Leadership of Data Annotation Teams

Ian McCulloh
Applied Physics Laboratory
Johns Hopkins University
Laurel, USA
Ian.mcculloh@jhuapl.edu

James Burck
Applied Physics Laboratory
Johns Hopkins University
Laurel, USA
James.burck@jhuapl.edu

Josef Behling
Applied Physics Laboratory
Johns Hopkins University
Laurel, USA
Josef.behling@jhuapl.edu

Michael Burks
Applied Physics Laboratory
Johns Hopkins University
Laurel, USA
Michael.Burks@jhuapl.edu

Jonathon Parker
Global InfoTek, Inc.
Tampa, FL USA
jkparker@mail.usf.edu

***Abstract*—Extracting (social) network data and conducting effective searches of large document collections requires large corpora of labelled, annotated training data from which to build and validate classifiers. As the importance and value of data grows, industry and government organizations are investing in large teams of individuals who annotate data at unprecedented scale. While much is understood about machine learning, little attention is applied to methods and considerations for managing and leading annotation efforts. This paper presents several metrics to measure and monitor performance and quality in large annotation teams. Recommendations for leadership best practices are proposed and evaluated within the context of an annotation effort led by the authors in support of U.S. government intelligence analysis. Findings demonstrate significant improvement in annotator utilization, inter-annotator agreement, and rate of annotation through prudent management best-practices.**

Keywords—machine learning; supervised learning; training data; inter-annotator agreement; leadership

I. INTRODUCTION

Machine learning (ML) and natural language processing (NLP) of intelligence documents has been the subject of substantial recent attention. The speed and volume of available intelligence is increasing at a faster rate than the intelligence community's (IC) ability to organize, catalog, and synthesize data [1-6]. As the capacity and availability of knowledge management systems grow, the IC requires the ability to efficiently search for relevant documents and return items that are more likely to contain items needed by analysts. Entity extraction is a related and equally important problem. Automated identification of actors, resources, organizations, and the relations between them, enable rapid construction of meta-networks used for strategic prioritization and network intervention planning [7-12]. ML and NLP can enable more efficient search to locate relevant documents.

ML and NLP approaches to the knowledge search and entity extraction problems typically require supervised learning [13-15]. Supervised learning requires a training data set that is used to develop classifiers that locate relevant content, as well as a test data set that is used to evaluate the veracity of the classifier. A high-quality data set used for training and testing classifiers is often referred to as a "gold standard data set" (GSDS).

Ensuring a GSDS meets high quality standards can be challenging. Creating such a data set not only requires close human attention, but also requires consistent human evaluation [16-18]. It requires that those humans developing the GSDS have a shared understanding of what data should be annotated and how it should be labeled.

The development of a GSDS may also require adherence to a production schedule. A data corpus must have sufficient sample size for inference and be available on time before a certain deadline. There exists a likely tradeoff between quality and schedule. Quality may be adversely affected by increasing the time annotators spend labeling data, increasing the number of annotators, changing standards during annotation, or by implementing authoritarian management practices. Leadership interventions must be carefully considered and objective quality and schedule metrics are required to evaluate the effectiveness of those interventions.

We introduce a specific annotation effort that supports a real-world intelligence requirement for the U.S. Government. Several objective metrics are then proposed to evaluate the quality and production schedule of an annotation team. Recommendations for management and leadership of annotation teams are provided to include those that were implemented during the real-world effort. Findings of effectiveness are provided, followed by recommendations for similar annotation efforts throughout government and industry.

II. BACKGROUND OF ANNOTATION EFFORT

One of the authors conducted a series of interviews with intelligence analysts throughout a specific community of interest and also reviewed existing ontologies to create a custom ontology. This ontology was reviewed by the government sponsor. The resulting ontology was then used for human annotation of intelligence documents.

The ontology was divided into six orthogonal taxonomies where each class in a lower level of the taxonomy represented a subclass of a parent class in the next higher level. Annotators were given the ability to interrelate classes using a set of predicates. A breakdown of each taxonomy, identified by its top-level class, is listed in TABLE I.

