Explicit and Implicit Requirements of Technology Evaluations: Implications for Test Data Creation

Lauren Friedman, Stephanie Strassel, Meghan Lammie Glenn
Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19103
{lf, strassel, mlglenn}@ldc.upenn.edu

Abstract

A multitude of approaches, methodologies and metrics exist for evaluating the performance of technologies like machine translation, speech recognition and information extraction. While metrics vary widely in their assumptions about what is being tested and how it should be measured, most technology evaluations rely crucially on a carefully constructed test data set that is both accurate and fully expressive of the phenomena being evaluated. Within this context, this paper explores some of the challenges of creating reference data for technology evaluations, highlighting many of the decisions and judgments that must be made with regard to data selection, difficulty, annotation, and quality. We discuss not only the fully articulated expectations for test data, but also the hidden assumptions and implicit requirements that affect test set creation. We use the GALE Machine Translation task as a case study in discussing these issues, occasionally drawing examples from other evaluations to illustrate various aspects of the problem.

1. Introduction

A multitude of approaches, methodologies and metrics exist for evaluating the performance of technologies like machine translation, speech recognition and information extraction. While metrics vary widely in their assumptions about what is being tested and how it should be measured, most technology evaluations rely crucially on a carefully constructed test data set. While some metrics require post-hoc manual assessment of system performance, even automatic metrics like BLEU and METEOR assume the existence of one or more gold standard references against which system performance can be compared. Different metrics vary in their requirements about the completeness of the reference data or the extent to which multiple “right answers” can exist, but nearly all assume that the reference data is both accurate and fully expressive of the phenomena being evaluated.

Within this context, this paper explores some of the challenges of creating reference data for technology evaluations. We use the GALE Machine Translation task as a case study in discussing these issues, occasionally drawing examples from other evaluations to illustrate various aspects of the problem.

On the surface, creation of test data for a task like machine translation is straightforward: take the set of evaluation documents and manually translate them. But like any task involving human judgment, “translation” is not a monolithic task and there are multiple decision points along the way. In the sections that follow, we discuss several of these decision points, considering not only the fully articulated requirements for test data – the type stated in an evaluation plan – but also hidden assumptions and implicit requirements that are equally important in constructing appropriate data for evaluation.

2. Data Selection

First, we consider the question: what data is appropriate for inclusion in the test set? From the perspective of a system developer, a good test set is one whose profile is reasonably similar to that of available training and devtest data. Project sponsors and customers, on the other hand, may expect systems to handle previously unseen challenges.

The ability of data creators to balance these two opposing requests is limited by the pre-determined collection epoch for each evaluation. Irrespective of stakeholders’ expectations, the profile of the final test set will be dictated at least partially by the pool of available data.. Some features of the evaluation set – its topic coverage, for example – will be necessarily distinct from what is found in training and devtest data. Thus the specification of a test epoch can automatically add novel challenges to the evaluation. Challenges introduced by the epoch constraint are features of the available data pool and outside of the control of data creators. While a narrowly defined evaluation epoch can increase difficulty, it also limits the range, scope, and variability possible within a test set.

Data creators are often in the difficult position of balancing these conflicting requirements and limitations when selecting data for inclusion in the test set. To make things still more challenging, the “profile” of any given set of data is highly multidimensional, including such components as language, dialect, genre, source, structure, topic, time epoch, document length, segment length, lexical variation, difficulty, etc. While some of these components (summarized in Table 1) are clear cut and unambiguous (e.g. document length), others are less well-defined.

	Unambiguously Specified in Typical Eval Plan?	**Directly Measurable / Testable During Eval Set Creation?**
Language	m	y
Dialect	n	m
Genre	m	m
Source	y	m
Topic	n	m
Epoch	y	y
Document Structure	n	m
Source Data Format	m	y
Encoding	y	y
Doc Length	y	y
Segment Length	y	y
Lexical Variety	n	m
Linguistic/Structural Complexity (e.g. syntax)	n	n
Overall Difficulty	n	n

Table 1: A subset of data features.

For the NIST Open Machine Translation Evaluation, for example, the evaluation plan developed by NIST included clear direction on goals, training conditions, test data, file formats, and performance metrics (NIST, 2008). Such a detailed evaluation plan is valuable not only for participating sites, but for data creators as well.

