Predicting Intelligibility from Fidelity in MT Evaluation

John White

Litton PRC
1500 PRC Drive
McLean VA 22102
white_john@prc.com

Abstract

Attempts to formulate methods of automatically evaluating machine translation (MT) have generally looked at some attrinbute of translation and then tried, explicitly or implicitly, to extrapolate the measurement to cover a broader class of attributes. In particular, some studies have focused on measuring fidelity of translation, and inferring intelligibility from that, and others have taken the opposite approach. In this paper we examine the more fundamental question of whether, and to what extent, the one attribute can be predicted by the other. As a starting point we use the 1994 DARPA MT corpus, which has measures for both attributes, and perform a simple comparison of the behavior of each. Two hypotheses about a predictable inference between fidelity and intelligibility are compared with the comparative behavior across all language pairs and all documents in the corpus.

Keywords

Evaluation; fidelity, intelligibility, DARPA corpus

The issues associated with automating MT evaluation are well known, both in terms of the need for having such a capability and the difficulties inherent in creating it. Several new studies into the possibility have emerged very recently, which attempt to capture an automatically measurable phenomenon associated with translation and extrapolate to all of the MT attributes that need to measured for particular tasks/stakeholders.

Each of these proposed methods for automatic evaluation appeal to one of two classic attributes of translation: fidelity (conveyance of the information in the source expression into the target expression) and intelligibility (how understandable the target expression is to a target-native speaker). Some of these approaches appeal to intelligibility by comparing MT output to models of expected English co-occurrences (e.g., Jones and Rusk, 2000; Corston-Oliver, 2001). Other approaches appeal to fidelity by, for example, determining whether the named entities in the source are correctly represented as named entities in the target (Hirschman et al., 2000).

These approaches show promise for capturing precise rapid measurements of the attributes they directly measure. However, the assumption that the findings can be extrapolated to other MT attributes (specifically, fidelity to intelligibility or vice versa) is based on a relationship between the two which is not yet demonstrated.

This paper investigates the possibility of that there is a sufficient correlation between fidelity and intelligibility that it may be eventually feasible to predict the value of one by (automatically) measuring the other. We look at some simple mapping of fidelity scores against intelligibility scores for the 1994 DARPA corpus, and find a first step toward making the association.

Fidelity and Intelligibility

There is no overt reason to suppose that there is ever a correlation between the two. An ersatz system that simply output President Bush's inauguration address regardless of the input would measure quite high in intelligibility but usually quite low in fidelity. An algorithm that simply listed out the place- and person names from a text, untranslated, would be perhaps optimally faithful, but far less intelligible than a translation.

However, there are at least two points where fidelity and intelligibility converge. As noted elsewhere (White 2000), an imaginary MT system that only output random dots, for example, is both maximally unintelligible and maximally unfaithful. At the other extreme, a text written in the target language in the first place is as faithful as it can be (not regarding the actual truth of the assertions in the documents), and at least within the range of intelligibility sufficient for any target-native speaker to recognize that it is a set of expressions of the target language. As illustrated in Figure 1, there is some divergence between fidelity and intelligibility in between the extremes, i.e., in the range of quality in which MT lies. The question that remains, and which is the subject of this paper, is how far fidelity and intelligibility diverge over a continuum of translation quality. If this divergence can be determined, then it will be possible to predict the fidelity of an MT output by measuring its intelligibility,

