GOLD STANDARDS IN ENTERPRISE KNOWLEDGE PROJECTS

Michael K. Bergman[1], Coralville, Iowa USA
July 18, 2016
AI3:::Adaptive Information blog

It is common — if not nearly obligatory — for academic researchers in natural language processing (NLP) and machine learning (ML) to compare the results of their new studies to benchmark, reference standards. I outlined some of the major statistical tests in a prior article [1]. The requirement to compare research results to existing gold standards makes sense: it provides an empirical basis for how the new method compares to existing ones, and by how much. Precision, recall, and the combined F1 score are the most prominent amongst these statistical measures.

Of course, most enterprise or commercial projects are done for proprietary purposes, with results infrequently published in academic journals. But, as I argue in this article, even though enterprise projects are geared to the bottom line and not the journal byline, the need for benchmarks, and reference and gold standards, is just as great — perhaps greater — for commercial uses. But there is more than meets the eye with some of these standards and statistics. Why following gold standards makes sense and how my company, Structured Dynamics, does so are the subjects of this article.

A Quick Primer on Standards and Statistics

The most common scoring methods to gauge the "accuracy" of natural language or supervised machine learning analysis involves statistical tests based on the ideas of negatives and positives, true or false. We can measure our correct 'hits' by applying our statistical tests to a "gold standard" of known results. This gold standard provides a representative sample of what our actual population looks like, one we have characterized in advance whether results in the sample are true or not for the question at hand. Further, we can use this same gold standard over and over again to gauge improvements in our test procedures.

'Positive' and 'negative' are simply the assertions (predictions) arising from our test algorithm of whether or not there is a match or a 'hit'. 'True' and 'false' merely indicate whether these assertions proved to be correct or not as determined by the reference standard. A false positive is a false alarm, a "crying wolf"; a false negative is a missed result. Combining these thoughts leads to a confusion matrix, which lays out how to interpret the true and false, positive and negative results:

Correctness	***Test Assertion***	
	Positive	***Negative***
True	**TP** True Positive	**TN** True Negative
False	**FP** False Positive	**FN** False Negative

These four characterizations — true positive, false positive, true negative, false negative — now give us the ability to calculate some important statistical measures.

[1]Email: mike@mkbergman.com

The first metric captures the concept of *coverage*. In standard statistics, this measure is called sensitivity; in IR (information retrieval) and NLP contexts it is called recall. Basically it measures the 'hit' rate for identifying true positives out of all potential positives, and is also called the true positive rate, or TPR:

$$TPR = TP/P = TP/(TP + FN)$$

Expressed as a fraction of 1.00 or a percentage, a high recall value means the test has a high "yield" for identifying positive results.

Precision is the complementary measure to recall, in that it is a measure for how efficient whether positive identifications are true or not:

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positives}}$$

Precision is something, then, of a "*quality*" measure, also expressed as a fraction of 1.00 or a percentage. It provides a positive predictive value, as defined as the proportion of the true positives against all the positive results (both true positives and false positives).

Thus, recall gives us a measure as to the breadth of the hits captured, while precision is a statement of whether our hits are correct or not. We also see why false positives need to be a focus of attention in test development: they directly lower precision and efficiency of the test.

That precision and recall are complementary and linked is reflected in one of the preferred overall measures of IR and NLP statistics, the F-score, which is the adjusted (beta) mean of precision and recall. The general formula for positive real β is:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

which can be expressed in terms of TP, FN and FP as:

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{true positive}}{(1 + \beta^2) \cdot \text{true positive} + \beta^2 \cdot \text{false negative} + \text{false positive}}$$

In many cases, the harmonic mean is used, which means a beta of 1, which is called the F_1 statistic:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

But F1 displays a tension. Either precision or recall may be improved to achieve an improvement in F_1, but with divergent benefits or effects. What is more highly valued? Yield? Quality? These choices dictate what areas of improvement need to receive focus. As a result, the weight of beta can be adjusted to favor either precision or recall.