TABLE I. ONTOLOGY OVERVIEW

	# Distinct Classes	*Maximum Depth*
Object	58	4
Event	20	2
Information Artifact	16	3
Quality	6	2
Role	33	2
Place	7	1

Six annotators were trained on the government approved ontology and subsequently annotated a sample set of documents. Annotation involved marking any span of text, i.e. n-gram, that matched an ontology class. Initially, the six individuals annotated the same documents from the sample corpus in order to evaluate consistency. If annotators disagreed on the annotations, then the standards for annotation may be unclear, there may be high ambiguity in ontology classes, or annotators may be experiencing fatigue. Any issues were resolved with dialogue between annotators and management. Each issue can affect the GSDS quality. Once annotators were able to achieve a sufficient level of consistency as measured by Krippendorff's Alpha [19-20], then the annotation standards and ontology quality were verified and the level of redundant document annotation was reduced. At this point, some level of overlap redundancy is still required to monitor quality throughout the process.

Annotators used a custom software application developed by the Johns Hopkins Applied Physics Laboratory (JHU/APL). This software is an ontology-enabled information extraction and knowledge discovery platform. For this effort, the tool was loaded with the specific ontology. The software application allows an annotation manager to assign documents to annotators and monitor their progress. The tool allows annotators to annotate, or "label", n-grams that refer to entities matching a specific class in the ontology. In other words, the classes in the ontology are employed as data labels for the corresponding n-gram. The tool also allows annotators to create associations and events making linkages between labelled items. Once document annotations are complete, the annotation manager can use the tool to generate both document and annotation statistics, including Inter-Annotator Agreement (IAA) metrics.

IAA is determined using statistics computed on documents annotated by multiple annotators. While the qualitative discussions held during annotator meetings do help identify some finer points of individual annotator confusion and disagreements, a more scalable and quantitative approach is required, to assist in prioritizing the most confusable ontology classes, and identifying which (if any) classes annotators tend to accidentally skip compared to other annotators.

As the annotation effort progresses, and quality is verified, the level of redundant overlap was reduced to increase the volume of annotated documents while ensuring an acceptable level of quality. The quality of annotation was reviewed weekly by the annotation team and monthly by the management team to monitor progress and ensure quality. During the pilot phase, regular qualitative discussion meetings were held with annotators to discuss and resolve questions and confusions about which ontology classes to use for annotating particularly challenging spans of text in the documents.

III. INTER ANNOTATOR AGREEMENT

Inter-annotator agreement (IAA) was calculated using Krippendorff's alpha (*K*-α) because it handles multiple raters, random overlap in rater-document assignments, document size variation, and the potential for missed labels. *K*-α is expressed,

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{\sum_{x=1,x'=1}^{X} o_{xx'}\delta(x,x')}{\frac{1}{n-1}\sum_{x=1,x'=1}^{X} n_x n_{x'}\delta(x,x')} \quad (1)$$

where, D_o is the number of observed disagreements between annotators and D_e is the number of expected disagreements by chance, X is the set of labeling options, x and x' represent annotation labeling options for two different annotators, o is the observed pair of labels assigned by two different annotators, δ is a difference function that equals 0 when $x = x'$ and 1 when $x \neq x'$, n_x and $n_{x'}$ are the number of observed annotation labels for each class and n is the total number of pairs between annotators.

K-α holds several advantages over competing alternatives. It allows for multiple raters without inflating the metric. A traditional pairwise comparison between raters will magnify the potential error at factorial scale. For example, with two raters there may be a single comparison between the two. With three raters, there are three comparisons, and with a dozen raters there are 66 comparisons. When the number of raters varies across the documents, this can be especially problematic. Documents with more raters would have greater negative leverage on the overall quality metric. *K*-α allows a more optimal quality metric for multiple raters, allowing greater consensus assessment and thereby greater quality confidence.

