


An Informal Guide To Implementing The Gold Standard (Draft that is not to be published)

AISHWARYA BALIVADA ^{1,2} HERBERT SCHILLING (MENTOR)¹ AND BRANDON RUFFRIDGE (MENTOR)¹

¹*NASA Glenn Research Center
21000 Brookpark Rd,
Cleveland, OH 44135*
²*Purdue University
610 Purdue Mall,
West Lafayette, IN 47907*

ABSTRACT

This informal guide will go over what the Gold Standard is and what the steps are to implement this into our machine-learning models. We will briefly go over what the best practices are when creating the Gold Standard dataset and what this would mean for information extraction in NASA PeTaL. We also do a side-by-side comparison of noisy data versus Gold Standard data. We find that manual human vetting leads to a richer Gold Standard dataset. This is supposed to serve as an informative guide to using the Gold Standard. The importance of having this dataset is stressed upon and must be made so we can avoid ambiguity and reach optimal performance in our machine learning model or artificial intelligence systems. Future advances will change how the Gold Standard is implemented, but for now this paper guides us to the best practices performed by others when implementing this type of dataset.

Keywords: Data Quality Control — AI and ML — Natural Language Processing — Information Extraction

1. INTRODUCTION

1.1. *The Gold Standard*

What is the Gold Standard? The Gold Standard are datasets that have been gone through human vetting. It often represents the "objective truth" (Petkova (2021)). The datasets are text from papers. We will be focusing on biomimicry papers for most of the examples throughout this journal. NASA PeTaL (Periodic Table Of Life) is an open-source artificial intelligence tool which advances biomimicry research. Biomimicry is an approach to using nature's designs and strategies for use in human-made products. The use of the Gold Standard is to have "rich" data in order to accurately train machine learning models.

Artificially intelligent systems are fed with various data sources, but often times there is a lot of ambiguity amongst the compiled dataset. For example, one paper in a dataset could talk about a topic, but they miss a few key pieces that the other paper covers on the same topic. This ambiguity is avoided via data linking techniques. This finds, matches, and combines data records covering the same topic. It also identifies entities that seem similar. The Gold Standard dataset is manually verified through two human raters to avoid bias. It is recommended to have a well-defined rating system with proper guideline for annotation. The steps to implement the Gold Standard will be discussed in the next subsection.

How does the Gold Standard help with information extraction? There are two types of texts: structured and unstructured. Unstructured text needs to be converted to machine-readable facts. This happens through a process called semantic annotation where text is tagged and enriched with metadata (Petkova (2021)). However, one obstacle would be the ambiguity of the annotations. So this all depends on how strictly the human raters can remove the ambiguity of data and avoid bias. This is how the Gold Standard is applied to unstructured texts. We in turn receive structured data and more "richer" datasets as a result. Though this may be time-consuming, it has been proven through multiple sources to ensure accuracy when training our machine learning models.

1.2. Implementation Of The Gold Standard: Best Practices On The Gold Standard According To Other Research Papers

We have defined the Gold Standard and discussed its importance. Now how can we implement it into our datasets? First and foremost, we should eliminate any journals that are not a part of the topic that we would study. Journal titles can give us an idea of whether this is the topic we should focus on. Sometimes titles cover various topics and eliminating them may sound counterintuitive, but we want our machine-learning models to not learn on ambiguous text. For example, if we are focusing on biomimicry papers, we will look through a database solely focusing on them and pick out any titles that are too ambiguous and isn't worth focusing on. Eliminating by title will narrow down text, but we will need to do more vetting to obtain a more focused dataset.

