Improving ML Training Data with Gold-Standard Quality Metrics

Leslie Barrett
Bloomberg LP
lbarrett4@bloomberg.net

Michael W. Sherman
Google
michaelsherman@google.com

ABSTRACT

Hand-tagged training data is essential to many machine learning tasks. However, training data quality control has received little attention in the literature, despite data quality varying considerably with the tagging exercise. We propose methods to evaluate and enhance the quality of hand-tagged training data using statistical approaches to measure tagging consistency and agreement. We show that agreement metrics give more reliable results if recorded over multiple iterations of tagging, where declining variance in such recordings is an indicator of increasing data quality. We also show one way a tagging project can collect high-quality training data without requiring multiple tags for every work item, and that a tagger burn-in period may not be sufficient for minimizing tagger errors.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Machine learning**; • **General and reference** → *Reliability*; *Evaluation*; *Empirical studies.*

KEYWORDS

reliability, interrater reliability, inter-rater reliability, machine learning, natural language processing, natural language understanding, nlp, nlu, artificial intelligence, training data, data quality, cohen's kappa, krippendorff's alpha, crowdsourcing, annotation, tagging, data collection

ACM Reference Format:
Leslie Barrett and Michael W. Sherman. 2019. Improving ML Training Data with Gold-Standard Quality Metrics. In *KDD '19: 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 05, 2019, Anchorage, AK*. ACM, New York, NY, USA, 4 pages.

1 BACKGROUND AND PROBLEM

Human-tagged[1] training data is essential to supervised machine learning. In the last few years, services like Mechanical Turk and Figure Eight (formerly known as CrowdFlower) have increased the ease of collecting hand-tagged data, but assessing the quality of tags remains an issue with taggers of all levels of expertise. With text in particular, ambiguity in the underlying data or the tagging instructions may cause tagging inconsistencies that affect the performance of machine learning models trained on the tagged data.

This problem of inconsistent tagged data has been addressed in the research community, resulting in advances in statistical measures of consistency and agreement among human taggers. Following Carletta [4], inter-rater agreement on language-tagging tasks has taken distributional data effects into account rather than simply comparing the percentage of overlapping tags. In particular, Cohen's kappa [5] was introduced as an agreement metric for tagging natural language data.

Further research revealed deficiencies in Cohen's kappa due to its inability to account for disagreements in the data despite accounting for distributional effects. Kappa is prone to inconsistencies in the presence of skewed data, giving rise to the "paradox problem" [7], and does not handle missing values. Krippendorff [8] addressed these problems with the introduction of Krippendorff's alpha, a metric incorporating disagreement among taggers which allows for for missing data and multi-tagger scenarios. In particular Antoine et al. [2] showed that Krippendorff's alpha [8] is a more reliable metric than kappa on natural language tagging tasks including emotions, opinions and co-references. Krippendorff's alpha (Figure 1) has since been widely adopted in the NLP community and has become the standard for tagging tasks based on language data.

Figure 1: Krippendorff's Alpha

$$A = \frac{D_e - D_o}{D_o}$$

Where D_o is the observed disagreement between taggers and D_e is the expected or chance disagreement.

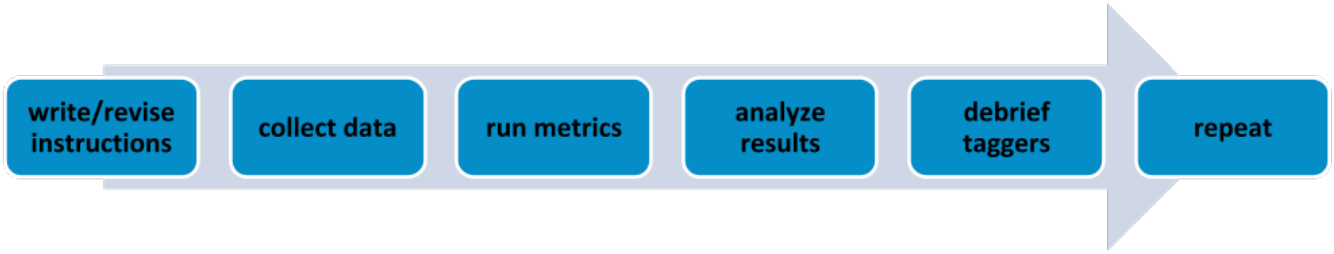
Many current studies take a single measurement of inter-rater agreement to validate a training set or provide an upper-bound on model accuracy. Generally, the main consideration is that the resulting tags are "reliable" and reproducible [3]. Others [11] use the results to alter the data, removing samples that are problematic for taggers.