Accuracy is another metric that can factor into this equation, though it is a less referenced measure in the IR and NLP realms. Accuracy is the statistical measure of how well a binary classification test correctly identifies or excludes a condition:

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.

All of the measures above simply require the measurement of false and true, positive and negative, as do a variety of predictive values and likelihood ratios. Relevance, prevalence and specificity are some of the other notable measures that depend solely on these metrics in combination with total population [2].

Not All Gold Standards Shine

Gold standards that themselves contain false positives and false negatives, by definition, immediately introduce errors. These errors make it difficult to test and refine existing IR and NLP algorithms, because the baseline is skewed. And, because gold standards also often inform training sets, errors there propagate into errors in machine learning. It is also important to include true negatives in a gold standard, in the likely ratio expected by the overall population, so as to improve overall accuracy [3].

There is a reason that certain standards, such as the NYT Annotated Corpus or the Penn Treebank [4], are often referenced as gold standards. They have been in public use for some time, with many errors edited from the systems. Vetted standards such as these may have inter-annotator agreements [5] in the range of 80% to 90% [4]. More typical use cases in biomedical notes [6] and encyclopedic topics [7] tend to show inter-annotator agreements in the range of 75% to 80%.

A proper gold standard should also be constructed to provide meaningful input to performance statistics. Per above, we can summarize these again as:

- **TP** = standard provides labels for instances of the same types as in the target domain; manually scored
- **FP** = manually scored for test runs based on the current configuration; test indicates as positive, but deemed not true
- **TN** = standard provides somewhat similar or ambiguous instances from disjoint types labeled as negative; manually scored
- **FN** = manually scored for test runs based on the current configuration; test indicates as negative, but deemed not true.

It is further widely recognized that the best use for a reference standard is when it is constructed in exact context to its problem domain, including the form and transmission methods of the message. A reference standard appropriate to Twitter is likely not a good choice to analyze legal decisions, for example.

So, we can see many areas by which gold, or reference, standards may not be constructed equally:

1. They may contain false positives
2. They have variable inter-annotator agreement
3. They have variable mechanisms, most with none, for editing and updating the labels
4. They may lack sufficient inclusion of negatives
5. They may be applied to an out–of-context domain or circumstance.

Being aware of these differences and seeking hard information about them are essential considerations whenever a serious NLP or ML project is being contemplated.

Seemingly Good Statistics Can Lead to Bad Results

We may hear quite high numbers for some NLP experiments, sometimes in the mid-90% to higher range. Such numbers sound impressive, but what do they mean and what might they not be saying?

We humans have a remarkable ability to see when things are not straight, level or plumb. We have a similar ability to spot errors in long lists and orders of things. While a claimed accuracy of even, say, 95% sounds impressive, applied to a large knowledge graph such as UMBEL [8], with its 35,000 concepts, translates into 1,750 misassignments. That sounds like a lot, and it is. Yet misassignments of some nature occur within any standard. When they occur, they are sometimes glaringly obvious, like being out of plumb. It is actually pretty easy to find most errors in most systems.

Still, for the sake of argument, let's accept we have applied a method that has a claimed accuracy of 95%. But, remember, this is a measure applied against the gold standard. If we take the high-end of the inter-annotator agreements for domain standards noted above, namely 80%, then we have this overall accuracy within the system:

.80 x .95 = 0.76

Whoa! Now, using this expanded perspective, for a candidate knowledge graph the size of UMBEL — that is, about 35 K items — we could see as many as 8,400 misassignments. Those numbers now sound really huge, and they are. They are unacceptable.

A couple of crucial implications result from this simple analysis. First, it is essential to take a holistic view of the error sources across the analysis path, including and most especially the reference standards. (They are, more often than not IMO, the weak link in the analysis path.) And, second, getting the accuracy of reference standards as high as possible is crucial to training the best learners for the domain problem. We discuss this second implication next.

How to Get the Standards High

There is a reason the biggest Web players are in the forefront of artificial intelligence and machine learning. They have the resources — and most critically the data — to create effective learners. But, short of the biggest of Big Data, how can smaller players compete in the NLP and machine learning front?

