible with standard manual evaluation approaches;

As part of the Direct Assessment baseline model, Amazon Mechanical Turk is used as a crowdsourcing platform. HIT contain hundred translation assessments each, one per screen. The worker is thus required to iterate through translations, rating them one at a time. Each hundred-translation HIT consists of:

- seventy MT system outputs;
- ten reference translations, corresponding to ten of the seventy system outputs;
- ten bad reference translations, corresponding to a different ten of the seventy system outputs ;
- ten repeat MT system outputs, drawn from the remaining fifty of the original seventy system outputs;

The order in which a worker access translation is controlled so that a minimum of forty assessments intervene between each member of a pair of quality-control items (bad reference versus MT system, or MT system versus MT system repeat, or reference translation versus MT system). The selection is further controlled so that each hundred-translation HIT contains approximately equal numbers of randomly-selected translations from each contributing MT system, so as to

Figure 1. Example of user interface from AMT for image annotation

Figure 2. Distribution of translation in a HIT

provide overall balance in the number of translations that are judged for each system. That is, no matter how many HITs each worker completes, they will return roughly the same number of assessments for each of the contributing systems. This helps avoid any potential skewing of results arising from particularly harsh or lenient assessors.

Approach to measuring assessor consistency is based on two core assumptions (Graham et al. 2017):

- (A) When a consistent judge is presented with a set of assessments for translations from two systems, one of which is known to produce better translations than the other, the score sample of the better system will be significantly greater than that of the inferior system.
- (B) When a consistent judge is presented with a set of repeat assessments, the score sample across the initial presentations will not be significantly different from the score sample across the second presentations.

Assumption A is validated based on the set of bad reference translations and the corresponding set of MT system translations and Assumption B based on the pairs of repeat judgments in each HIT. The bad reference translations that are inserted into each HIT are deliberately degraded relative to their matching system translations, on the assumption that a measurable drop in the assessor's rating should be observed.

The null hypothesis to be tested for each Amazon Mechanical Turk worker is that the score difference for MT system translations is not less than the score difference for bad reference pairs. To test statistical significance, the paired t-test is used. With lower p values indicating more reliable workers (that is, greater differentiation between repeat judgments and bad reference pairs). $P < 0.05$ as a threshold of reliability is used to access the workers.

V. Direct Assessment for crowdsourcing Image Captions

As part of this approach, Amazon Mechanical Turk is used as a crowdsourcing platform. Direct Assessment is technically modified such that each HIT contains thirty translation assessments shuffled randomly and seventy Images, one per screen. The worker is thus required to iterate through each hit (thirty randomized translation first and seventy images later). Each HIT consists of:

- seventy images,
- ten reference translations,
- ten bad reference translations corresponding to ten of the MT system outputs,
- ten MT system outputs

Figure 3. Distribution of translation and images in a HIT from the experiment

The null hypothesis to be tested for each Amazon Mechanical Turk worker is that

- (A) The score difference for system output is not less than the score difference of bad references. To test statistical significance, the paired t-test is used.

As part of this experiment, in addition to null hypothesis A, each Amazon Mechanical Turk worker is also tested for the null hypothesis that

- (B) The score difference for reference is not less than the score difference of bad references. To test statistical significance, the Mann-Whitney U test is used

The quantity of work involved per HIT was communicated to workers prior to their acceptance of a HIT. An additional specification was that only native speakers of the language complete HITs, in this case it was English.The payment was at the rate of US$0.99 per HIT. Due to the anonymous nature of crowdsourcing, it is of entirely possible for workers requiring the necessary skills to employ someone else to complete his/her test. Besides, qualification tests do not provide any assurance that skilled workers do not carefully complete experiments, and then aggressively optimize earnings at a later stage. In contrast, the quality-control mechanism we employ does not rely on one-off tests but applies quality checks across all of the HITs provided by each worker.

Since the quality-control mechanism, could be too high a bar for some genuine workers to meet, we do not use or recommend its use as the sole basis for accepting or rejecting HITs. Indeed, some workers may lack the necessary literacy skills to complete evaluations effectively or annotate images accurately. In the assessment (Graham et al. 2017), they have rejected only the insincere workers who they believe tricked the system into earning money. Deceitful workers are identified by comparison of the worker's mean score for reference translations, genuine system outputs, and bad reference translations. For example, the random-clickers had these mean scores extremely close. Another approach which helped to identify insincere workers was adequacy assessment. When a reference translation appears like the item to be assessed, it was identical to the reference translation displayed on the screen. Although such items appeared to be too obvious to be useful in assessments, they, acted as a further modest hurdle for workers to meet and caught many workers.