Often times, datasets that are picked out can be loosely categorized, leading to ambiguous data that will make information extraction involve more time-consuming steps. So, there should be a way to classify the papers in datasets based on relevance to the topic you want to focus on. A method was found that involves classifying 250 journals. The method involved rejecting all the journals that were catalogued and interchangeable with other adjective Multidisciplinary. Adjective Multidisciplinary is a one-topic journal that can correlate with other subject disciplines. It is possible to group journals based on one criterion that human vetters would like to use to determine whether the journal is relevant to the topic. This criterion is called a category. Normally, there are multiple categories present to create "rich" datasets. The requirement for the Gold Standard dataset is the need for the papers in the dataset to already be assigned to a different attributes and their values. Another way to achieve Gold Standard as described by López-Vázquez et al. (2022), is based off of article networks and how the architecture of these networks are used to analyze different papers. If we want accuracy and less time-consumption when classifying text, then a second must for a Gold Standard dataset is to have partial classification. Partial classification are models that show data classes, but aren't containing all aspects of that class.

A Gold Standard will be important in automatic text classification and can be obtained manually or automatically. The preferred method by most is through human experts that make a network of connections between text or documents to identify a field of knowledge. To label the Gold Standard data, there are three main steps: pre-labeling, labeling, and post-labeling. When we pre-label we use a binary classification, which determines whether the paper is or isn't talking about the specific topic you want to acquire data on. Then, experts familiar with the subject of the articles' terminology evaluate each text or document by comparing it to a specific definition of the subject that the dataset would cover (López-Vázquez et al. (2022)). For example, for biomimicry papers, we want the expert to evaluate the papers by comparing it to a definition of biomimetic functions. This will answer whether the paper is talking about biomimicry or not. This is just one example of this method being applied to the NASA PeTaL project.

In order to have a proper evaluation, we need to come to a consensus between the experts that vetted the text. So, we select reviewers and articles at random from a set of text and perform two-stage random sampling. Binary classification has three metrics: Simple Agreement, Krippendorff Alpha Index, Gwet AC Index. All of them assess the accuracy and how reliable the final results are. The metrics are determined after the voting scheme amongst the random experts for the pre-labeling classification portion. When we label, stratified sampling builds a second data set. Stratified sampling is a sampling method that decreases sampling error. Through random selection of articles and experts, the experts receive personalized systematically selected articles. Three evaluations are made from various experts. Anonymity was achieved by having two other experts outside of the random selection to view selected articles (López-Vázquez et al. (2022)). Binary classification was determined by majority vote. For post-labeling, reliability indices are computed. There is no outliers found, so the final class was determined by majority vote amongst the experts.

The reviewers used three categories when viewing and labeling the data: "purpose", "how", and "what". All the three categories allowed experts to perform sentence-level classifications of different document's abstracts, in order to figure out where the author's emphasis is. "Purpose" is the motivation behind the work. "How" is the methodology of the work. "What" is the result of the work (López-Vázquez et al. (2022)). If the article focused on the results, then the article is focusing on the application of the work. If the "how" is greatly emphasized then experts go through a second stage of vetting, where they compare the content with the definition that defines the binary classification problem. In our case, we want to know whether the content of the document highlights biomimetic functions. We evaluate this binary classification problem by comparing it to the definition of biomimetic functions and what classifies as biomimetic functions. "Purpose" was not considered in López-Vázquez et al. (2022) vetting process for labeling and post-labeling our dataset.

Experts mark the sentences that are considered "purpose", "how", and "what". Then, they perform a word count for each category they marked. If the word count for "what" is greater than the word count for "how", then López-Vázquez et al. (2022) determines the papers are not what they are looking for and eliminate the abstract from the list. If the word count for "how" is greater than the word count for "what", the expert goes through a second stage of questioning. That involves checking if the abstract, for example, correlates with any aspects to the definition of biomimetic functions. If any of the questions affirm the definition, then the text is a biomimetic function. Similar words will need to be left to the expert to write in a comments section of their evaluation (López-Vázquez et al. (2022)). There should be an original criteria which can be amongst the experts.

Results of classification are often times analyzed using indicators compared with reference values in that source. The Cohen kappa index identifies the classification quality. The Holsti reliability coefficient divides the total cases and finds a consensus between experts' evaluation. Krippendorff's alpha takes values in between $[-1,1]$, with 1 being the best agreement and 0 being random classification. AC_1 uses the same interval, but its purpose is to correct random coincidences. It is an indicator used by most experts where easily classifiable documents will be classified correctly, but hard classifiable documents would be random uniform distributions (López-Vázquez et al. (2022)). The Cohen kappa index often serves as a guideline for Krippendorff's alpha. There is a re-sampling technique called Bootstrap and it randomly extracts and shows a percentage of classified cases for each re-sample. This at the very least avoids ambiguity and repetition within different documents in the sample.