Missing labels can also be an issue in text annotation. Annotation requires two cognitive tasks, recognition and classification. An annotator must first recognize an n-gram in the document as an object that requires annotation. This is the recognition task. The annotator must then classify the recognized n-gram by correlating it with a class in the ontology. This is the classification task. Because *K*-α can handle missing data, the metric can be used to measure whether quality issues are due to recognition or classification. This is an important distinction for aiding annotators in improved labeling quality.

K-α scales differently than other IAA metrics. The key difference is the benchmark for random chance. In a percent agreement metric, the random chance is a function of the number of choices that an annotator could select. Evaluation of this benchmark is challenging, because it is conditioned on the ontology structure. For example, there is only one choice for person class, but many choices for subclasses of the equipment class. The random choice benchmark in *K*-α is a function of the number of annotators. When there are only two annotators,

the benchmark for random choice is -1. As the number of annotators gets very large, the random choice benchmark converges to 0. Thus, direct comparison between percent agreement and K-α is difficult. The range of K-α is -1 to 1. An industry and academic benchmark for good IAA using K-α is 0.67 whereas very good IAA is 0.8 [19]. A rough conversion from these industry standards to percent agreement would approximately be 84% and 90% respectively. Thus, we will use K-α of 0.67 and 0.8 as the target metrics for the GSDS.

IV. Proposed Quality Metrics

When developing a GSDS, it is important to understand potential sources of annotation error so that corrective measures can be introduced to increase IAA to an acceptable quality level. It is also important to understand performance schedule metrics to estimate when the GSDS will be complete and identify any risk to schedule that might be incurred with interventions to improve quality. We propose metrics that are separated into two categories: quality metrics and schedule metrics. Both are important for managing an annotation effort.

A. Quality Metrics

Recognition measures the extent to which annotators identify n-grams that should be labelled according to the ontology. An annotator might fail to recognize a n-gram as needing annotation, but if he or she were presented with the n-gram, they would classify it the same as other annotators. Recognition can be measured by comparing K-α across a set of documents with the K-α value calculated only on n-grams that have been labeled by multiple annotators, even if the labeling is not consistent. For example, if annotators 1 and 2 assign a class A to an n-gram and annotator 3 assigns no class, the standard K-α measure would indicate disagreement. When calculating recognition, however, the K-α measure would only be calculated for annotators 1 and 2 who both correctly classified the n-gram, making agreement 1.0. By comparing the K-α calculated only on labeled items with K-α including overlooked items, the impact of recognition on annotator agreement can be measured. The reduced number of n-grams to consider affects the sample space for evaluation and therefore the metric. If recognition is a problem, then K-α will improve for the unanimously labelled condition.

Scale complexity measures error that stems from cognitive limitations in the number of elements that a person can distinguish between in an annotation task. Given no context or structure, annotators will have more difficulty in correctly labeling items with larger ontologies [21]. The number of items that can be distinguished can be improved with context, annotator expertise, and with classification into a deeper ontology [22]. The scale complexity error can be approximated by comparing the K-α computed on a set of documents to the K-α when the ontology is collapsed to high level classes. A collapsed ontology would revert labels to their parent class. For example, two annotators may agree an item is a car, but disagree on whether it was a Toyota or a Honda, or whether it is a Camry or Corolla. Improvement in K-α following ontology collapse is a measure of scale complexity error.

Rater Vitality measures how well individual annotators agree with the larger group of annotators. This measure scores their impact to the overall K-α score. The K-α measure is calculated for all annotators, K_A, and for all annotators except an individual i, K_i. The vitality of annotator i is $v_i = K_A - K_i$. Thus, annotators with a positive vitality score, increase consensus, while annotators with negative vitality scores decrease consensus. This measure allows the annotation manager to identify annotators that may be conducting inconsistent annotations compared to other annotators. Coupled with metrics for recognition and scale complexity, the annotation manager can identify potential issues and conduct retraining.