The sheer number of data variables and types, however, makes it impossible to fully account for the effect of individual components and the various interactions among them. Data creators endeavor to build a test set according to specifications described in an evaluation plan. But the "ideal" balance of components remains elusive since the impact of certain factors is not yet known – and in some cases cannot be fully known – and the various components are often non-orthogonal.

While all efforts are made to meet any explicit expectations, blindly following only the expectations specified in an evaluation plan does a disservice to the program. Without understanding finer points about the data itself, the goals of the evaluation, and the design of the evaluation metrics, data creators might make choices during test set construction that have unintended consequences. Having detailed expectations stated explicitly in an evaluation plan is essential, but it's not enough. Since decisions on subtler points of the data will always be necessary, data creators must have well-rounded knowledge of all aspects of an evaluation.

For instance, in a typical translation task we assume that the source and target languages are constant between the training and test data partitions. Confirming the language of a given set of documents seems trivial, but there can be hidden challenges. For example, in the case of Arabic, some informal genres like weblogs may show a substantial amount of colloquial Arabic mixed with Modern Standard Arabic. The amount of dialect mixture and the particular dialects represented can vary widely from one source to the next, from one individual document to the next, and even within a single document.

A test set unwittingly selected from dialect-heavy documents, sources or genres may be significantly more challenging than the training data.

3. Test Set Difficulty

The question of test set difficulty is particularly important for evaluations that include "go/no-go" performance targets, such as the DARPA GALE program, since the program's continuation depends in part on the ability of translation systems to meet these pre-defined targets. Fair and accurate quantification of performance and measurement of progress require a test set whose make-up is carefully controlled and fully intentional. In GALE, unsurprisingly, considerable effort is devoted to selecting an annual test set whose difficulty is closely matched to the previous year's test set. The selection process begins with human annotators reviewing a pool of candidate documents, making judgments about language, dialect, genre and topic category; annotators also give a preliminary document difficulty rating on the Interagency Language Roundtable (ILR) scale (Clifford et al, 2004). The selection may be further refined by a series of automatic diagnostics to calculate log-perplexity and tri-gram hit-rate for documents in the candidate pool, in order to identify those that are outliers when compared to the rest of the selection pool and/or previous MT evaluation sets. TER (translation edit rate [Przybocki, Sanders, & Le, 2006]) may also be calculated for translated candidate documents as another measure of test set difficulty.

This approach – with several stages of data analysis and filtering – ensures that as many components as possible are known factors when building the final test set. However, even with all data features available to aide the selection process, the measure of "difficulty" is by no means straightforward. MT systems have different weak points and will find different areas of the data especially challenging. Assessing disparate data components when constructing a test set is important in order to balance test difficulty for all evaluation participants, but also to provide evaluation coordinators and sponsors with a reliable metric for gauging actual performance improvements over time.

The continued growth of multi-year programs, such as GALE, is somewhat constrained by the need for consistent test data; since the performance targets are set from the beginning, the difficulty of test sets for all phases must match that of the first in order to reliably measure progress. For example, if Phase 1 data is found to be too difficult, that inflated level of difficulty will be preserved for the duration of the program; otherwise, any conclusions drawn from trends in performance over subsequent phases will be untenable. Although the data selection process for GALE has become lengthier and more complex with each year's evaluation, there will always be unknowns, and matching difficulty from one phase to the next remains a significant challenge.

A "progress set" offers one alternative approach to the problem of measuring improvement against a test set that is different each year. While GALE does not include a progress set, the DARPA EARS Program introduced the idea of designating a subset of evaluation data that remains blind for the duration of a program (Strassel, 2004). While this progress set introduces a new list of challenges – including the long-term sequestration of data – it does offer a fixed yardstick for the measure of progress over time. Whether the potential benefits of a progress set outweigh its added costs and complications is an open question.

4. Data Annotation and Quality

Assuming the question of test data selection has been settled, the selected data is typically annotated in some fashion – transcribed, translated, tagged for entities – to create the gold standard reference. Here too there are a multitude of challenges for the data creator in ensuring the test set is well-matched to the evaluation. The goals of the evaluation must be utterly explicit in terms of what is being measured and how; it is also important for data creators to understand the desired application for the technology being evaluated. All of this has a bearing on what the reference should consist of and how it should be created, but often these goals are only defined in the broadest of terms.