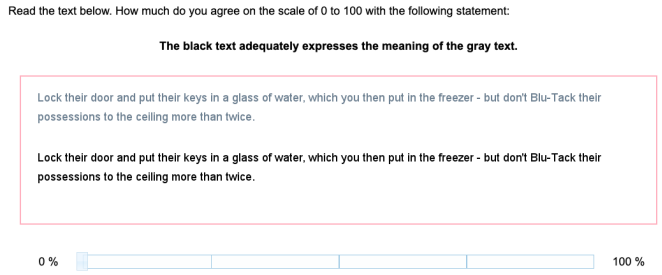


Figure 4. Example of a REF displayed to AMT worker to rate

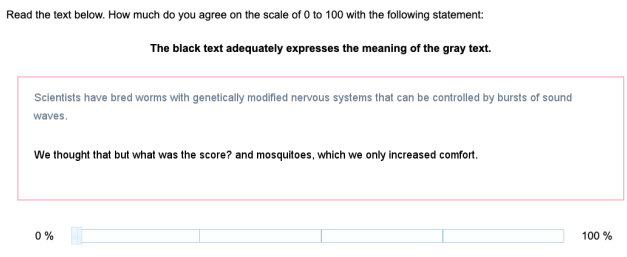


Figure 5. Example of a BAD˙REF corresponding to system output to AMT worker to rate

1) Experiment- Stage 1: The experiment is carried out to evaluate the initial setup. One HIT (30 Translation Assessment and 70 images) is posted on AMT ,and 15 unique workers are asked to annotate similar images of the HIT. The workers are not made aware of quality-control employed in the HIT. Once the HIT's are completed, the works are post-processed as part of Direct Assessment.

Table I
Percentage of people cleared DA in stage 1

	A(t-test)	B(U-test)
No of Workers who has satisfied assumption	40%	60%

Each workers reliability is accessed based on assumptions A and B Out of fifteen workers, six has cleared the DA when considering assumption A and nine has cleared the considering assumption B. The captions annotated by workers are analyzed in the next stage of post-processing.

- 6 out of 15 workers are identified as reliable by DA with assumption A, where only 4 out of 6 have annotated the image properly.
- 9 out of 15 workers are identified as reliable by DA with assumption B, where only 6 out of 9 have annotated the image properly.

The analysis of captions shows that there is no significant difference between the length of captions submitted by reliable workers satisfying assumption A and assumption B.Data Statistics on caption length submitted by reliable workers is shown in Figure 8 and Figure 9.

2) Experiment- Stage 2: The experiment setup is evaluated in stage 1 by small sample. As part of stage 2, the complete dataset is used.

- The dataset is divided 14 HITs, with each Hit consisting of 70 images,
- eight workers are asked to complete each HIT,
- 7840 captions are collected for 980 images, with each image consisting of 8 captions.

The works are collected, and the reliability of the workers is evaluated by DA.

- Each worker is evaluated with assumption A and assumption B.
- Each worker is rated manually depending on the quality of their captions as shown in table II

As part of the post-processing, the works of each crowdsource worker are evaluated statistically with t-test for assumption A and U-test for assumption B. At the end of evaluation 19 out of 112 workers have identified as reliable by DA with assumption A and 45 out of 112 workers have identified as reliable by DA with assumption B.

The worker ratings and captions of the reliable workers identified by DA are analyzed. The length of the captions by reliable workers ranges from 1 to 29.

VI. Results and Analysis

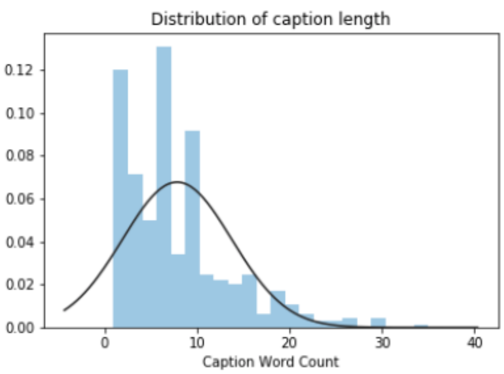
The images are annotated swiftly and economically by crowdsourcing.

	Assumption A	Assumption B
1	A pretty white horse surrounded by a rocky background and pretty flowers.	A pretty white horse surrounded by a rocky background and pretty flowers.
2	A white horse beside a flower.	A white horse beside a flower.
3	White horse.	White horse.
4	A white horse stands near a rocky structure.	A white horse stands near a rocky structure.
5		The white mayor stops to look back over the grass
6		A beautiful horse smells the flowers