Rater Consistency measures how challenging documents are to annotate. This potential error is measured by having an annotator review the same document on separate days and compare the consistency of annotation. This can be measured by percent agreement, since multiple raters are not involved and the percent agreement metric is easier to interpret for annotator training. The measure can be calculated for the entire ontology and collapsed ontology structure to evaluate whether potential sources of error are due to recognition or classification problems. The annotation manager can use this measure to identify retraining requirements. When documents have a low percent labelling agreement from multiple annotators, these documents should be investigated to qualitatively understand why the annotation is more challenging for the identified documents. Insights can be used for retraining. Calculation of this metric requires annotators to repeat labeling of previously viewed documents, which reduces the resources to label additional documents. For this reason and several schedule constraints, this measure was not calculated for the example project discussed in this paper.

Leverage Risk measures a potential bias that can occur when using a completed GSDS due to ontology classes consisting of items that occur infrequently in the corpus. The relative low-density of these items creates greater leverage for these rare ontology items when using the GSDS to develop classifiers. There are two solutions for correcting this issue: increase the size of the GSDS to capture the rare ontology items; or, bias the sample of documents to include rare ontology items at greater frequency. The problem with the first option is that it can exponentially increase the cost for annotating a GSDS and will involve oversampling common ontology items. The problem for the second issue is that the biased selection of documents may not be representative of the overall sample, which could lead to challenges of veracity in the final GSDS corpus. While this error cannot be directly measured, it is possible to identify low density annotation classes and items. These classes and items should be considered for removal from the ontology, increase of GSDS size, or biased sampling, depending on the importance of the items to the corpus.

B. Schedule Metrics

Schedule metrics are important for planning the number of annotators required to develop a GSDS of sufficient size by a specified deadline. They also help an annotation manager

identify potential schedule risks incurred by implementing measures for quality improvement.

Basic annotation statistics include the **number of documents annotated**, **average words per document**, **average labels per document**, and **total number of annotated n-grams**. Using these basic statistics, the **annotation density** is the average labels per document divided by the average words per document.

Rates can be calculated using basic annotation statistics and noting the time involved in annotation. The annotation software system records the time annotators spend logged into the system, completing annotation tasks. The **document rate** is the number of documents divided by the time spent annotating. The **labeling rate** is the annotation density multiplied by the document rate.

There are several annotator dependent metrics. The **average number of annotators per document** is a measure of document overlap. Higher overlap may be necessary to produce quality metrics. When quality metrics indicate sufficient quality, the overlap can be reduced to increase the document and labeling rates to meet schedule requirements. If quality rates fall, then annotator overlap should be increased to investigate potential threats to quality until issues are identified and resolved.

The **distraction rate** is an annotator specific metric, measuring the average number of sessions the annotator uses to annotate a document. If an annotator does not interact with the annotation software for 6 minutes, the system logs them out. The annotator must log back in to complete the task, which indicates an additional session. When an annotator must walk away from their annotation task, their concentration and focus is disrupted which may adversely affect quality. Conversely, if an annotator spends too much time conducting annotation, they may become fatigued, which may also adversely affect quality.

The final metric is **utilization**. This is also an annotator specific metric which compares the amount of time an annotator is logged into the annotation software to the amount of timed billed to a customer or project. We operationalize this as the number of hours logged into the annotation software divided by the number of hours billed to the annotation project.

V. FINDINGS

The annotation effort for the U.S. Government was conducted over six months and involved 10 annotators. During the pilot phase, there were only six annotators. The *average number of annotators per document* was high at 3 annotators per document to generate quality metrics for establishing effective annotator training standards and guidance. The production phase was conducted over four months and included 10 annotators. The ontology consisted of nine parent classes and 150 total ontology elements.