In a translation task for instance, the goal is known to be the production of "high quality" MT. But how important is fluency versus completeness or precision of meaning? Ideally, all of these features are present in a high-quality translation, but – in reality – they are often at odds. All of the possible MT goals that the FEMTI framework (King, Popescu-Belis, & Hovy, 2003) identifies require different emphases during the creation of the evaluation set. The desired use of the MT technology and the context within which it will be applied shape the priorities of the system developers and the evaluators, and these same details must also guide the data creators.

If the goal of the evaluation is to generate readable translations, the data creator might be tempted to heavily emphasize fluency when producing the reference translations. But a measure such as readability is difficult to quantify and almost entirely dependent upon the intended use of the data. A domain expert might prefer that subtleties of meaning be preserved even at the expense of fluency, while a novice reader might reverse these preferences.

The target consumer should guide these choices but is often an unknown quantity. Even when the audience is known, its needs are not always fully articulated or understood. And if the consumer of the translations is not a human at all but another downstream application (information retrieval, summarization, entity extraction, etc.), readability becomes something else entirely; both fluency and semantic accuracy could become secondary concerns if the preservation of word order, for example, is a requirement for a downstream task. Even when the technology goal itself is straightforward, a brittle evaluation paradigm with too many competing requirements will make the data creation task unmanageable.

The question of quality is central in test data creation. The term gold standard implies that the resulting resource is the best that humans can produce. But while there are several ways translation quality can be measured, there is always a subjective component. A universally accepted objective standard for human translation quality is probably untenable since, at least in the context of technology evaluations, translation quality must always be judged in terms of its intended use. The translation that will be most useful to the target consumer and the translation that will evaluate an MT system most fairly are not one and the same. Consequently, as with the question of data selection, the opinions of system developers and project sponsors do not always dovetail on what constitutes high quality data.

In addition to the lack of consensus in defining data quality, hidden assumptions can impede appropriate creation of a gold standard for any given evaluation. What kinds of humans, with what skills or training and with what kind of infrastructure, are expected to produce the gold standard? For instance, a run of the mill commercial translation will represent the work of one, or perhaps two, translators. But for many evaluation paradigms, the gold standard translation represents the collective effort of a much larger team; in the GALE program, gold standard translations require a series of manual passes by at least six individuals:

1) source-language dominant bilingual translator produces a preliminary translation emphasizing accuracy;
2) target-language dominant bilingual translator revises the translation to improve fluency;
3) source-language dominant bilingual annotator checks translation for errors and omissions;
4) source-language dominant bilingual senior annotator checks for remaining errors, improves fluency, corrects and standardizes named entities;
5) target-language dominant bilingual annotator improves fluency and adds translation variants where required;
6) target-language monolingual annotator reviews for fluency and flags questionable regions.

By any reasonable definition, the GALE gold standard translations can be said to be high quality, but the quality is in some ways artificial. The final references, as the product of a carefully constructed team, are far beyond the scope of what a single human translator could generate. Thus the MT is not scored against a human translation that could in any way be considered representative, but against a composite translation that is polished an almost unreasonable number of times.

This laborious process for gold standard creation was defined with the specific requirements of the GALE evaluation firmly in mind. The GALE evaluation metric is HTER, defined as the minimum number of edits one must make to the MT output so that it has the same meaning as the gold standard reference and is equally understandable (Przybocki, Sanders & Le, 2006). Given this metric, the gold standard references for GALE have properties that are not required for many other MT evaluations, and are not frequently found in run of the mill commercial translations. For example, when the source text's meaning is ambiguous (e.g., verb tense is not expressed in Chinese), variants are added to the gold standard translation. In a standard translation, a translator would resolve ambiguities based on context and judgment, but the GALE gold standards require that the presence of this ambiguity is carefully preserved. Similarly, idioms are translated both literally and figuratively. The final references are meant to be not only fluent and accurate, but also completely inclusive of all reasonable interpretations of the source. This approach seeks to address the "multiple correct answers" problem of translation and ensure the fairest possible evaluation of MT systems.

Another dimension of test set quality is the consistency of the reference annotation. For annotations that require multiple passes by multiple judges like the GALE gold standard translations described above, it is difficult to imagine what "consistency" would mean, or how it could be measured. With a metric like edit distance, a high level of consistency is not really possible, expected, or even desirable. The multiple passes on GALE evaluation translations, for example, actually take inconsistency as a baseline assumption; each stage of quality control is intended to produce output that differs from – and improves upon – the previous stage. The expectation of this approach is not consistency between annotators, but rather the consistency of this group as a whole. While the group may not be internally consistent, the consistency between this group and other similarly-constructed groups can be expected to be greater than the consistency between two individuals.

