[*MT evaluation: basis for future directions*. Proceedings of a workshop...2-3 November 1992, San Diego]

NEW DIRECTIONS: Automatic Evaluation of Translation Quality: Outline of Methodology and Report on Pilot Experiment

Henry S. Thompson
Human Communication Research Centre
University of Edinburgh

1. Introduction

The original motivation for the work reported here is the desire to improve the situation with respect to evaluation of the performance of computer systems which produce natural language text. At the moment there are few if any concrete proposals for appropriate metrics or methodologies. The domain chosen to explore a possible solution to this problem was that of machine translation, as it offered both the most obvious source of relevant material and the most pressing need for such evaluation.

I start from the premise that fast, accurate, automatic evaluation methods are of vital importance in the development process for any large-scale natural language processing application. Historically there has been little emphasis on evaluation in the machine translation community, and although that is now starting to change, the methods proposed are not automatic, thus not fast, nor in most cases is there any obvious way to test their accuracy—that is to say, the statistical significance of their results.

2. A New Methodology

Most evaluation amounts to measurement against a standard. Direct evaluation of the quality of translation has historically been achieved by human experts by comparing the candidate translation against their expectations, possibly with an eye on a 'standard' translation or a set of guidelines. Starting with the ALPAC Report, and very occasionally thereafter, some efforts at statistical processing have been included in this process, with several human evaluators marking candidate translations on three-, five-, nine-point, etc., scales of fidelity, intelligibility, and so on. I know of no attempt to automate this process, with the possible exception of work done in Beijing (Dong & Shiwen 1991), presumably because any such effort would have involved comparison with a standard, but the range of acceptable translations is usually so large that this obviously would not work.

To overcome this problem, the new methodology takes the simple approach of using multiple standards. That is, instead of comparing the candidate translation against a single standard, it compares against a *set* of standards. Furthermore, the methodology is such that the *effective* size of the standard is much greater than its actual size.

Comparison is in terms of simple string-to-string distance between clauses, measured by well-known dynamic programming techniques with respect to an inventory of primitive operations—e.g., deletion, insertion, and substitution. This is, of course, far too crude a measure, but the use of a standard set rather than a single standard compensates somewhat for this crudeness.

For the time being, the method operates at a paragraph level, although alternatives could be imagined. Several alternative approaches within the broad area of comparison with a standard set are possible. Those I have begun to explore are described below, together, where appropriate, with results from a pilot experiment in which a standard set of 44 English translations of three paragraphs drawn from two French texts were used.

2.1 The Simple Method

Each of two versions of this method starts by constructing a triangular submatrix of distances, with one entry for each pair drawn from the set composed of all the standards and candidates. Each such distance is simply the normalized distance between the optimal alignment of clauses[1] between the two texts. That is to say, if, for example, one text consists of clauses a, b, c, d, and e, and the other of u, v, w, x, y, and z, then once again we use dynamic programming to find that alignment of clauses—say, a + b with u + v + w, c with x and d + e with y + z—such that the sum (or other appropriate monotonic function) of the distances between the three pairs of strings is a minimum over all possible alignments.

On one version, the minimum or average of the distance from a candidate to the members of the standard set is taken as its score. In the pilot experiment, the difference between these two was not significant, both correlating around .55 with a human scoring of the first paragraph and .2 with the human scoring of the second.[2]

For the other version, the entire matrix was processed by a Multidimensional Scaling package (MDS(X) by Coxon et al.) to explore the dimensionality of the variation in distances. Such an approach attempts to assign coordinates in, for example, 3-space to each translation so that the order (nonmetric scaling) or actual value (metric scaling) of the intertranslation distances from the matrix are respected. Preliminary results suggest that for a reasonably accurate model (stress d-hat< .15) four dimensions are required, whether metric or nonmetric scaling is used. This in itself does not give a measure for an individual translation. Two approaches to this are possible but have yet to be explored: either using the contribution to the stress allocated to all the distances involving the candidate in the standard decomposition, or else comparing the overall stress with and without the candidate's row for a given dimensionality.

2.2 The Compound Method

Even with 44 translations of quite short paragraphs (between 20 and 70 words in length), it was noteworthy that no two translations were identical. But at the clause level, some identities, and many very near identities, were observed. If the standard set were treated not as 44 paragraphs but rather as 44 times six clauses, we can take advantage of this and effectively increase the size of the set many-fold by allowing a candidate translation to match against a compound or synthetic target composed of clauses from *different* members of the original set.

If we treat the complete set of clauses from the standard set as available for matching against each clause (or pair of clauses, etc.) or the candidate, we run the risk, especially in a large paragraph, of using the same clause twice, or using clauses in manifestly illegitimate order. But given that the members of the standard set are not themselves aligned with one another, except indirectly, it would not be trivial to enforce a strict sequentiality constraint. Rather than attempt this, the algorithm used for the pilot simply enforces that the clauses chosen must be strictly increasing by midpoint, percentage-wise.

The correlation of this compound measure, again taking each of the 44 texts in turn as the candidate and measuring it with the remaining 43 as the basis for the compound standard, was significantly better than the simple methods described above: .50, .30, and .53 for the three test paragraphs.

---

[1] For the purposes of discussion, take a paragraph to be separated into clauses by any nonbracket punctuation, although actually there is a lot of room for maneuver here.

[2] For this and subsequent correlation tests, all correlations reported are significant at the p<.005 level and were measured by treating each member of the standard set in turn as the candidate and measuring it against the rest. The source of the human scores is discussed below in section 3.

3. Human Evaluation

Two different approaches to human evaluation of the standard set were tried. In the first, or traditional, approach, paragraphs were marked on a scale from 0 (not a translation) through 3 (a good translation). This was not felt to give adequately find judgments, but increasing the resolution of the scale did not seem possible, as many comments at the Les Rasses workshop confirmed. The alternative approach, suggested by a colleague familiar with similar tasks in psycholinguistics, is called *magnitude estimation.* This amounts to focusing the human rater on relative merit, with the emphasis on ratio judgments, as opposed to the absolute judgments required in the scalar approach. Experience in other domains suggests that this approach is both intersubjectively reliable and relatively insensitive to order effects, despite its apparent simplistic character. The following two paragraphs, extracted from the instructions for a further rating pilot experiment I hope to carry out soon, convey the basic technique:

> To do this, read each translation carefully. After you have read the first one, assign it a number which reflects impressionistically how good a translation you think it is. Use any scale you like. As you read each successive translation, assign it a number which reflects its quality relative to the quality of the first translation you read. Just write the scores in the left margin next to the paragraphs as you go.
>
> For example, if you assign a 12 to the first translation, and the second one seems to you to be twice as good, you would assign it a 24. If the third appears only a tenth as good as the first, you should assign it a score of 1.2. In other words, in assigning scores, focus on the ratio of goodness in each case to the original, rather than trying to arrange them all on some linear scale.

Although much further work needs to be done to validate this approach to human rating of translation quality, it is clearly promising, not only because it appears to give comparable results to traditional scalar approaches while providing better resolution, but also because it takes much less time to perform.

4. Conclusions

Especially given that no attempt was made to remove less-than-wonderful translations from the standard set, and that one paragraph (the second of the three) was clearly unusual in the demands it placed on translators and evaluation methods alike, the results are very encouraging. It seems at least possible that with the idea of evaluation based on standard sets we are well on the way to the goal of a fast, automatic measure of translation quality which correlates well with human evaluations. As a side benefit, we may also have uncovered in magnitude estimation a more reliable and less costly approach to human evaluation.