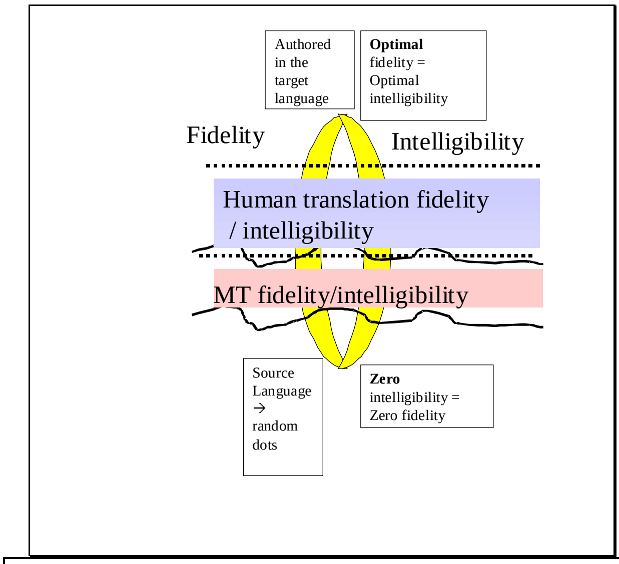
Authored in the target language
Optimal fidelity = Optimal intelligibility
Fidelity
Intelligibility
Human translation fidelity / intelligibility
MT fidelity/intelligibility
Source Language → random dots
Zero intelligibility = Zero fidelity

Figure 1: convergence of fidelity and intelligibility at the extremes, undetermined in between.

and/or vice versa.

DARPA MT Evaluation Measures

The Defense Advanced Research Projects Agency (DARPA), as part of its Human Language Technologies Program, undertook a series of evaluations of prototype, commercial, and operational MT systems (White 1995; Doyon et al. 1998). The last and most comprehensive of these, in 1994, resulted in a sizable body of parallel texts, in the source language (French, Spanish, or Japanese), expert human translations (two for each text) and raw MT outputs from several systems in each language pair.

Three measures were taken of the translations, each by 100 monolingual, English-speaking evaluators.

Adequacy, in which evaluators were given texts arranged with an expert translation on one column, MT output (or control) on another column, and a space for scoring, on a 1-5 anchored scale. The evaluators determined the extent to which meaning conveyed in a segmented portion of the expert translation (generally sub-sentence) was conveyed in the MT output text.

Fluency, in which evaluators looked at output texts and scored on an anchored 1-5 scale each sentence, on the extent to which the sentence was intuitively acceptable to a native speaker, was well formed, grammatically correct, and makes sense in the context of the overall text.

Informativeness, in which the evaluators read an output or control, and then answer multiple choice questions, like a reading comprehension test, but crafted to test the text rather than the reader.

The DARPA corpus has value for determining the possible relationship of intelligibility and fidelity, because each of 1800 translations has a score for all three measures, appropriately controlled against human factor biases. So the potential exists for analyzing the behavior of any one measure against either of the others; in this paper we look at the comparison of adequacy (an apparent fidelity measure) and fluency (an apparent intelligibility measure).

Fidelity and Intelligibility from the DARPA Corpus

As part of the 1994 evaluation, analysis-of-variance and Pearson product-moment processes were performed on the results. Among other things, these analyses indicated a correlation between fluency and adequacy scores. However, these do not differentiate correlations that might be observed in overall poor, translations, overall good ones, and the vast set of translations in between.

We posit two hypotheses for the establishment of a predictable correlation between fidelity and intelligibility:

- Fluency (and therefore intelligibility) increases in a near linear fashion with adequacy (fidelity), and thus the value of one is readily predictable from the other at all points on a quality continuum; or
- Fluency and adequacy converge at the extremes, and are much less correlated in between, but the algorithm is discoverable by which one measure can predict the other along the continuum.

Figure 2a is a line graph showing the relationship of fluency scores to adequacy for all translations (i.e., all translations in all language pairs from all systems). It is not entirely evident from visual inspection of the chart that either hypothesis is supported. However, some observations can be made:

- the low values for adequacy appear to converge with the low values for fluency, and the high appear to converge with high values
- visually, the apparent mean of fluency seems to rise with adequacy.