Sometimes manual evaluation are difficult to obtain due to the different experts involved and are time-consuming. This is where we use automation. Ontology Matching tools are data sources. The features of the proposed model, in a binary manner, translate where mapping was absent (0) or present (1) in the output. Reference alignment produces target class and supervised learning to classify and validate the mapping of two ontologies. Ontology Matching tools output patterns which use the ground truth as the reference alignment. Ground truth is the goal or target for training a model with labeled data. Reference alignment is key to ontology alignment evaluation. Manual reference alignments for the Ontology Matching systems have high performance. Three ontologies can be matched pair-wise creating three alignments with the reference alignment being extracted from an external source (Lima et al. (2020)). Normally, different datasets are randomly generated. They each contain a majority of alignments for training and 2 to 3 alignments for testing. The reference alignment are true positive mappings for this case. Randomly sampling pairs of entities that aren't there in the reference alignment are negative examples. We take all mappings of these examples that the Ontology Matching tool finds and we find out two different sampling strategies: oversampling and undersampling. Ten-fold cross-validation with a grid-search of hyperparameter tuning over a few machine learning techniques is a baseline (Lima et al. (2020)) to this evaluation of data. Feature extraction of Ontology Matching tools for training and testing was carried out to see if a model generalizes with other tasks.

Another method used by Lima et al. (2020) was training/testing data of different domains. The majority vote was calculated and taken into consideration for all the methods mentioned. As a result, cross-validation receives good F1-scores. The feature-extraction performed well with a better F1-score when the model was trained on the data of a certain domain. The last method of training and testing on different domains received high F1-scores and the machine learning models outperformed (Lima et al. (2020)) the previous methods. We can conclude that a model trained on alignment tasks from another domain gives us a better score, which will help us when evaluating the Gold dataset. This evaluation metric allowed us to determine whether these automated techniques are optimal when evaluating our Gold Standard dataset.

2. NOISY DATA VERSUS GOLD STANDARD DATA

Noise has a "regularization effect" that improves machine learning models' validity, but will too much of noise cause issues in our models? Models trained on noisy labels can be on par with models trained on the Gold Standard Data, but the models don't perform well due to the ambiguity. After thorough studies by Tekumalla & Banda (2021) on deep learning models which used preprocessed train/test sets, the transformer model-BERT-outperformed other models in precision, accuracy, recall, and F-measure (Tekumalla & Banda (2021)). The BERT model exceeded the results with Gold Standard dataset versus noisy dataset. We can conclude supervised and weak supervised learning is best with large annotated Gold Standard datasets. Whether it is done manually or automatically is up to the vetters to decide. Manually curating the dataset through multiple vetting seems to be optimal for our machine learning models and artificially intelligent systems.

3. FUTURE WORK AND CONCLUDING REMARKS

146 One of the projects NASA PeTaL is focusing on is information extraction. In particular, we want a named-entity
147 recognition that is able to identify a biomimetic function and extract it so it is included in our biological summaries
148 of the text. Some of the different indicators mentioned in the above subsection, can be used when creating a Gold
149 Standard dataset. Measuring confidence can benefit the expert's opinion on each article (López-Vázquez et al. (2022)).
150 This can be a step that we can take after the manual human vetting. Technology may automate this process, but the
151 human vetting process is proven to be the solution for acquiring rich data.

REFERENCES

- 152 Lima, B., Branco, R., Castanheira, J., Fonseca, G., & 154 López-Vázquez, C., Gonzalez-Campos, M.,
155 Bernabé-Poveda, M., et al. 2022, IEEE
153 Pesquita, C. 2020, CEUR-WS 156 Petkova, G. 2021, ontotext: Making Sense of Text and Data
157 Tekumalla, R., & Banda, J. 2021, Springer Link