Automatic Evaluation of Computer Generated Text: Final Report on the TextEval Project

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

July 2, 1996

Abstract

The report describes computational and psychological experiments carried out under grant number GR/H29179 from the Science and Engineering Research Council of Great Britain. These would not have been possible without the gracious assistance of Ian Mason of Heriot Watt University, Edinburgh. We aimed 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.

Introduction

The project: The TextEval project was designed 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 chose to work on translations, but in principle the techniques would apply to any body of suitably

> Par contre, le fait d'etre le seul bloc economique et industriel important dans le monde a devoir trouver une solution a de tels problemes peut aussi donner a l'Europe un avantage economique considerable.
>
> On the other hand, Europe may derive considerable economic advantage from its position as the world's only large economic and industrial bloc faced with the need to find solutions to such problems.

Figure 1: A good translation

comparable texts, such as alternative versions of an instruction sheet, or multiple answers to a suitably structured class assignment. One follow up study which we will attempt if possible is to apply our techniques to a corpus of student-generated texts described by Charney and Carlson (1994). In this work psychology methods sections are elicited from large classes of students who watch a filmed account of an experiment. Like translations, these texts can plausibly be regarded as being attempts to convey a particular message, so the ways in which they vary should be similar to the ways in which translations vary.

This report: The purpose of this final report is to describe our progress toward a framework for the automatic evaluation of translation quality, and indicate how our understanding of this framework has developed in the course of the explorations which we undertook during the TextEval project. 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

It is evident that some translations are better than others. The reader will probably be prepared to agree that the translation in figure 1 is more satisfactory than that in figure 2.

We are reasonably confident about this judgement, because we have arranged for numerous alternate versions of these passages including these sentences to be subjectively rated by translators, and the extracts shown are consistently rated at the top and bottom of the range respectively.

Some possible reasons for these judgements are the following:

- **Literalness:** The second translation follows the French very literally, Par contre, le fait d'etre le seul bloc economique et industriel important dans le monde a devoir trouver une solution a de tels problemes peut aussi donner a l'Europe un avantage economique considerable.

On the other hand, the fact of the existence of the single economic and industrial bloc of world importance which has made it necessary to find a solution for these vexed problems, has also given Europe a considerable economic advantage.

Figure 2: A less satisfactory translation

and thereby becomes unidiomatic in English, while the first involves an extensive reformulation into appropriate and idiomatic English.

- **Inaccuracy:** The second sentence incorrectly conveys the impression that an economic advantage has already materialised, while both the French and the first translation indicate that this is merely a possibility.
- **Processing:** The relative clause in the second sentence is hard to interpret, since the relative pronoun is separated from its antecedent (the "fact"). This is liable to cause confusion, since the reader may initially attempt to attach the relative pronoun to an alternative antecedent (the "single industrial and economic bloc").

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 (1966). 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. We explored 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. Once we have established a metric, we may apply two approaches to the generation of a rank ordering. In the first approach we chose 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 started by generating a pairwise distance matrix, then use multi-dimensional scaling (Coxon 1982; Hatch and Lazaraton 1991) 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 captures the relationships described by the distance matrix.

Sparse data

A running theme will be a series of encounters with problems which arise primarily from the sparsity of the data which we are able to obtain. There are two aspects to this problem. The first is familiar in statistical language processing: no text is ever long enough to be an entirely adequate representation of the language within which it is written. Unfortunately there is a limit to the length of the texts which a class of students are able to translate as a class assignment. Although techniques exist for obtaining useful frequency information from fairly small corpora (Dunning 1993), it is clear that the feasibility of the techniques developed in this project hinges crucially on the availability of large amounts of text in multiple versions.

The second aspect of the sparse data problem is more specific to this project: we do not have a huge number of versions of any of the translations. In no case were we able to elicit human judgements of more than 17 versions, and for some of the texts these were all the usable translations which we had. The space of possible translations of a text appears to be an enormous one; and the same is apparently true even of the spaces of very good and/or very bad translations. Our investigations indicate that a very large number of data points would be required in order to make reliable and useful discriminations between translations drawn from the different classes. The compound distance metric idea explained below goes some way towards alleviating this shortage of translations, but at the expense of potentially introducing artefacts of its own

Human Experiments

The basis for evaluation: We chose to build our evaluation metrics on the basis of empirical data collected from expert translators. This data consisted both of translations and of judgements of these translations. Human judgements are valuable for the following reasons:

1. We need to inspect differently rated texts in order to get ideas about the criteria which human raters may be using in order to make their

judgements. These can form the basis of automatic techniques.

2. Human evaluations are the standard against which automatic evaluation techniques need to be judged. We therefore need reliable human evaluations in order to in turn evaluate the automatic techniques.
3. Most of our techniques depend on the idea of augmenting the effective size of the standard set by forming composite translations formed from one or more genuine translations. In taking this route we involve ourselves in heuristic search of a large and rather unstructured space of possible translations. Knowing about human judgements allows us to impose a degree of structure on this space, for example by searching for differences between documents which tend to make the same discriminations between documents as do the human raters.