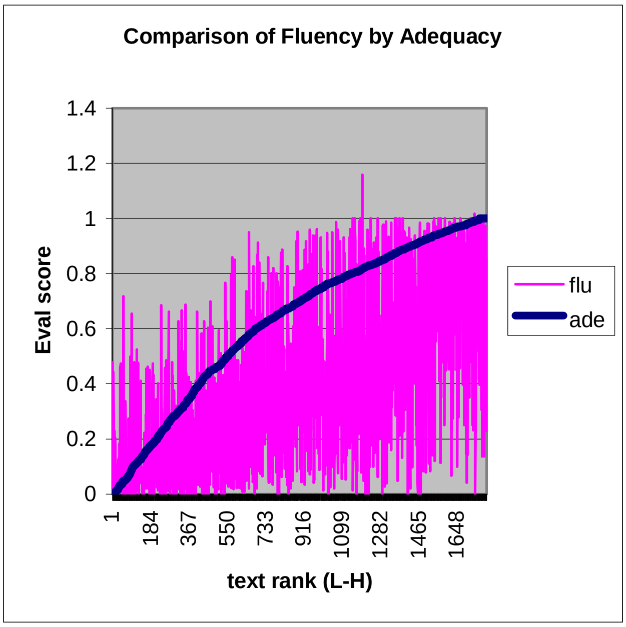

Figure 2a. Fluency scores compared to adequacy curve. All pairs, all texts.

Figure 2b simplifies the picture by taking the means in clusters of 10 from lowest to highest. Here we see the general upward trend of the two measures, but less of the convergence effect. Figure 3a and 3b show the relationship from the vantage of adequacy scores to fluency, all texts and 10-text cluster averages, respectively.

