AUTOMATIC EVALUATION OF COMPUTER GENERATED TEXT: A PROGRESS REPORT ON THE TEXTEVAL PROJECT

Chris Brew, Henry S. Thompson

Human Communication Research Centre
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, SCOTLAND
hthompson@edinburgh.ac.uk

ABSTRACT

We present results of experiments designed to assess the usefulness of a new technique for the evaluation of translation quality, comparing human rankings with automatic measures. The basis of our approach is the use of a standard set and the adoption of a statistical view of translation quality. This approach has the ability to provide evaluations which avoid dependence on any particular theory of translation, which are therefore potentially more objective than previous techniques. The work presented here was supported by the Science and Engineering and the Social and Economic Research Councils of Great Britain, and would not have been possible without the gracious assistance of Ian Mason of Heriot Watt University, Edinburgh.

INTRODUCTION

The TextEval project aims to explore and develop a new approach to the automatic evaluation of computer-generated texts, based on the use of standard sets.We believe that fast, accurate and automatic evaluation methods are vital to the development of any large piece of natural language software, and note that current methods, which involve extensive intervention by human experts, are too costly to be a routine part of the development cycle. For pragmatic reasons we are working on translations, but in principle the techniques would apply to any body of suitably comparable texts, such as alternative versions of an instruction sheet.

Our purpose in this presentation is to describe a framework for the automatic evaluation of translation quality. We will start by presenting the results of an evaluation experiment, then use the results of this to assess the usefulness of several alternative automatic metrics.

TRANSLATION QUALITY.

Although it is not clear precisely what translators' subjective judgements are based on when they report on the quality of translations, it is worthy of comment that such judgements can be elicited at all, since attempts to prescribe detailed guidelines for the scientific evaluation of translation quality have been problematic since the ALPAC report [1]. It is arguable that any particular set of guidelines will inappropriately constrain the evaluator by imposing the theoretical preconceptions of the author of the guidelines. In the current project we are exploring the possibility of producing automatic evaluations without any prior commitment to a theory of translation.Our method of evaluation depends only on the essentially statistical hypothesis that good translations will tend to be more similar to each other than will bad ones. Pre-requisites of this approach are the availability of a suitable corpus of translations and the choice of a similarity metric. In this paper the metrics which we describe will be (negative log) probabilities Strictly this makes them dissimilarity metrics.

Once we have established a metric, we may apply two approaches to the generation of a rank ordering. In the first approach we find an appropriate method of combining the elements of a set of translations, then measure the distance of each individual translation from the composites formed by the respective remainders of the standard set. Under our hypothesis the better the translation is the lower will be its distance from the composite formed by the remaining data.In the second approach we start by generating a pairwise distance matrix, then use multi-dimensional scaling [3,4(p 498)] to reduce the data to one dimension. This produces not only a linear ordering over the items tested but also a measure of the extent to which this linear ordering captuures the relationships described by the distance matrix.

We will introduce our framework by describing an experiment on the extraction of

translation quality judgements from human beings, along with the analysis which demonstrates that these judgements do indeed reflect some objective reality about the translations. We will then carry out a very similar analysis of results obtained from automatic metrics.

HUMAN EXPERIMENTS

Subjects were a class of final year translation students from an established translation course at a Scottish University. We restricted our attention to translations made by native speakers of English. It can safely be assumed that these students have a high level of competence in both French and English. In Experiment 1 they were asked to make judgements about the quality of translations prepared by others, while Experiment 2 required them to assess the quality of translations produced by classmates. Since we are taking a theory neutral approach, we did not offer the subjects any guidelines for their judgements other than asking them to assess the quality of the translations. Since translation courses assume basic linguistic competence and concentrate their efforts on showing translators how to preserve the message of a text in translation [5],one might expect that the basis on which their judgements are made might be on global impressions of the translation rather than on small details. We took as our starting point the technique of Magnitude Estimation [7], long used in the social sciences for evaluative tasks where forced choice scoring is difficutlt or inappropriate. It is ideal for our purposes as it is robust, validatable and order insensitive.

Experiment 1

The first experiment made use of translations which had previously been elicited by electronic mail for use in a pilot study [8]. The volunteers who submitted translations differed considerably in background and experience of translation, and there were concomitant substantial differences in the quality of the translations which they produced. The original corpus consisted of 44 translations of the same piece (a report on the opportunities and dangers provided by Europe's peculiar position as a multilingual community).. For the experiment reported here we selected 10 translations spanning the quality range of the corpus as a whole. From each of these translations we drew an extract corresponding to 111 words of French. The translations varied in length from 86 to 129 words .

