Invited Talk

THE PRIMACY OF CORE TECHNOLOGY MT EVALUATION

John S. White
PRC Inc.
1500 PRC Drive
McLean, VA 22102
white_john@prc.com

Abstract

Much of the evaluation of machine translation today is focused on the current reality that there are different types of MT, each suitable to certain uses and users, and not to others. In light of this view, black box evaluation of heterogeneous core algorithms underlying MT, as has been done in the DARPA MT Evaluation series, seems to lack value. This paper claims not only that such core technology evaluations are of value, but that they are ultimately of more inherent value than other forms of evaluation in anticipating a future when a fully automatic, high quality translation capability will be closer to reality.

One of the benefits of the renewed interest in machine translation evaluation over the last few years has been the variety of categorizations of MT uses and users. These delineations have added significant clarity to the strengths and weaknesses in modern MT approaches, and helped to focus not only on the areas that need the most work, but also on the parts of the MT process that promise the best opportunities for improvement in time, cost, and user satisfaction. It is clearly true that, as has often been expressed (e.g., Somers 1993), translation support tools will bring more benefits to users in the near term than improvement in the translation algorithms. But somewhere in the heart of hearts of the field is the belief that that is not the way it will always be. Someday, we cannot help but believe, applications that translate languages will work as well as applications that translate WordPerfect files into MS Word. Awareness of this goal, or vision (or hallucination), has led me to make the claim that the delineations of system type, system purpose, and user type, and the evaluations that purport to measure them, are ultimately of less importance than the "core technology" that will eventually lead to the achievement of the goal.

In this presentation I will address a criticism that has been raised against attempting to measure very different MT approaches using the same set of black-box methods. In so doing, I intend to discuss two reasons why methods intended to evaluate the core translation engines of MT systems are not only valid in their own right, but better in the long run than methods that measure peripheral or evolutionary characteristics.

Background. Different people involved in the translation process need to know different things about an MT system. Buyers need to know about cost and support; managers need to know about cost-effectiveness and (ideally) about user satisfaction. Developers need to know about performance, improvement, and regression against some set of benchmarks. Translators need to know about pre-/post-editing tools and lexical update. For each type of information that a person in the translation loop needs to know, there should be evaluation measures, at least some of them specific to the particular attribute the person needs.

Add to these the current (and currently accurate) claim that the different uses to which MT is to be put—publication, scanning, gisting, text extraction, to name a few —are strongly correlated with the underlying algorithms for translation. Here, the presumption has been that different MT designs (direct, transfer, statistical, knowledge-based, etc.) are somehow better for some uses of MT and worse for others. Thus a knowledge-based, human-interactive system might be presumed to have a higher quality, and be better for publication-style translation tasks, though relatively slow, and perhaps with a somewhat limited semantic/pragmatic domain coverage. On the other hand, a direct or statistical system that produces single-pass, batch output, should be much faster and therefore more amenable to rapid information retrieval and fast scanning applications. If that is so, the argument goes, there should be no point in evaluating different MT designs by the same metric, since this would fail to capture the use-specific merits of particular designs (e.g., Hovy 1994).

Meanwhile, the MT evaluation program of the Defense Advanced Research Projects Agency (DARPA), begun in 1991 and continuing today under various auspices, intends to evaluate just the 'core technologies' of MT systems, in a series of black-box methods that attempt to be oblivious to the underlying MT system design, while factoring out effects of user interface, user competence, robustness of the implementation, and even the languages translated. A goal of the DARPA program was to find the translation algorithm with the best potential for becoming the underlying engine of the fully automatic, high quality translation (FAHQT) system of the future.

The DARPA MT evaluation methods collect human subjective judgments of sufficient size and granularity to permit generalizations about the performance of the translation components of research and commercial MT systems. The methods in the evaluation suite are necessarily black-box: the DARPA series contained systems of radically different design, different underlying linguistic theory (and lack thereof), and, of course, translated different languages.

There are three measures in the DARPA suite, intended to reach different aspects of correctness of translation, using judgments of monolingual evaluators made on English outputs of the various MT systems and human control translations.

- *Adequacy.* The objective of the adequacy evaluation is to measure the extent to which all of the content of a text is conveyed, regardless of the quality of the English in the output. In this evaluation, evaluators compare the content of an expert translation against the content of a machine translation.

- *Fluency.* The fluency evaluation measures human judgments of how 'native-like' a translation is, in context of a whole text, but without direct regard to the accuracy of the translation.

- *Informativeness.* The informativeness evaluation measures a system's ability to produce a translation from which users can glean information they seek. This evaluation used a multiple choice format by which evaluators indicated the successful communication of facts in the translation (see also Church and Hovy, 1991).