Few studies have focused on determining the degree of underlying ambiguity in the data itself. Dumitrace et al.[6] create a model to represent a crowdsourcing system in three main components—workers, input units and tags. Their model proposes to explain how noise in any one of the three components influences the other components and the overall tagged data quality. In particular, it is one of the only studies to separate the noise generated by taggers from

[1]Many terms are commonly used to refer to a human manually enriching data, including "annotator", "rater", "coder", "tagger" and "labeler". For consistency we arbitrarily use "tagger" and related forms, other than when referring to "inter-rater agreement".

KDD '19, August 05, 2019, Anchorage, AK

Figure 2: Tagging Task Design Process



noise due to the inherent ambiguity of the data and the tagging task itself.

In the present study, we build upon this idea by attempting to reduce noise due to the tagging task and the taggers as much as possible, leaving only noise due to data ambiguity. Additionally, we specifically avoid replacing taggers or tagging every work item multiple times. These two common techniques for increasing tag quality suffer from the assumptions that tag creation is low cost, and that tagging task administrators are overly willing to replace individual taggers. Neither of these assumptions is necessarily the case, especially when a tagging task requires special expertise.

To create high-quality tagged data with these restrictions, we use observations of inter-rater agreement in three different contexts throughout the data collection process. First, we use inter-rater agreement measurements to design the tagging task. Next, we use inter-rater agreement measurements to determine when our taggers are fully educated[2] on the tagging task (rather than a fixed-length burn in period). Finally, we use inter-rater agreement measurements to monitor our educated taggers through the collection of the tagged data. Taking inter-rater agreement measurements in these three contexts allow for the creation of high quality data when practical concerns make collecting multiple tags on the entire dataset infeasible, while also reducing the risk associated with taking only a single inter-rater agreement measurement.

2 METHODOLOGY AND DATA

2.1 Tagging Task Description

In our tagging task, taggers were given pairs of sentences and asked to rate the quality of the paraphrase between the two sentences on a 1-4 scale, where 1 represents no similarity and 4 represents a perfect paraphrasing. This use of an ordinal tag to quantify shared meaning between two sentences is described in previous literature as Semantic Textual Similarity [1] and is closely related to paraphrase detection.

Our untagged data is from a corpus of legal text. This resulted in two specific challenges. First, we could not rely on general notions of "paraphrase" and "similarity", as these have different meanings in the legal domain than in a general context. Second, legal expertise was required to meaningfully tag the data, so taggers needed to be sourced independently (at considerable expense) rather than relying on crowdsourcing platforms.

[2]We use the term "education" and related forms to describe tagger instruction rather than "training", to avoid confusion with "training data".

2.2 Tagging Task Design

Before we began collecting large amounts of data, we wanted to make sure our tagging task was well specified enough that multiple taggers would give most sentence pairs identical scores. To accomplish this, we went through five rounds of tagging task design (Figure 2). Each round started with a set of instructions for the tagging task, followed by two experienced attorneys using the instructions to tag a set of data. We then analyzed the tagged data (including inter-rater agreement metrics). The analysis was used to debrief the attorneys, with a focus on determining what instruction misunderstandings led to disagreed tags. Finally, the insights from the debrief were used to create a new set of instructions for the next round.

After achieving a Krippendorff's alpha above .8 (.889) on the fifth round, followed by a debrief with positive feedback from the attorneys, we decided our tagging task was clear enough to proceed. Krippendorff's >.8 is considered "almost perfect agreement" based on the Landis & Koch scale [9], a commonly used benchmark for interpretation.