Subjects were asked to respond in two modalities, line production and numerical estimation. The use of two modalities was originally motivated by the requirements of alternative analytical method based on magnitude estimation. For present purposes it can be regarded as a somewhat elaborate means of eliciting two judgements of each translation from each rater.

Under both modalities the subject is asked to compare a series of translations with a reference translation which remains present throughout the sequence of tests. In the case of line production the reference translation is associated with a pre-drawn line of a particular length, and the subject is asked to indicate an assessment of the current target translation by producing a line which is longer or shorter than the line associated with the reference translation. In the case of numerical estimation the the reference translation is described by a number, and the subject is asked to indicate their assessment by producing a number whose ratio to the reference number best reflects the relative quality of the target translations.

24 subjects took part. The first part of the session was taken up with a calibration exercise designed to familiarize the subjects with the responses required. In the second part of the session we asked for judgements of the 10 translations under both modalities, allowing 90 seconds for each judgement. Because of the limited time available, rating was paced by the experimenter

ANALYSIS

Experiment 1

In order to establish whether there is consistency between the we use Kendall's coefficient of concordance [4 pp 454-456]. This is a measure of the agreement in ranking between the raters, and is associated with a χ^2 statistic for which a significance test is available.For both line production and numerical estimation the results were highly significant. (For Line Production Kendall's $W = 0.3049 \chi^2(23) = 70.121$ $p < 0.001$) Numeric Estimation Kendall's $W = 0.3263$ $\chi^2(23) =$

75.054 $p < 0.001$) indicating that there is a measure of agreement between raters. It would be a surprise if this were otherwise.

We also carried out an experiment as a baseline for our hypothesis about similarity metrics.The probabilistic distance metric which we used was the matched t-test [4 pp 287-290]. For each pair of translations we used the t-test to calculate the probability that these translations were rated the same by our subjects, then transformed the probabilities into negative log space to form the distance metric.We then reduced this data to a single dimension using multidimensional scaling.

We wished to assess the extent to which the linear scale produced by the multidimensional scaling describes the variability in the data. For line production the r^2 value was 0.693, indicating that the single linear dimension accounts for this proportion of the variance in the pairwise distances. For numerical estimation the corresponding figure was $r^2 = .645$. Multidimensional scaling is essentially an exploratory technique, therefore it is inappropriate to read too much into these figures, but they are large enough to suggest that there is some connection between the input data and the results of scaling.

For numerical estimation the Spearman rank correlation between the result produced by scaling and the original data was 0.95 and Pearson's ρ was 0.96. For line production the corresponding statistics were Spearman ρ = 0.41 Pearson $\rho = 0.098$, but these low figures were entirely due to the fact that scaling had assigned a very low score to a single very high quality translation. Our basic hypothesis allows, and indeed encourages this, because all that the scaling process knows is that this translation is very different from the others. In supplying only inter-translation distances we have effectively suppressed the information that this translation is better rather than worse than the generality of translations. Once this aberrant translation is ignored the statistics become Spearman ρ=0.95 Pearson ρ=0.93, indicating the same good fit as in the line production case.

In general, whenever we apply the sacling technique we need to be alert to the possibility that outliers will be classified in the wrong extreme region of the scale. For the purposes of evaluation of machine translations it is unlikely that this will be a major problem, since it is unlikely that any current system will produce results much better than those of the human beings in the standard set. On the other hand, it is certainly possible that we will encounter the converse problem, which is that machine translations may be so bad as to cast doubt on the relavance of comparisons based on the differences between human translations, which are typically of much higher quality.

Experiment 2

In Experiment 2 the same procedure was followed as for Experiment 1,In both cases the texts which were used were previously unseen by the raters, having been given as an exercise only to the other half of the subject pool. Since slightly more time was available, rating was self paced, although raters were advised to spend no more than two minutes on each translation. The text was 127 words of French extracted from a report on economic prospects for Europe, with the length of the translation extracts used ranging from 87 to 143 words.The standard set consisted of 14 translations, and the total number of raters was also 14

Experiment 3

In Experiment 3, which was carried out in the same session as Experiment 2, the text was 137 words of French, extracted from the annual report of a European initiative for the dissemination of technical information. When translated this produces between 115 and 155 words of English. In this experiment there were 9 raters providing judgements on a set of 16 translations.Analysis for Experiments 2 and 3

In experiments 2 and 3 we again found prima facie evidence that the raters were agreeing on something. For experiment 2 the coefficient of concordance was 0.188 for 14 subjects using line production ($\chi^2(12) = 31.61$ $p < 0.01$) and 0.206 ($\chi^2(12) = 34.61$ $p < 0.001$) using numerical estimation.For experiment 3 the figure for line production was 0.145 ($\chi^2(8) = 18.5333$ DF= 8 $p < 0.025$), and for numerical estimation 0.25026 $\chi^2(8) = 32.0333$ $p < 0.001$).

