[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland;
ed. Kirsten Falkedal (Geneva: ISSCO).]

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

Henry S. Thompson
Human Communication Research Centre
University of Edinburgh

0 Introduction

The original motivation for the work reported here is the desire to improve the situation with respect to the 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.

1 A New Methodology

Most evaluation amounts to measurement against a standard. For direct evaluation of the quality of translation, this has historically been achieved by human experts, comparing the candidate translation against their expectations, possibly with an eye on a 'standard' translation or a set of guide-lines. 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 etc.

point 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 (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 set 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 forty-four English translations of three paragraphs drawn from two French texts were used.

1.1 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 normalised distance between the optimal alignment of clauses[1] between the two texts. That is to say, if e.g. 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 distances 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]

[1] For the purposes of discussion, take a paragraph to be separated into clauses by any non-bracket punctuation, although actually there is a lot of room for manoeuvre 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

For the other version, the entire matrix was processed by a Multi-dimensional 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 e.g. 3-space to each translation so that the order (non-metric scaling) or actual value (metric scaling) of the inter-translation 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 non-metric 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.

1.2 The Compound Method

Even with forty-four translations of quite short paragraphs (between twenty and seventy 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 e.g. forty-four paragraphs, but rather as forty-four 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) of 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 one with another, except indirectly, it would not be trivial to enforce a strict sequentially constraint. Rather than attempt this, for the pilot the algorithm used 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 forty-four texts in turn as the candidate and measuring it with the remaining forty-three as the basis for the compound standard, was significantly better than the simple methods described above: .59, .30 and .53 for the three test paragraphs.

turn as the candidate and measuring it against the rest. The source of the human scores is discussed below in section 2.

2 Human Evaluation

Two different approaches to the human evaluation of the standard set were tried. In the first, 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 fine judgements, 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 psycho-linguistics, is called *magnitude estimation.* This amounts to focussing the human rater on relative merit, with the emphasis on ratio judgements, as opposed to the absolute judgements required in the scalar approach. Experience in other domains suggests that this approach is both inter-subjectively reliable and relatively insensitive to order effects, despite its apparent simplistic character. The following two paragraphs are extracted from the instructions for a further rating pilot experiment I hope to carry out soon, and 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.

3 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.

Appendix. Exemplary data

The French originals and three translations drawn from those I collected are given below for the three texts used whose evaluation results are described in this report. In each case I've chosen three translations which human evaluators ranked high, middle and low, and given the percentile ranking of the translation, first by the human evaluator and then by the meta-distance measure. For example, the notation 33/81 means that the human score was 33 when scaled from 0 to 100, and the meta-distance measure 81.

Letter para. 3

French Original

Je vous remercie d'avoir bien voulu participer à ce colloque dont la Commission tirera le plus grand profit et vous prie d'en trouver ci-joint le compte-rendu.

Good translation 100/96

Thank you for having agreed to take part in this workshop, which will be of the greatest benefit to the Commission. Please find the report attached.

Mediocre translation 33/81

I would like to thank you for agreeing to take part in this colloquium which will be of immense benefit to the Commission and am pleased to enclose herewith the programme.

Poor translation 7/0

I request your help in participating in this colloquy which gives the Commission the biggest profit and request a finding jointly with the complete account.

Letter para. 4

French Original

J'attire votre attention sur le fait que le document "LIFE" qui vous a été distribué contient une bibliographie importante en annexe et je vous serais particulièrement reconnaissant si vous pouviez m'indiquer quelques références méritant d'y être ajoutées.

Good translation 71/100

I draw your attention to the fact that the document "LIFE" which was distributed to you contains a large bibliography as an appendix. I would be particularly grateful if you could let me know of further references worthy of inclusion.

Mediocre translation 64/64

Please note that the document entitled LIFE which had been distributed to you contains as an annex a large bibliography. I would be extremely grateful if you could point to references which would be worthy of inclusion.

Poor translation 0/47

I draw your attention to the making of the LIFE document which contains as distributed an important bibliography annexed and I wish you to particularly recognized if you indicate to me each references which deserve.

Para. para 2

French Original

Les langues constituent le véhicule de l'information, notamment de l'information économique. La création d'un marché unique européen demande que tous les partenaires participant aux activités économiques puissent avoir accès aux informations mises à leur disposition dans des langues autres que la leur et qu'inversement ils puissent communiquer les informations qu'ils destinent à des personnes ne parlant pas leur langue. C'est le problème du transfert de l'information entre les langues, autrement dit de la traduction.

Good translation 100/91

Languages constitute the vehicle of information, in particular of economic information. The creation of a single European market demands that all partners participating in economic activities have access to information placed at their disposal in languages other than their own, and that conversely they can communicate information to others not speaking their own language. This is the problem of the transfer of information between languages, in other words of translation.