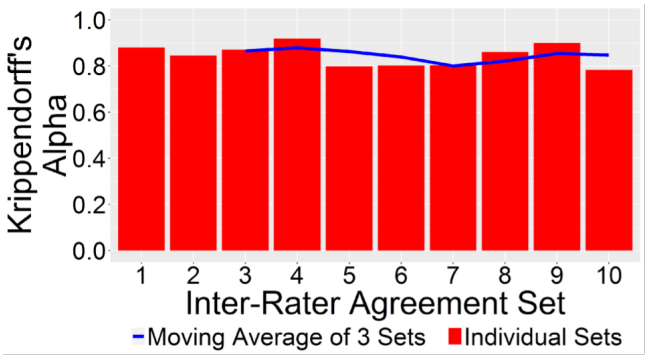
2.3 Tagger Education

For the tagging task design we used experienced attorneys as taggers, but this was financially infeasible for collecting a large amount of data. Instead, a group of five less experienced legal services professionals was engaged for subsequent tagging. First, this group was educated by the experienced attorneys who had done the tagging for the task design. Next, we recorded Krippendorff's alpha on 10 sets of "education data" to ensure the group of five taggers was producing data of sufficient quality. Each of these 10 sets of tagger education data had 50 sentence pairs (for a total of 500 sentence pairs), with every pair tagged by every tagger. While there were inconsistencies in Krippendorff's alpha across the education data sets, Figure 3 shows that a 3-set moving average of Krippendorff's alpha was always above .8, suggesting the group of five taggers was reliability tagging data. This "burn-in" phase has become common practice with crowdsourced tagging [10].

2.4 Data Collection and Tagger Monitoring

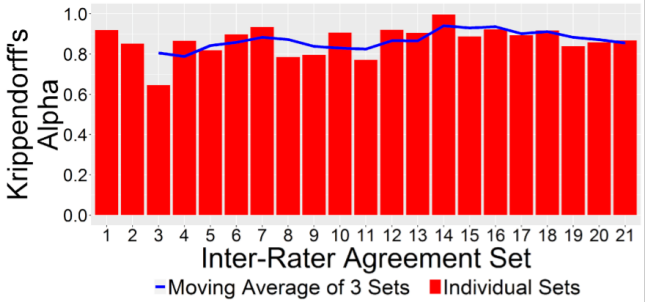
The tagging of the main dataset (35136 sentence pairs) took place over three months. Every sentence pair in the main dataset was tagged by only a single tagger. To ensure the quality of the data remained high, we collected smaller "monitoring datasets" (ranging from 60 to 150 sentence pairs) twice each week. Each sentence pair

Figure 3: Krippendorff's Alpha of Tagger Education Data



in a monitoring dataset was tagged by multiple taggers. The number of sentence pairs in the monitoring datasets assigned to each tagger was adjusted to correspond to the proportions of the main dataset tagged by each tagger, since the taggers worked at different speeds. Discrepancies in the monitoring data were inspected by experienced attorneys, who then instructed individual taggers on how they could improve.

Figure 4: Krippendorff's Alpha of Monitoring Datasets



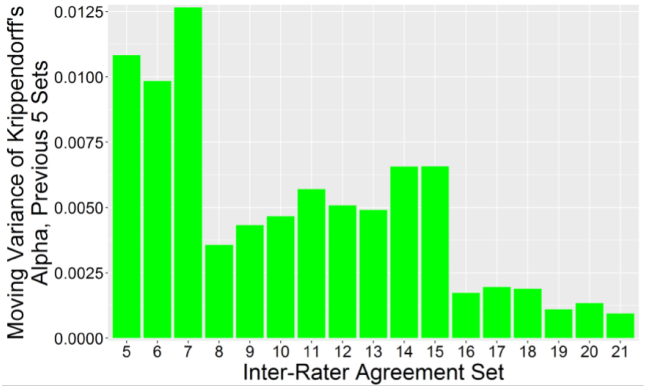
Krippendorff's alpha scores for the 21 monitoring datasets and a moving average are in Figure 4. We also calculated the moving variance of the Krippendorf's alpha values from the five previous monitoring datasets, which is seen in Figure 5.

3 DISCUSSION

Figure 3 shows relatively consistent performance by all five taggers during their education. This suggests the time invested in tagger education and tagging task design was well spent, as there is no obvious gain in agreement across these 500 tagged sentence pairs of education data. This led us to believe burn-in was complete and the tagging task was fully understood by the taggers. Initially, the alpha scores on the monitoring datasets in Figure 4 appeared to confirm this.