Today, we have high-quality (but with many inaccuracies) public data sets ranging from millions of entity types and concepts in all languages with Wikipedia data, and a complementary set of nearly 20 million entities in Wikidata, not to mention thousands more of high-quality public datasets. For a given enterprise need, if this information can be coherently organized, structured to the maximum extent, and subject to logic and consistency tests for typing, relationships, and attributes, we have the basis to train learners with standards of unprecedented accuracy. (Of course, proprietary concepts and entity data should also figure prominently into this mix.) Indeed, this is the premise behind Structured Dynamics' efforts in knowledge-based artificial intelligence.

KBAI is based on a curated knowledge base eating its own tail, working through cycles of consistency and logic testing to reduce misassignments, while continually seeking to expand structure and coverage. There is a network effect to these efforts, as adding and testing structure or mapping to new structures and datasets continually gets easier. These efforts enable the knowledge structure to be effectively partitioned for training specific recognizers, classifiers and learners, while also providing a logical reference structure for adding new domain and public data and structure.

This basic structure — importantly supplemented by the domain concepts and entities relevant to the customer at hand — is then used to create reference structures for training the target recognizers, classifiers and learners. The process of testing and adding structure identifies previously hidden inconsistencies. As corrected, the overall accuracy of the knowledge structure to act in a reference mode increases. At Structured Dynamics, we began this process years ago with the initial UMBEL reference concept structure. To that we have mapped and integrated a host of public data systems, including OpenCyc, Wikipedia, DBpedia, and, now, Wikidata. Each iteration broadens our scope and reduces errors, leading to a constantly more efficient basis for KBAI.

An integral part of that effort is to create gold standards for each project we engage. You see, every engagement has its own scope and wrinkles. Besides domain data and contexts, there are always specific business needs and circumstances that need to be applied to the problem at hand. The domain coverage inevitably requires new entity or relation recognizers, or the mapping of new datasets. The nature of the content at hand may range from tweets to ads to Web pages or portions or academic papers, with specific tests and recognizers from copyrights to section headings informing new learners. Every engagement requires its own reference standards. Being able to create these efficiently and with a high degree of accuracy is a competitive differentiator.

SD's General Approach to Enterprise Standards

Though Structured Dynamics' efforts are geared to enterprise projects, and not academic papers, the best practices of scientific research still apply. We insist upon the creation of gold standards for every discrete recognizer, classifier or learner we undertake for major clients. This requirement is not a hard argument to make, since we have systems in place to create initial standards and can quantify the benefits from the applied standards. Since major engagements often involve the incorporation of new data and structure, new feature recognizers, or bespoke forms of target content, the gold standards give us the basis for testing all wrinkles and parameters. The cost advantages of testing alternatives efficiently is demonstrable. On average, we can create a new reference standard in 10-20 labor hours (each for us and the client).

Specifics may vary, but we typically seek about 500 true positive instances per standard, with 20 or so true negatives. (As a note, there are more than 1,900 entity and relation types in Wikidata — 800 and 1,100 types, respectively — that meet this threshold. However, it is also not difficult to add hundreds of new instances from alternative sources.) All runs are calibrated with statistics reporting. In fact, any of our analytic runs may invoke the testing statistics, which are typically presented like this for each run:

```
True positives:  362
False positives:  85
True negatives:  19
False negatives:  45
```

```
+--------------+------------+
| key          | value      |
+--------------+------------+
| :precision   | 0.8098434  |
| :recall      | 0.8894349  |
| :specificity | 0.1826923  |
| :accuracy    | 0.7455969  |
| :f1          | 0.84777516 |
| :f2          | 0.8722892  |
| :f0.5        | 0.82460135 |
```

When we are in active testing mode we are able to iterate parameters and configurations quickly, and discover thrusts that have more or less effect on desired outcomes. We embed these runs in electronic notebooks using literate programming to capture and document our decisions and approach as we go [9]. Overall, the process has proven (and improved!) to be highly effective.