Criticism of core technology evaluation. The DARPA goals, and particularly the methodology, have enjoyed significant visibility and healthy criticism over the years, much of which has helped in its evolution and increased focus on core technologies (White et al. 1994). But certain criticisms remain of core technology evaluation, which I summarize in two sets:

Core technology evaluation doesn't measure everything that need measuring. Core technology evaluation does not measure the cost of getting or using a system, nor the potential for upgrading to new languages, nor the solvency of the producer, nor the time (or keystrokes, or calories) to pre- and post-edit. And these are all things that need to be known about MT systems, at least today;

Core technology evaluation doesn't measure anything that needs measuring. There is no useful purpose served by comparing different MT algorithms. The designs underlying different types of MT engines are inherently more amenable to certain uses for machine translation, and less amenable for others. Comparing radically different systems with the same metric is like trying to compare "a bulldozer and a Rolls Royce and a dune buggy and a motorcycle" (Hovy op.cit.).

Core technology, change, and vision. I will not say that the measures we have used for core technology evaluation are exactly right yet, or that they will always be the right ones. As MT algorithms improve, the results of each of the measures will converge—more fluent translations are unlikely to have less accurate coverage of content. In fact, we have shown elsewhere in this conference that there is a certain affinity between the DARPA measures and particular uses of MT in today's end-to-end integration context (White and O'Connell 96). Whatever the methods are in the future, they should be uniformly applicable to any core algorithm, against any language pair, regardless of intended end-use.

But I will raise two major arguments in response to these criticisms about core technology evaluation in general, as follows: 1). circumstances change, for external reasons, that can make a lot of what we need to measure today irrelevant in the future; and 2). a real breakthrough in core technology will cause a change in our vision of the whole idea of MT uses, users, and types. The remainder of this discussion presents these arguments in turn.

1. *The world changes in ways that we do not anticipate.* We have learned many times, in just the last few years, that what we thought was going to be important turned out not to be. We don't actually know what users will need in the future, nor what expectations about cost, support, and so on will be. These are all factors that are outside of the control of the evolution of MT itself, but will have a profound influence on its future directions.

Less than a decade ago the idea of being able to type plain English commands to a computer seemed like the most obvious of benefits to be gained from the field of natural language understanding. It occurred even to the geequiest of us that

```
% (find -name \[Nn\]ew\.\* -user jwhite -atime +2 -exec  vi -print) >& /dev/null
```

was better represented as

% *could you get that file I was working on yesterday, I think it started with "new something"...*

Of course, today we know that these are both wrong. What we language types didn't know then was that we all actually preferred to communicate without any written commands at all. It turned out that the much simpler idea of pointing and clicking on icons was also much preferable to typing anything at all, natural language or not. So all of the work in parsing, dialogue maintenance, meta schema modeling, etc. that went into such designs was of little importance to command interface or database interface per se. But the core technology work in these areas did turn out to be of great use to speech recognition. So research, development, and evaluation of the core capability of natural language interface remained useful even though the use of it changed in an unanticipated way.

2. *The 'uses of MT' paradigm is a necessary evil, not a fundamental truth.* We are indebted to a recent paper by Steven Krauwer (Krauwer 1993) in which he characterizes certain fallacies of MT Evaluation in terms of evaluating the performance of a car by evaluating its transmission. He notes that the idea of testing something that is a piece of a larger process involves mounting that component on a test framework. The fallacy comes when you compare the results of that test with a test of the framework without the component.

The point is well taken; at this moment in history we must find a way to evaluate just the properties of the translating component, in the context of the interaction of the component with the rest of the translation process.

But we all hope that the future will be different in this regard: machine translation will work. That is, there will be FAHQT on demand as either a word processing tool or as a 'preprocessor' to information detection and extraction systems (or whatever downstream information handling systems will exist then). I may grudgingly concede that the word-processor tool may afford some (monolingual!) user interaction (a spell checker does, after all), but in general the tool just renders the input language into some other language, without human translator interaction, resulting in text that is entirely and easily readable by a target monolingual human (and of course by any systems that can manipulate such text).

If we think about this vision, we realize that a lot of assumptions change. The need to evaluate the effect of human interaction on cost, efficiency, required expertise, etc., becomes far less important, for example. And the whole idea of 'uses/types of MT' experiences a needed catharsis. If MT "worked" in accordance with this vision, we wouldn't have to distinguish between the MT uses of publishing, scanning for relevance, and rendering of messages. Any single MT algorithm that "works" will handle all of those uses just fine.