However, after examining the moving variances in Figure 5, the taggers continue to improve throughout the collection of the main dataset. Variance drops at both the 8th and 16th monitoring dataset, and appears to converge at a very small value at the 20th monitoring

Figure 5: Moving Variance of Krippendorff's Alpha of Five Monitoring Datasets



dataset, despite mostly high (>.8) values of Krippendorff's alpha for most of the monitoring datasets (Figure 4). This tells us that burn-in was not fully complete until very far into the tagging process.

We believe the primary driver of decreasing moving variance of Krippendorff's alpha in Figure 5 was the review of disagreeing tags in the monitoring datasets by experienced attorneys. Each example of disagreement in the monitoring datasets was considered by an experienced attorney, who then provided feedback to taggers about how they could improve. Although none of these interventions were dramatic enough to result in changes to the tagging task, these interventions did result in more consistent tagging by the five taggers based on the reduced moving variance of Krippendorff's alpha as seen in Figure 5. We hypothesize that specific interventions around the 3rd and 11th monitoring datasets were especially useful to the taggers, although we have no record of the content of conversations between the experienced attorneys and taggers.

The data further show the importance of sampling inter-rater agreement more than once when collecting multiple tags on every work item is infeasible. In this tagging task, not only was the variance of Krippendorff's alpha across multiple monitoring datasets initially high, but agreement improved over time as variance declined.

Our results also suggest there is a minimum threshold for agreement on a given dataset. This suggests common interpretations of "strength of agreement", like the one proposed in Landis and Koch [9], may only be appropriate in limited contexts where the data and tagging task are unambiguous.

We note that some of our sample sizes were small, potentially biasing Krippendorff's alpha towards outliers. In other words, drawing too few sentence pairs for each monitoring dataset could result in a monitoring dataset with a balance of "easy" and "difficult" sentence pairs different than the balance in the entire data. In the future we would consider collecting monitoring datasets with a larger number of work items, although we believe using moving metrics of agreement statistics (rather than considering each monitoring dataset's agreement individually) mitigates some of the impact of outliers.

4 CONCLUSION

Our results show it is possible to improve the quality of a human-tagged paraphrase detection dataset through multiple rounds of inter-rater agreement analysis with tagger-specific interventions based on disagreed work items. Our results also show the rolling variance of Krippendorff's alpha on monitoring datasets decreased as the taggers became more experienced. As this variance decreased, the agreement values came closer to reflecting the "true" ambiguity in the data as opposed to ambiguity contributed by outside factors like poor instructions and inadequate tagger education.

The continued drop in moving variance of Krippendorff's alpha across multiple monitoring datasets implies conventional approaches to burn-in, which focus primarily on tagging a small number of work items at the beginning of a data collection project, may not be sufficient when dealing with highly ambiguous data and/or a difficult tagging task (both which are common with language data). Additionally, our results suggest that tagger education plus a burn-in period may not maximize tagger performance without continued monitoring, and that continued monitoring of taggers reveals tagger-improving interventions that an initial analysis could miss.

Furthermore, we show a feasible alternative to having multiple tags on each work item by conducting periodic agreement studies as described. The collection of tagged data is often a cost barrier to building machine learning models, and our results show a considerable amount of savings is possible without compromising data quality.

In the future we would repeat this experiment for other tagging tasks with a view to eventually developing a more robust interpretation scale for agreement metrics like Krippendorff's alpha that take into account ambiguities in the data. For example a "reliable" result may be one in which the metric's variance is reduced by a certain amount over a given set of iterations rather than a fixed point that applies unilaterally.

REFERENCES

[1] Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task*. The Association for Computational Linguistics, 385–393.

[2] Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation.. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. ACM, 550–559.

[3] Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.

[4] Jean Carletta. 1996. Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics* 22, 2 (1996), 249–254.

[5] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[6] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018)*. CEUR-WS, 11–18.

[7] Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology* 43, 6 (1990), 543–549.

[8] Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA, USA.

[9] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[10] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), 859–866.

[11] Philipp Schaer. 2012. Better Than Their Reputation? On the Reliability of Relevance Assessments with Students. In *Third International Conference of the Cross-Language Evaluation Forum for European Languages Initiative*. Springer, 124–135.