Two quality issues were identified in the pilot phase, recognition and scale complexity. TABLE II. reports the IAA metrics in a confusion table. The top left cell displays the overall *K*-α measure for all documents and annotators. The right column displays the *K*-α measure when only unanimously labeled items are used as in the recognition metric. The bottom row displays the *K*-α measure when the ontology is collapsed to the high-level parent classes as used in the scale complexity metric. The recognition error is therefore, 0.804 - 0.638 = 0.17. The scale complexity error is therefore, 0.702 - 0.638 = 0.06. Note there is an interaction effect and the combined recognition and complexity error is 0.919 - 0.638 = 0.281, which is greater than the sum of the two errors. This implies that the scale complexity may contribute to annotator recognition.

TABLE II. IAA CONFUSION TABLE

	Agreement	***Correcting for Recognition***
Agreement	0.638 (~82%)	0.804 (~90%)
Correcting for Complexity	0.702 (~85%)	0.919 (~96%)

Annotator overlap was reduced to increase the production rate and periodically increased as the new annotators were trained to ensure they meet quality standards. Overlap was also increased periodically to verify quality was maintained. TABLE III. provides the same quality metrics that were displayed in TABLE II. for the subsequent production phase.

It can be seen that the *K*-α values in TABLE III. are lower in the left column than TABLE II. , but the right column remains comparable. This is indicative of increased recognition error. Upon investigation, the recognition error is due to new annotators joining the effort. There was very little overlap in document annotation between experienced annotators. This finding illustrates how the metric can be used to monitor quality and signal potential changes.

TABLE III. IAA CONFUSION TABLE

	Agreement	***Correcting for Recognition***
Agreement	0.511 (~75%)	0.803 (~90%)
Correcting for Complexity	0.558 (~78%)	0.936 (~97%)

Quality issues can be further investigated by inspecting the rater vitality of individual annotators. Figure 1 illustrates a plot of vitality, where the horizontal axis represents rater vitality and the vertical axis represents the rater vitality, corrected for recognition error. These two vitality scores are correlated with a coefficient of determination of $R^2 = 0.48$. Annotators with a negative, horizontal residual are plotted to the left of the trend line, such as annotator A, and have problems with recognition. Annotators with a negative, vertical residual are plotted below the trend line, such as annotator B, and have problems with correctly labeling text. The vitality plot can be used to identify potential problem annotators, diagnose the training deficiency, and develop an appropriate retraining intervention.

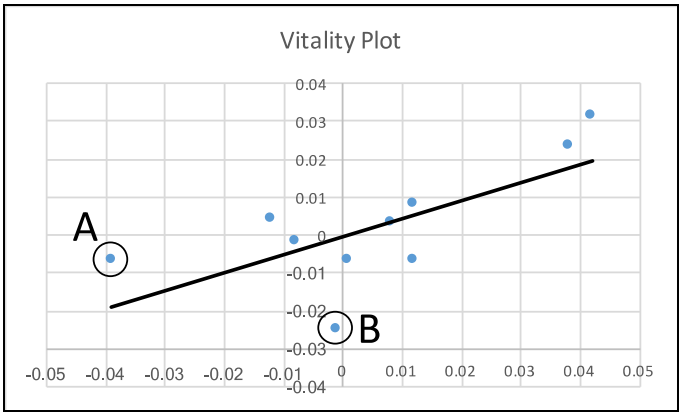

Fig. 1. Rater Vitality Plot

Utilization is a concern associated with delivering a GSDS on time and on budget. Utilization rates fluctuated between 60-100%. There is a drop in utilization to 22% for one day, which corresponds to a day where annotators were required to conduct training on annotation standards and was anticipated. Other fluctuations are affected by bringing new annotators into the effort and training them accordingly. As the annotators move into a performance phase, the utilization exceeds 80%, which deemed acceptable to the project manager and customer.