Figure 6. captions by reliable workers from stage1 for an image

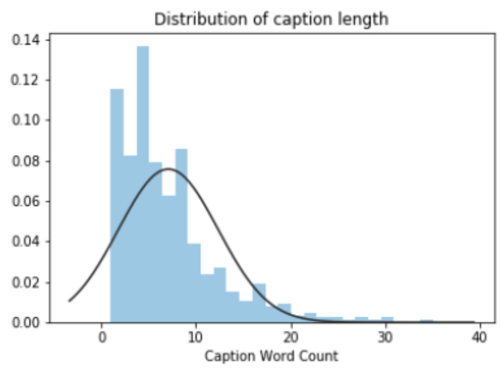
	Assumption A	Assumption B
1	A girl with black hair wearing a black Gucci shirt is holding her hand up so it looks like she is holding a red and green Christmas package that is hanging in the background.	A girl with black hair wearing a black Gucci shirt is holding her hand up so it looks like she is holding a red and green Christmas package that is hanging in the background.
2	A girl wearing a black Gucci top.	A girl wearing a black Gucci top.
3	A white Chinese girl close to a box.	A white Chinese girl close to a box.
4	An Asian woman poses in front of a display.	An Asian woman poses in front of a display.
5		A girl is holding what looks to be a drum.
6		A model is holding a present that is hanging from the ceiling.

Figure 7. captions by reliable workers from stage1 for an image



mean	7.86
std	5.90
min	1.0
25%	4.0
50%	7.0
75%	10.0
max	35.0

Figure 8. caption length from stage1 based assumption A



mean	7.08
std	5.26
min	1.0
25%	3.0
50%	6.0
75%	9.0
max	35.0

Figure 9. caption length o from stage1 based assumption B

- when DA is employed with assumption A for crowd-sourcing image annotation, only 17% of the workers is identified as reliable. The reliable workers felt difficult to differentiate system output from bad references.
- when DA is employed with assumption B for crowdsourcing image annotation, 40% of the workers are identified as reliable. The reliable workers can differentiate reference from bad reference easily
- Although is it easy to identify reliable workers without DA, by evaluating the caption length. Some workers copied irrelevant, lengthy sentences from the web to annotate images.
- when the reliable workers are rated manually by the quality of their annotation and analyzed. Quality works are identified more by DA with assumption B than DA with assumption A.

VII. Conclusion

...to be added xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Table II
Sample representation of data from post processing of stage 2

Worker ID	A	B	Comments	Ratings
1	✓	✓	Captions are very good	85
2		✓	Captions are very good	90
3			Captions are very bad	10
4	✓	✓	Captions are good	79
5	✓		Captions are bad and worker has skipped many	21
6			Captions are good, but worker has failed DA	64
7		✓	Captions are very good	89
8	✓	✓	Captions are average	53
9	✓	✓	Captions are good,worker has faced technical glitch	72
10			Captions has copied irrelevant sentences from the web	0

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

References

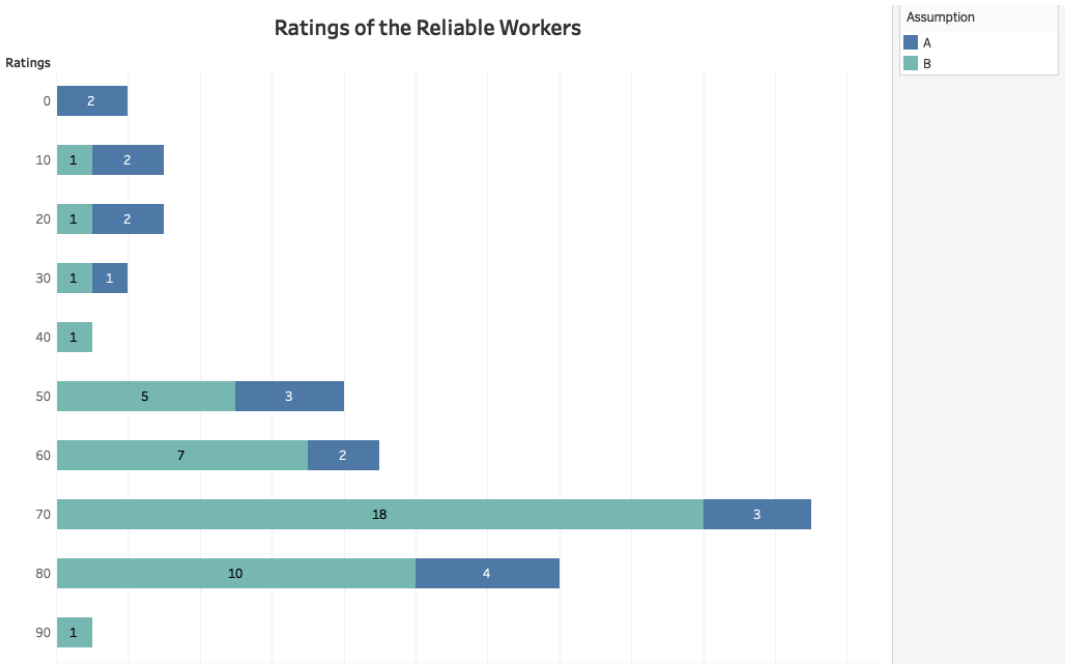


Figure 10. Rating Analysis of Reliable Workers Identified by DA