Other tasks are superficially more straightforward, like orthographic transcription of audio data. As part of the DARPA EARS program in 2004, LDC undertook a careful study of inter-transcriber consistency, using the RT-03 English current test set (Strassel, 2004). Each evaluation file was transcribed by two annotators working independently, and the resulting transcripts were compared using the standard scoring software developed by NIST for the program's speech-to-text evaluation (NIST, 2004). While consistency was good, it was by no means perfect: the broadcast news genre showed a word disagreement rate of 1.1%, while conversational telephone speech showed 4.3% disagreement. These numbers are quite low in absolute terms, but given go/no-go performance targets of 5-10% word error rate and better for STT systems, it is critical to establish a baseline for human "performance". For more complex tasks, consistency rates are typically lower.

Performance targets are being set higher and higher; can machine error rate be reasonably expected to drop as low as – or lower than – the rates of human variation? Or should systems only be expected to perform somewhere within the range of typical human error? The urgency of resolving this issue rises as the gap between machine performance and human consistency narrows with each evaluation campaign.

5. Conclusion

The challenges for test set creation discussed in the sections above are not unique to evaluation data; they are relevant to any linguistic resource created for a particular purpose. With evaluation data, however, the stakes are typically higher and so the pressure on data creators is more intense. This is often coupled with a shorter timeline for developing evaluation data (compared to training data), which can be quite challenging given the primary emphasis on quality and the increased importance of consistency. As a result, the overall cost for test data creation is typically many times higher than training data created for the same evaluation. For GALE MT for instance, gold standard references are roughly ten times more costly (in dollars and time) than training data references, even though the training data can also be characterized as high quality.

The process for creating training data, though the end product is certainly high quality, only minimally resembles the gold standard creation process – even within the same program. While test set creation is so intensive precisely because the stakes are so high and the margin for error is so low, the effect of the schism between these two approaches to data creation needs to be further interrogated. The protocols for evaluation data could not reasonably be applied to training data, given the high volumes required of the latter. But what is the significance, if any, of training systems on data that is constructed differently, with different quality standards, than the test data that will ultimately be used to evaluate them?

The creation of gold standard references is so resource-intensive that even scaling up or supporting multiple evaluations at once becomes an inordinate challenge. The significantly higher costs of test set creation are only justifiable if higher quality can be shown to correlate with fairer evaluation – a correlation that is nearly impossible to prove. The high cost of evaluation data creation further underscores the importance of clearly defining the goals of the evaluation, fully informing data creators of program requirements, and then closely matching the test data to these needs and goals.

6. Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

7. References

Clifford, Ray, Neil Granoien, Douglas Jones, Wade Shen, & Clifford Weinstein (2004). "The effect of text difficulty on machine translation performance: a pilot study with ILR-rated texts in Spanish, Farsi, Arabic, Russian and Korean." In: *LREC-2004: Fourth International Conference on Language Resources and Evaluation*, Proceedings, Lisbon, Portugal, 26-28 May 2004; pp.343-346.

King, M., Popescu-Belis, A. and Hovy, E. 2003. "FEMTI: creating and using a framework for MT evaluation." In:

AMTA (2003), 224-231.

National Institute for Standards and Technology (2004). NIST RT-04 Spoken Language Technology Evaluation. http://www.nist.gov/speech/tests/rt/rt2004/fall/index.htm

National Institute for Standards and Technology (2008). *The 2008 NIST Open Machine Translation Evaluation Plan 2.4.* http://www.nist.gov/speech/tests/mt/2008/doc

Przybocki, Mark, Gregor Sanders, & Audrey Le (2006). "Edit distance: a metric for machine translation evaluation." In: *LREC-2006: Fifth International Conference on Language Resources and Evaluation*. Proceedings, Genoa, Italy, 22-28 May 2006; pp.2038-2043

Strassel, Stephanie (2004). "Linguistic Resources for Effective, Affordable, Reusable Speech-to-Text." In: *LREC-2004: Fourth International Conference on Language Resources and Evaluation*. Proceedings, Lisbon, Portugal, 26-28 May 2004.