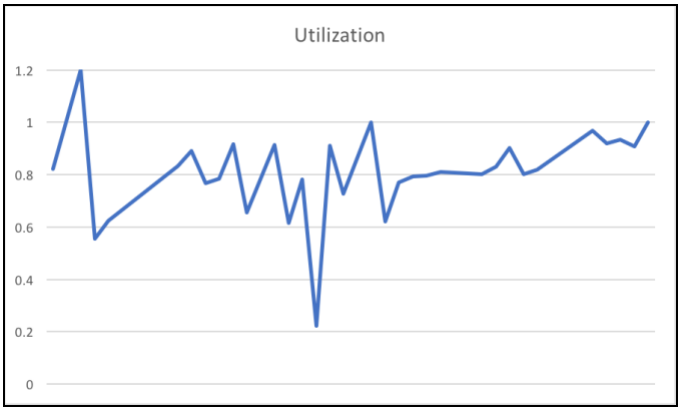

Fig. 2. Annotator Utilization Rate Over Time

The distraction rate was regressed against the rater vitality using data from the initial pilot phase and again during the performance phase. During the pilot phase annotators would review a document in as many as 15 different sessions that might span more than 2 days. During this phase, there was a statistically significant negative correlation between distraction rate and rater vitality at the $p < 0.0001$ significance level. This indicates that increased distraction was associated with worse IAA. The annotation manager implemented a policy that documents must be annotated within the same day. During the implementation of this policy, there was no statistically significant relationship between distraction rate and rater vitality. The trend between the relationship actually reversed indicating that some breaks in the annotation may actually improve annotation quality.

VI. Conclusions

As more and more businesses and government agencies begin to leverage the power of machine learning to automate document search, the management of annotation efforts will increase in importance. There are a number of potential errors that can be introduced in an annotation task. We have presented five within this paper. There are also multiple metrics that must be monitored to both ensure on-time delivery and proper resourcing, as well as to measure the impact of performance improvement interventions. We have presented 10 with this paper. Executives wishing to leverage machine learning to automate knowledge management processes must be equipped with these metrics in order to monitor annotation efforts and to make effective risk and resourcing decisions. The quality metrics are also useful for evaluating the degree of reliability present in a completed GSDS.

The use of the proposed quality and schedule metrics were demonstrated on a real-world annotation effort. While potential quality issues are inherent in any annotation effort, the proposed quality metrics were shown to be useful in identifying and quickly resolving issues. The schedule metrics were also shown to be useful for identifying schedule risk. Used together, the annotation manager was able to make in course corrections, conduct annotator retraining, clarify standards, and adjust the number of annotators to deliver a high quality GSDS on time and on budget.

There are several limitations inherent within this project. While the authors believe we have proposed a comprehensive list of potential metrics to monitor, no doubt other metrics will emerge. For each of the quality concepts we have presented, there are multiple ways of calculating equivalent metrics. The exploration of potential metrics is well beyond the scope of this paper, which is intended to introduce the concept of quality management in GSDS construction. This work was demonstrated on a single, resource intensive application. Other applications may exhibit quality problems in different areas. Our hope, however, is that this paper introduces an important concept in machine learning – management of annotation.

Leaders must not rely on machine learning as a black box. It is important that managers understand the basic concepts involved in machine learning in general and the construction of reliable training corpora in particular. By understanding these concepts, advancements in management practices can be tailored to machine learning application. We believe this will increase management's confidence in the methods, ensure more effective use of the approach, and expand the use of machine learning to new application areas.

Acknowledgment

Preparation of this manuscript was supported by the Office of Naval Research, Grant No. N00014-17-1-2981/127025. The annotation effort was supported under a contract from the U.S. Government. Special thanks to the annotation team and annotation managers supporting the paper.

REFERENCES

[1] Akhgar B, Saathoff GB, Arabnia HR, Hill R, Staniforth A, Bayerl PS. "Application of big data for national security: a practitioner's guide to emerging technologies." Butterworth-Heinemann; 2015 Feb 19.

[2] Chen CP, Zhang CY. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information Sciences. 2014 Aug 10;275:314-47.

[3] McCue C. "Data mining and predictive analysis: Intelligence gathering and crime analysis". Butterworth-Heinemann; 2014 Dec 30.