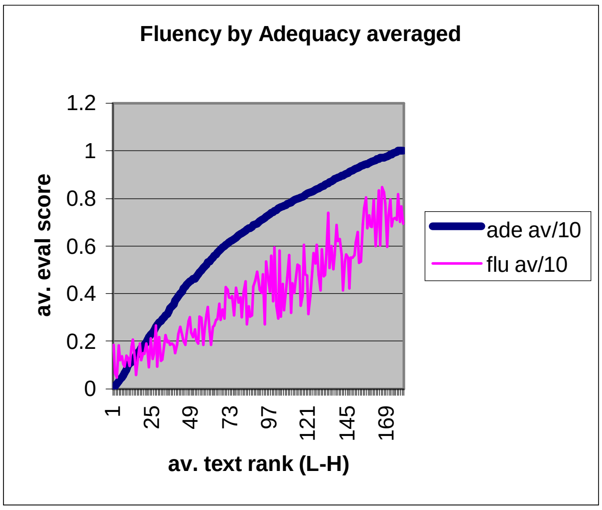

Figure 2b. means of groups of ten: fluency compared to adequacy curve

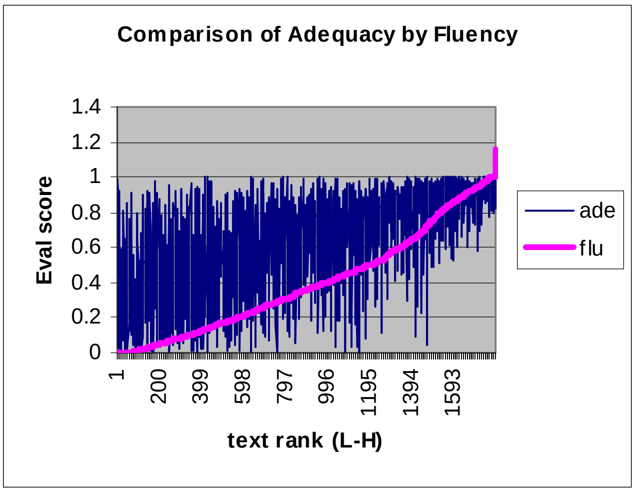

Figure 3a. Adequacy compared to fluency curve

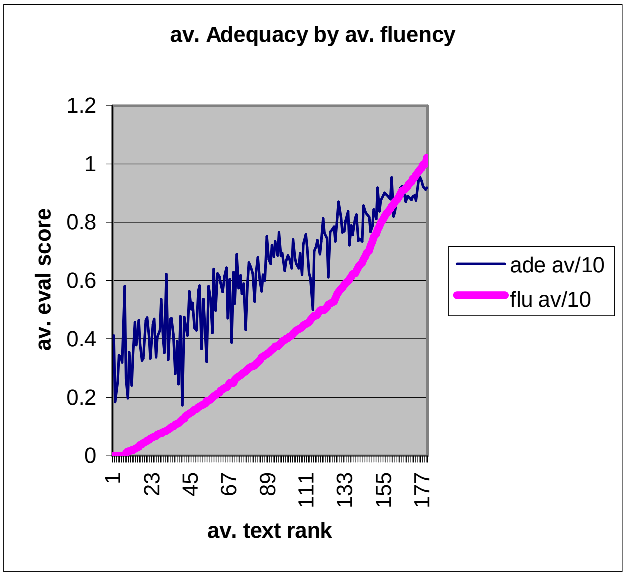

Figure 3b. means: adequacy compared to fluency curve.

The most salient observation is that where adequacy is low, fluency is low; and where fluency is high, adequacy is high. The implications are unidirectional (it can't be said that fluency is always high when adequacy is), but appear to lend support to the second hypothesis: there is a convergence at the extremes. We do not know if there is some theoretical extension of each range that allows the bi-directional claim to be made. For instance, if the corpus contained texts that were even less fluent than the ones it has, we might expect the adequacy scores to settle to zero at some point to the hypothetical left of the chart in 3b, and texts that were even more informative than the best ones in the corpus might ultimately converge with fluency to the right of the chart in 2b.

It may be worth speculating that the inverted mirror-image impression reflects a fundamental nature of fidelity and intelligibility, given a similar task (both the adequacy and fluency measures had the rater express values on a 1-5 scale after examining a section of the text). For instance, it may be that it was always at least somewhat possible to glean some information out of at least some portions of quite unintelligible text, hence the lack of convergence at zero in 3b. Similarly, it may be that even texts superbly adequate for registering information will not always be judged by all people at all times to be the best possible way to express something. And so, as 2b. suggests, fluency does not completely converge at the high end with adequacy.

Discovering an algorithmic relationship

Returning to the hypotheses, it appears that we cannot tell, with this data organized in this way, whether there is a predictor from intelligibility to fidelity because of a linear relationahip, or because of some divergence between the extremes that can be characterized and formulated, or whether there is no predictable relationship at all in between the extremes. The data from the DARPA measures appear to support both the linear and diverging hypotheses, in that each measure rises with the other, though with wild variation along the way. From this, considerable hope remains that it may be possible to decompose phenomena associated with the measures to isolate what aspects of fidelity may be predictable from intlligibility, and vice versa. Ultimately, it will be possible to evaluate MT automatically, using a handful of easily captured behaviors, enabling researchers, deveopers, and users to determine immediately the status and potential for MT approaches and systems.

.

Bibliographical References

Corston-Oliver, S., Gamon, M., and Brockett, C. (2001). A Machine Learning Approach to the Automatic Evaluation of Machine Translation. Proceedings of the 39th Annual Conference of the Association of Computational Linguistics. Toulouse, France.

Doyon, J., Taylor, K., & White, J. 1998. The DARPA Machine Translation Evaluation Methodology: Past and Present. *Proceedings of AMTA-98*. Philadelphia, PA.

Hirschman, L., Reeder, F., Burger, J., & Miller, K. 2000. Name Translation as a Machine Translation Evaluation Task. *Proceedings of the Workshop on Machine Translation Evaluation, LREC-2000.*

White, J. (1995). Approaches to Black-Box Machine Translation Evaluation. Proceedings of MT Summit 1995. Luxembourg.

White, J. (2000). Toward an Automated, Task-Based MT Evaluation Strategy. Proceedings of the Workshop on Evaluation, Language Resources and Evaluation Conference, LREC-2000. Athens, Greece.