For purposes of inspection even a small number of evaluated translations may be of considerable assistance, while the other two would benefit from as large a corpus as it is practical to collect.

The experiments: Subjects were a class of final year translation students from an established translation course at Heriot-Watt 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 (Hatim and Mason 1991), 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 (Lodge 1981), long used in the social sciences for evaluative tasks where forced choice scoring is difficult or inappropriate. It is ideal for our purposes as it is robust, validatable and order insensitive.

The first experiment made use of translations which had previously been elicited by electronic mail for use in a pilot study (Thompson 1991). 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 we selected 10 translations spanning the quality range of the corpus as a whole.

Modalities: 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 an 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.

Results: More details of the results of these experiments can be found in our contribution to the DARPA HLT workshop (Brew and Thompson 1994). In brief they suggest that when our subjects provided ratings of translations they are achieving a measure of agreement. We have also used this data, in combination with a reduction to distance matrices using t-tests, to demonstrate that multidimensional scaling is indeed capable of recovering an appropriate linear order from a matrix of probabilistic distance measures.

Automatic Metrics

Introduction: 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 (1993) 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.

Using part-of-speech tags As an alternative to simply counting words, we have used the Xerox part-of-speech tagger (Kupiec 1992) 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 multinomials: In three 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. More details of these experiments can be found in Brew and Thompson (1994) . In essence our finding was that no one of these techniques can produce reliable imitations of human judgements on translations of the length which we attempted.

Using human data to improve automatic metrics

Extended bigrams: 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 indicated that the results of the multinomial approach are not improved by this manipulation, so we chose to abandon this approach in favour of work more likely to produce definite results from the available corpus.

Using context: 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. We have not investigated this in detail, since it would suffer from problems of sparse data to a still greater extent than the multinomial approach. A second 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.

Shaped document clustering: We investigated a slightly more refined version of the multinomial idea, in which we carried out a search for words which can act as effective discriminators between classes of translations identified by the human raters. The first step was to identify these classes. To do this we constructed a single link dendrogram (Jardine and Sibson 1971) on the basis of the similarity matrix from the human judgements. This dendrogram clustered translations in the way which one would expect on the basis of their ratings by human beings. We then carried out word counts for each class and searched for words whose frequency was significantly different (on Dunning's likelihood ratio test) between the classes. In essence what we are doing here is picking those discriminators which are most likely to allow us to imitate the human judgement. The first step in this procedure is to verify that we can use the chosen discriminators to emulate human performance on the training corpus, and the next is to test their ability to carry out the appropriate discrimination on previously unseen translations. Unfortunately we were unable to carry out the first step, so the second fell by default. The may be several reasons for this failure, one being the sparseness of the data or the inadequacy of the metric, but another being the possible inappropriateness of the clustering scheme. It might be worthwhile to try alternative clustering schemes, such as those considered by Hughes and Atwell (1994).

The significance of commas: The only robust finding from this is that the machine translation programs to which we have access are loath to use punctuation other than the full stop, and that the usage of commas is frequently a diagnostic for highly rated human translations. This should be no surprise, since commas occur mainly when clauses are combined using conjunctions, the reason for such use of conjunctions is often to ensure that their translation does not appear as a collection of apparently unrelated sentences[1] and the machine translation systems in question have no useful notion of rhetorical purpose, or indeed any other information which operates at a level above that of individual sentences. The correct use of commas contributes to the cohesion of the text.

Other cohesive devices: In essence the multinomial approach outlined above provides a metric of texts analogous to the power spectrum of a speech signal. It says which words are there, but utterly disregards the way in which these words are organized into a sequence. This is far too post-millennial a view. 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, have the potential to pick out the same general type of information which was uncovered by the inspection of the frequency of commas.

Synonym chasing: As well as the approach based on selective use of word counts to select discriminators, we also considered an approach based on the use of word clustering and context to identify word-word substitutions. The techniques used by Finch (1993) and Hughes and Atwell (1994) can derive classes of potential synonyms from inspection of bigram statistics. We can also use a cheaper technique which makes use of alignment information. We will provide more detail on alignments in the next section.

For the moment take as read that we can obtain alignments which indicate which words of one translation are aligned with which of the other. A large subset of what we are going to call *potential synonyms* can be found by looking for situations in which a mismatch between words occurs wedged between a pair of exact matches. Some of these pairs of potential synonyms will be near synonyms (like “committee” and “group”) while others will be antonyms which pattern similarly (like “deficit” and “surplus”). We attempted to use such potential synonyms as discriminators in the same way that we did with frequency information. Potential synonyms are given a rating related to the difference in rated quality of the translations from which they are drawn. Carrying out this process over a large enough set of

[1] This is really part of a larger goal, students are taught (Hatim and Mason 1991) to value highly translations which correctly convey rhetorical purpose, and appropriate use of conjunctions is one means by which this purpose can be conveyed.

translations produces an equivalence class of words which seem to be being used interchangeably. The measure of translation quality can be obtained by picking out occurrences of members of the equivalence classes from the target text, then assigning a rating on the basis that the target text should score more highly if it uses members of the equivalence class which are characteristic of good translations than bad.