[4] Savas O, Sagduyu Y, Deng J, Li J. "Tactical big data analytics: challenges, use cases, and solutions." ACM SIGMETRICS Performance Evaluation Review. 2014 Apr 17;41(4):86-9.

[5] Stewart J. "Strong Artificial Intelligence and National Security: Operational and Strategic Implications." Naval War College Newport Ri Joint Military Operations Dept; 2015 May 18.

[6] Symon PB, Tarapore A. "Defense intelligence analysis in the age of big data." Joint Forces Quarterly—JFQ. 2015;79:4-11.

[7] Agichtein E, Gravano L. "Snowball: Extracting relations from large plain-text collections." In Proceedings of the fifth ACM conference on Digital libraries 2000 Jun 1 (pp. 85-94). ACM.

[8] Carley KM, Morgan GP, Levine J. "Socio-Cultural Cognitive Mapping." Center for the Computational Analysis of Social and Organizational Systems. 2017 Aug 31.

[9] Diesner J. "Network Construction based on Structured and Unstructured Text Data" in ConText.

[10] Diesner J, Evans CS, Kim J. "Impact of Entity Disambiguation Errors on Social Network Properties." In ICWSM 2015 May (pp. 81-90).

[11] Diesner J, Carley KM. "Words and networks." Encyclopedia of Social Networking. 2011:958-61.

[12] McCallum A. "Information extraction: Distilling structured data from unstructured text." Queue. 2005 Nov 1;3(9):4.

[13] Beitzel SM, Jensen EC, Frieder O, Lewis DD, Chowdhury A, Kolcz A. "Improving automatic query classification via semi-supervised learning." In Data Mining, Fifth IEEE international Conference on 2005 Nov 27 (pp. 8-pp). IEEE.

[14] Buro M. "Improving heuristic mini-max search by supervised learning." Artificial Intelligence. 2002 Jan 1;134(1-2):85-99.

[15] Janikow CZ. "A knowledge-intensive genetic algorithm for supervised learning. " Machine learning. 1993 Nov 1;13(2-3):189-228.

[16] Brodley CE, Friedl MA. "Identifying mislabeled training data." Journal of artificial intelligence research. 1999;11:131-67.

[17] Engelson SP, Dagan I. "Minimizing manual annotation cost in supervised training from corpora." In Proceedings of the 34th annual meeting on Association for Computational Linguistics 1996 Jun 24 (pp. 319-326). Association for Computational Linguistics.

[18] Wiebe JM, Bruce RF, O'Hara TP. "Development and use of a gold-standard data set for subjectivity classifications." In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics 1999 Jun 20 (pp. 246-253). Association for Computational Linguistics.

[19] Hayes, Andrew F., and Klaus Krippendorff. "Answering the Call for a Standard Reliability Measure for Coding Data". Communication Methods and Measures 1, no. 1 2007: 77-89.

[20] Krippendorff, Klaus (2011). "Computing Krippendorff's Alpha-Reliability." Retrieved from http://repository.upenn.edu/asc_papers/43

[21] Kinnell A, Dennis S. "The list length effect in recognition memory: An analysis of potential confounds." Memory & Cognition. 2011 Feb 1;39(2):348-63.

[22] Smith SM, Vela E. "Environmental context-dependent memory: A review and meta-analysis." Psychonomic bulletin & review. 2001 Jun 1;8(2):203-20.

[23] Piatko, C., Burck, J., Behling, J., McCulloh, I. "Distributed Common Ground/Surface System –Gold Standard Data Set (GSDS): Metrics And Quality Assurance Assessment Plan" Johns Hopkins Applied Physics Laboratory Technical Report No. AOS-17-0441. Dec 2017.

[24] Piatko, C., Burck, J., Behling, J. "Distributed Common Ground/Surface System –Gold Standard Data Set (GSDS): Annotation Guidelines" Johns Hopkins Applied Physics Laboratory Technical Report No. AOS-17-0440. Apr 2017.