Mediocre translation 60/79

Languages are the vehicle for the transfer of information, and particularly of economic information. The creation of a European common market necessitates that all economic partners have access to information communicated to them in languages other than their own, and conversely, that they be able to transmit information to people who do not speak their language. The problem is the transfer of information between languages: in a word, translation.

Poor translation 40/52

Languages represent the information vehicle, particularly for economic information. The creation of a unique European market requires that all partners, participating to economic activities, could have access to available information in other language then their own and conversely they could communicate information to people of a different language. This is the information transfer problem across languages, in other words the one of translation.

References

Alshawi, H., D. Carter, B. Gambäck and M. Rayner, 1991a (to appear). "Translation by Quasi Logical Form Transfer", 29th Annual Meeting of the Association for Computational Linguistics, University of Berkeley, California.

Language and Machines. 1966. Computers in Translation and Linguistics. Washington D.C.: Division of Behavioural Sciences, National Academy of Sciences, National Research Council. Publication 1416. A Report by the Automatic Language Processing Advisory Committee (The ALPAC report).

Falkedal, K. 1990. *Evaluation Methods for Machine Translation Systems: An Historical Overview and a Critical Account,* Technical Report, Swisstra, Geneva.

Falkedal, K. and King, M. 1990. *Using Test Suites in Evaluation of Machine Translation Systems,* in Proceedings of COLING 90, Helsinki.

Gambäck, B., H. Alshawi, D. Carter and M. Rayner, 1991b (to appear). *Measuring Compositionality in Transfer-Based Machine Translation Systems,* Work-
shop for Evaluation of Natural Language Processing Systems, University of California, Berkeley, California.

Gervais, A. 1980. *Evaluation du système-pilote de traduction automatique TAUM-AVIATION.* Ottawa Canada: Bureau des traductions, Secrétariat d'Etat. Rapport final.

Heid, U. 1988. *Evaluation der französisch-deutschen SYSTRAN-Übersetzung.* Stuttgart: IMS. Vorhabenskizze.

Heid, U. 1990. *Evaluation und Verbesserung der Sprachrichtung Französisch-Deutsch des Maschinellen Übersetzungssystems SYSTRAN.* Bericht des IMS für den Zeitraum 1.7.89 - 30.4. Vorversion.

Hildenbrand, E. and Heid, U. 1990. *Ansätze zur Ermittlung der linguistischen Leistungsfähigkeit von maschinellen Übersetzungssystemen.* Zur Entwicklung von Französisch-Deutschem Testmaterial für SYSTRAN. Paderborn. Talk presented at Linguistisches Kolloquium.

King, M. 1989. *A Practical Guide to the Evaluation of Machine Translation Systems,* Technical Report, Swisstra, Geneva.

King, M. 1990. *A Workshop on Evaluation: Background Paper.* In Proceedings from the Third International Conference on Theoretical and Methodological Issues in MT, pp.255-259. Linguistic Research Center, University of Texas at Austin.

Knowles, F. 1979. *Error analysis of Systran output – a suggested criterion for the 'internal' evaluation of translation quality and a possible corrective for system design.* In Snell (ed.) Translating and the Computer, pp.109 - 134. North-Holland Publishing Company.

Leick, J. M. and Schroen, D. 1978. *Quelques résultats statistiques d'une évaluation sommaire du système de traduction automatique Systran.* CETIL, CCE. Information document.

Miller, G. A. and Beebe, J. G. 1958. *Some Psychological Methods for Evaluating the Quality of Translations.* Mechanical Translation, v. 3, pp.73-80.

Pallett, D. S. 1988. *Types of evaluation methodology.* Talk presented at the workshop on Evaluation of Natural Language Processing Systems, Wayne, Philadelphia December 8-9.

Pfafflin, S. M. 1965. *Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments.* Mechanical Translation, v. 8, pp.2- 8.

Samuelsson, C. and M. Rayner, 1991 (to appear). *Quantitative Evaluation of Explanation-Based Learning as an Optimization Tool for a Large-Scale Natural Language System.* 12th International Joint Conference on Artificial Intelligence, Sydney, Australia.

Shiwen, Y 1991. *Automatic Evaluation of Output Quality for Machine Translation Systems,* this volume.

Slocum, J. and al. 1985. *An Evaluation of METAL: the LRC Machine Translation System.* In Proceedings of the Second Conference of the European Chapter of the Association for Computational Linguistics, pp.62 - 69. Geneva.

Van Slype, G. 1979. *Critical study of methods for evaluating the quality of machine translation.* Bruxelles and CCE: Bureau Marcel Van Dijk.

Wilks, Y. and LATSEC Inc. 1979. *Comparative Translation Quality Analysis.* LATSEC Inc.. Final Report. Contract F33657- 77-C-0695.

“The Evaluation and Systems Analysis of the SYSTRAN Machine Translation system.” 1977. New York: Battelle Colombus Laboratories, Rome Air Development Center, Air Force Systems Command, Griffiss Air Force Base. RADC-TR-76-399 Final Technical Report.