We could conceivably lower the requirement for 500 true positive instances as we see the underlying standards improve. However, since getting this *de minimus* of examples has become systematized, we really have not had reason for testing and validating smaller standard sizes. We are also not seeking definitive statistical test values but a framework for evaluating different parameters and methods. In most cases, we have seen our refererence sets grow over time as new wrinkles and perspectives emerge that require testing.

In all cases, the most important factor in this process has been to engage customers in manual review and scoring. More often than not we see client analysts understand and detect patterns that then inform improved methods. Both us, as the contractor, and the client gain a stake and an understanding of the importance of reference standards.

Clean, vetted gold standards and training sets are thus a critical component to improving our client's results — and our own knowledge bases — going forward. The very practice of creating gold standards and training sets needs to receive as much attention as algorithm development because, without it, we are optimizing algorithms to fuzzy objectives.

Acknowledgements

This article was originally posted on the *AI3::Adaptive Information* Web site at http://www.mkbergman.com/1964/gold-standards-in-enterprise-knowledge-projects/. This version has been edited and reformatted slightly for PDF distribution. We thank Cognonto Corporation for making this content freely available.

---

[1] M.K. Bergman, 2015. "A Primer on Knowledge Statistics," *AI3:::Adaptive Information* blog, May 18, 2015.

[2] By bringing in some other rather simple metrics, it is also possible to expand beyond this statistical base to cover such measures as information entropy, statistical inference, pointwise mutual information, variation of information, uncertainty coefficients, information gain, AUCs and ROCs. But we'll leave discussion of some of those options until another day.

[3] George Hripcsak and Adam S. Rothschild, 2005. "Agreement, the F-measure, and Reliability in Information Retrieval." *Journal of the American Medical Informatics Association* 12, no. 3 (2005): 296-298.

[4] See Eleni Miltsakaki, Rashmi Prasad, Aravind K. Joshi, and Bonnie L. Webber, 2004. "The Penn Discourse Treebank," in *LREC*. 2004. For additional useful statistics and an example of high inter-annotator agreement, see Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel, 2006. "OntoNotes: the 90% Solution," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pp. 57-60. Association for Computational Linguistics, 2006.

[5] *Inter-annotator agreement* is the degree of agreement among raters or annotators of scoring or labeling for reference standards. The phrase embraces or overlaps a number of other terms for multiple-judge systems, such as *inter-rater agreement*, *inter-observer agreement*, or *inter-rater reliability*. See also Ron Artstein and Massimo Poesio, 2008. "Inter-coder Agreement for Computational Linguistics," *Computational Linguistics* 34, no. 4 (2008): 555-596. Also see Kevin A. Hallgren, 2012. "Computing Inter-rater Reliability for Observational Data: An Overview and Tutorial," *Tutorials in Quantitative Methods for Psychology* 8, no. 1 (2012): 23.

[6] Philip V. Ogren, Guergana K. Savova, and Christopher G. Chute, 2007. "Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition," in *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, p. 2325. IOS Press, 2007. This study shows inter-annotator agreement of .75 for biomedical notes.

[7] Vaselin Stoyanov and Claire Cardie, 2008. "Topic identification for fine-grained opinion analysis." In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 817-824. Association for Computational Linguistics, 2008. shows inter-annotator agreement of ~76% for fine-grained topics. David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin 2010. "Automatic evaluation of topic coherence." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100-108. Association for Computational Linguistics, 2010, shows inter-annotator agreement in the .73 to .82 range.

[8] UMBEL (Upper Mapping and Binding Exchange Layer) is a logically organized knowledge graph of about 35,000 concepts and entity types that can be used in information science for relating information from disparate sources to one another. This open-source ontology was originally developed by Structured Dynamics, which still maintains it. It is used to assist data interoperability and the mapping of disparate datasets.

[9] Fred Giasson, Structured Dynamics' CTO, has been writing a series of blog posts on literate programming and the use of Org-mode as an electronic notebook. I have provided a broader overview of SD's efforts in this area; see M.K. Bergman, 2016. "Literate Programming for an Open World," *AI3:::Adaptive Information* blog, June 27, 2016.