In experiment 2 multidimensional scaling produced r^2 values of 0.882 for line production and 0.906 for numerical estimation, while in experiment 3 the

corresponding results were 0.791 for line production and 0.743 for numerical estimation.

The correlations with the original input data were all significant ($p < 0.01$) For the 14 subjects of Experiment 2 line production gave Spearman $\rho = 0.88$, Pearson $\rho = 0.94$, while numerical estimation gave Spearman $\rho = -0.96$ Pearson $\rho = -0.98$ (The negative correlation is not a problem, since multi-dimensional scaling involves an arbitrary choice of direction for the linear scale). For the 9 subjects of Experiment 3 line production gave Spearman $\rho = 0.85$ Pearson $\rho = 0.94$ while numerical estimation gave Spearman $\rho = 0.65$ Pearson $\rho = 0.88$. In the case of numerical estimation multidimensional scaling has been slightly less successful in matching the original data, but all correlations are still significant.

Discussion

The experiments which we have carried out suggest that when our subjects provided ratings of translations they are achieving a measure of agreement,and that multidimensional scaling is indeed capable of recovering an appropriate linear order from a matrix of probabilistic distance measures.

AUTOMATIC METRICS

In this section we illustrate our approach to the construction of automatically applicable metrics.The first type of model is a simple multinomial. In this model we focus exclusively on the frequency distribution of the words within a corpus. Given two multinomial distributions a technique described by Dunning [2] makes it possible to calculate the log probability that the two distributions are drawn from the same model. This is a distance metric analogous to the t-test which we used in the analysis of human judgements.In one variant of our technique, which we call the direct approach, we measure the probability that each translation is drawn from the same distribution as a composite formed from the remainder of the standard set, while in the other variant we calculate pairwise distances between each version, again using multi-dimensional scaling to reduce the matrix to a linear order. If the results of either of these approaches are a good match for human performance, then there is some suggestion that word population influences subject judgements.

As an alternative to simply counting words, we have used the Xerox part-of-speech tagger [6] to assign part-of-speech tags to the words of the translations. We can now apply the same multinomial techniques as we did with words. What we are doing here is to collapse across the equivalence classes which the tagger has identified. This metric is of interest, since if it matches human performance better than does the word-based metric then there is some suggestion that word-class statistics influence subject judgements.

]The final multinomial model which we have considered is again based upon information which is available within the Xerox part-of-speech tagger, but instead of collapsing across the parts of speech actually assigned we use only information contained in the tagger's lexicon.This takes the form of *ambiguity classes*, which are statements about the sets of possible parts of speech which can in principle be assigned to a particular lexical item. In contrast to the metric based on tags, but like that based on words, this metric is insensitive to actual way in which words are used in a given translation,but depends only on the words used. The difference from the word-based metric is simply that we have used the tagger's lexicon to collapse across equivalence classes of similar words.

EXPERIMENTS WITH AUTOMATIC METRICS

In these experiments we explored the usefulness of the three types of multinomial model , using both the direct approach (in which each translation is reduced to a vector of counts, and each count compared against the aggregate counts for the rest of the corpus), and the multidimensional scaling approach based on the reduction of pairwise counts to a single linear scale.

Experiment 4

In this experiment we used the 10 texts which had previously been used for Experiment 1. Under both modalities the direct approach produced small negative correlations for all types of multinomial model, but none of the correlations were significant.

Multidimensional scaling of the distance matrices produced by part-of-speech tagging, word counting and the counting of lexical ambiguity classes yielded the information that the proportion of variance accounted for in the reduction to a linear scale was r^2 = .492 for part of speech tags, r^2= .508 for words and r^2 = .473 for classes.Although this indicates that there was some linear component to the scales induced by the distance metrics, correlation analysis revealed that for this translation at least none of these scales seemed to mimic human performance.

Experiment 5

Experiment 5 used the 14 translations previously used in Experiment 2. The direct approach again failed to reveal any metric which correlated significantly with human performance using either line production or numerical estimation.

In the scaling approach we found that the linear reduction of the tag counts accounted for a larger fraction of the variance in its matrix than did the ambiguity classes in theirs. The linear reduction of the word count distance accounted for so little of the variance that the linear scale must be viewed as worthless irrespective of any correlation with human ratings (Tags r^2 = .465,Classes r^2=.258,Words r^2 = .083).