So questions about different types, different uses, and different users of MT diminish, and in some cases disappear when applied against this vision. This is because many of these distinctions are not based on the realities of the *context* of MT, as implied, but on the realities of the *limitations* of MT.

Let us return to the automotive metaphor, augmenting it a bit to speak of the core algorithm of an MT system as the engine of the MT process.

We may observe a couple of things about this metaphor. First, when we have a better engine, fewer of us, in fewer walks of life, care about evaluating it, diagnosing it, knowing how to fix it, etc. All of us who roamed the earth in the late Sixties not only knew about our own Volkswagen engines, but could tell what VW engines our friends had by the subtle differences in bumper shape, curve of the windshield, etc. —things that had nothing to do with engines but could help us diagnose each other's engine symptoms. We all had the same third-party repair book and had all fashioned our own static timers from flashlight bulbs. We hesitated to commit to long trips, and avoided high speeds, and didn't go anywhere without our set of special purpose, handmade tools, because of the likelihood that our engines would fail on the road.

Today, I don't even know for sure what kind of engine my car has, and would not be in the least tempted to diagnose or fix it if something ever did go wrong. This difference may be because I realized that my knowledge of engines never really helped that much; but it is more likely the case that today's engines simply work better than the old VW ones did. So the questions of evaluating engine performance against the intended uses of my car have become irrelevant. The engine just works, and it will work for everything a car is supposed to do, without carrying around special purpose tools that only I know how to use. In the same way, a translation algorithm that just "works" will work for everything an MT system needs to do.

Now this appears to run into our old "bulldozer-Rolls Royce-dune buggy-motorcycle" metaphor. After all, all of these vehicles do different things, and all of them probably have different sorts of engines in them. So shouldn't different end-use requirements necessitate different criteria for core algorithms? But the observation about vehicle differences is not the same one as the current argument about different translation algorithms for different translation purposes. The engines in bulldozers, Rolls Royces, dune buggies, and motorcycles all *work;* it is equally unacceptable for a bulldozer engine to chronically misfire as it is for a car. In either case the engine should never fail to run, move gasoline and water around, turn the crankshaft, etc. By contrast, the engines in today's MT systems only work part of the time, and their best use is predicated on trying to make the most of their shortcomings, not on the intrinsic differences in the end-use tasks themselves.

Certainly, we may someday realize that one type of correctly working MT engine is for some reason more suited to some purposes, and some other correctly working MT engine more suited others. Cost, for instance, will be eternally relevant. But we are not close enough to that bridge to be concerned with crossing it yet. And until we are, we need to realize that the necessary evils that have us dividing the universe of machine translation for today's purposes but not constrain our understanding that things will change, and some of the things will change for the better.

Today, we must continue to bear in mind the many different things we need to know about an MT system, and evaluate them with the right measures. And we must also bear in mind the types/uses/users model in developing evaluation methodologies. But most of the long list of factors are ephemeral, as we have discussed, for reasons both internal and external to MT itself.

Conclusion. Not only do the types/uses/users criticisms of core technology evaluation evaporate in the context of a vision of a working MT algorithm, but we can make the stronger claim that core technology evaluation actually facilitates the eventual accomplishment of the vision, in ways that other types of evaluation do not. The goal of DARPA, and subsequent uses of the methodology, has been to help determine which engine design shows the best promise of actually working. Black-box comparison of the different designs is entirely appropriate for making such decisions without regard for theoretical or software engineering preferences. Core technology evaluation transcends both the external conditions of today and the internal limitations of today's systems to provide a basis for selecting the best candidates for the future accomplishment of the vision all of us, even in our most cynical and skeptical moments, hope for.

References

Somers, Harold. 1993. Current Research in Machine Translation. *Machine Translation 7.*

Hovy, Eduard. 1994. Discussion on "Apples, Oranges, or Kiwis? Criteria for the comparison of MT Systems". Panel Discussion, Muriel Vasconcellos, moderator. In M. Vasconcellos (ed.) *MT Evaluation: Basis for Future Directions.* National Science Foundation.

White, J.S. and T.A. O'Connell. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas.*

Church, K.W. and E.H. Hovy. 1991. Good Applications for Crummy Machine Translation. In J.G. Neal and S.M. Walter (eds.) *Proceedings of the 1991 Natural Language Processing Systems Evaluation Workshop.* Rome Laboratory Final Technical Report RL-TR-91-362.

White, J.S. and T.A. O'Connell. 1996. Adaptation of the DARPA Machine Translation Evaluation Paradigm to End-to-End Systems. *Proceedings of AMTA-96.*

Krauwer, S. 1993. Evaluation of MT Systems: A Programmatic View. *Machine Translation 8.*