For line production the linear scale based on part of speech tags correlates significantly with the human data (ρ= .4713 p =0.0495). The other two metrics did not show significant correlations.For numerical estimation the same pattern emerged , with only the metric based on tags correlating significantly with human performance(ρ = -0.6159 p = 0.0362).

Experiment 6

This used the 16 translations from experiment 3, again subjecting them to the three multinomial distance metrics based on tags, classes and words.In the direct approach, for line production only the rating based on tags was significantly correlated (ρ = -0.5679; p = 0.0109) while for numerical estimation both tags and ambiguity classes produced significant results (For tags: ρ = -0.5555 ; p = 0.0029.For classes: ρ=0.2107 ; p = 0.0477)

The scaling approach showed the same pattern as experiment 5 The tag-based metric generated a linear scale which accounted for 0.673 of the variance in the distance matrix, while ambiguity classes did substantially worse (r^2 = .288) and the linear scale derived from words was essentially worthless (r^2= .071).As in Experiment 4, neither the metric based on ambiguity classes nor that based on tags produced significant correlations with the human data, and while marginally significant correlations did emerge for the word based metric, these have to be discounted because of the extremely low r^2 value produced in scaling.

FURTHER WORK

Once we have access to numerical information about human preferences we can be more sophisticated about the way in which we form the composite standard against which each version is assessed. In particular, rather than summing the frequencies with which words, tags or classes occur, we can generate a weighted combination of the frequencies of the individual elements, where the weights reflect the ratings assigned by human beings. Preliminary results indicate that the results of the multinomial approach are not greatly improved by this manipulation.

There are other ways of moving beyond simple multinomial models. One is the use of limited context to extend the multinomial approach to bigrams and beyond. A second is to make direct use of the frequency information encoded in the Xerox tagger to provide ratings of translations. A third approach is to radically extend the multinomial approach to allow counts of arbitrary text features, including word frequencies, bigrams,and so on,then to select a subset of the various features which closely mimics human judgements. Given the small size of the available training corpus (a total of under 10,000 words) it will be necessary to keep the number of degrees of freedom fairly small if over-training is to be avoided.

In essence the multinomial approach provides a metric of texts analogous to the power spectrum of a speech signal. There is clearly room for metrics which are analogous to phase information, i.e. those which take account of the connections between words. Such measures, such as the distance between successive occurrences of the same word, or that between an an anaphor and its antecedent, are of particular interest because students of translation are

taught [5] to pay particular attention to the problem of ensuring that their translation does not appear as a collection of apparently unrelated sentences. It is an open question whether they in fact make use of this training in carrying out the judgements described here, but if they do it should be the case that metrics of textual cohesion will be a useful complement to the essentially lexical metrics described here.

We are also pursuing the approach introduced in [8] of using string edit distance as the basis for automatic evaluation, and hope to report on this at the meeting.

CONCLUSIONS

There is some evidence that ratings based on multinomials are capable of capturing some human intuitions about translation quality. Although this varies from text to text, it does appear that the part of speech information which can be obtained by automatic tagging represents a promising way of collapsing across equivalence classes of words. By contrast, the results of multidimensional scaling using word counts suggest that there is too much irrelevant information in these counts to allow an automatic system to make much use of them in rating translations. Both the direct approach and that using multidimensional scaling show some success, although each failed on the translation for which the other succeeded.

REFERENCES

1 ALPAC. 1966. *Language and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee,Division of Behavioral Sciences, National Research Council, National Academy of Sciences, Washington D.C.

2 Coxon A.P. 1982. *The User's Guide to Multidimensional Scaling*, Exeter, NH, Sage Publications.

3 Dunning, Ted. 1993. "Accurate Methods for the Statistics of Surprise and Coincidence" *Computational Linguistics* 19(3) pp 61-74, March,

4 Hatch, E. and Lazaraton, A. 1991. *The Research Manual: Design and Statistics for Applied Linguistics* Newbury House, London.

5 Hatim, B. and Mason, I. 1991. *Discourse and the Translator* Longman, London.

6 Kupiec, J. 1992 "Robust part-of-speech tagging using a hidden Markov model." *Computer Speech and Language*, 6(3):225--242, July.

7 Lodge, M. 1981. *Magnitude Scaling: Quantitative Measurement of Opinions.* Sage Publications, Beverley Hills, London.

8 Thompson, H. S. 1991. "Automatic Evaluation of Translation Quality: Outline of Methodology and Report on Pilot Experiment". In K. Falkedal (ed) *Report from the Evaluators Forum*. ISSCO, Geneva, to appear.