Once again we must consider the length of text and the number of translations which would be needed to achieve robust equivalence classes. Certainly the texts which we tried were far too short, and while the equivalence classes were interesting to inspect, it did not prove possible to generate worthwhile ratings automatically. It may be possible to make further progress with a weaker version of this,in which every pair of words in the cross product of the contents of aligned clauses is considered as a potential synonym. This is essentially a variant of the IBM approach to statistical MT (Brown et al. 1988). We have not attempted this since it is computationally demanding and likely to be hamstrung by the small size of the available corpus.

Edit distance

The technique which we have explored in the greatest detail is the edit distance measure which we originally proposed in the pilot study. The basis of this approach is the dynamic programming technique for string-to-string correction used by Wagner and Fischer (1974). This calculates the minimum cost sequence of edit operations for transforming one string into another.

The implementation: Because of the heavy dependence of the project on calculations similar to edit distance, we have developed a high-level framework for declarative specification of dynamic programming problems. This uses the Common Lisp macro system to hide most of the detail of implementation. We show the essence of the definition of the simplest version string edit measure in figure 3. The crucial elements of this procedure are `define-move`, and `do-moves`. The former allows declarative specification of the form

```
(define-move <move-name> <xstep> <ystep>
             (<x> <xseq> <y> <yseq>)
     <penalty-expression>)
```

where `<penalty_expression>` is a Lisp form which may use `<x>`, `<y>`, `<xseq>` and `<yseq>` to calculate the penalty which should be applied in moving from `(<x>,<y>)` to `(<x>+<xstep>,<y>+<ystep>)`.

```
(defun wagner-fischer(xseq yseq)
   "Return the dynamic programming matrix for the
    Wagner-Fischer string-to-string correction algorithm"
  (define-move :match 1 1 (x xseq y yseq)
    "Calculate penalty for match of
     source and target elements."
    (let ((xel (inrange-elt xseq x))
  (yel (inrange-elt yseq y)))
      (if (and xel yel (eq xel yel))
  0 ; exact match
  *mismatch-penalty*))) ; mismatch
  (define-move :ins 0 1 (x seq1 y seq2)
    "Calculate penalty for insertion into target"
    *omission-penalty*)
  (define-move :del 1 0 (x seq1 y seq2)
    "Calculate penalty for deletion from target"
    *omission-penalty*)
  (do-moves xseq yseq
    (:match arr i xseq j yseq)
    (:ins arr i xseq j yseq)
    (:del arr i xseq j yseq)))
```

Figure 3: The string edit measure

The second element is `do-moves` , which is responsible for filling out the dynamic programming matrix by applying the relevant moves. Since this is an abstract interface to the search for a least cost path, the low level details of the search code can be varied without compromising the clarity of the higher level code which specifies the nature of the particular instance of the dynamic programming problem [2]. This partitioning of labour makes it relatively easy to experiment with alternative distance metrics.

Using Part-of-Speech Tags Given that we are already using the Xerox tagger to segment clauses, it is convenient to also make use of the part-of-speech tags to provide a more refined version of the edit distance metric which assigns a higher penalty when neither word nor part of speech match (we call this a *full mismatch*) than when we align different words carrying the same part of speech *a partial mismatch*. Using this extra information aids considerably in the generation of good alignments.

Efficiency: Although the Wagner-Fischer technique is efficient, it becomes impractical to apply it to long strings, since the amount of computation required is proportional to the product of the lengths of the strings being matched [3]. We therefore choose to segment the text into clauses before applying the edit distance measure, which makes the process tractable by reducing the length of the strings which are compared.

Segmentation: We made use of the tokenizer which comes with the Xerox tagger (Kupiec (1992)) to achieve a preliminary segmentation into clauses. Although this tokenizer has some ability to discriminate between sentence final punctuation and similar punctuation forming part of abbreviations, the effect is best characterized as defining clauses as anything delimited by commas, full stops and other similar punctuation.

Clause merging: Since the alignment of the clauses which are picked out by punctuation may not be correct, we made use of two versions of Gale and Church's text alignment algorithm (Gale and Church 1993). One of these is the standard version in which clauses are merged according to their length in characters, while another is an adaptation of our own in which the distance metric based on clause length is replaced by edit distance. In either case, once the best segmentation of clauses had been

[2] It turns out that the high level code has to be decorated with type declarations in order to run well in CMU Common Lisp, so what we have presented is something of a simplification

[3] Improvements are possible, but depend on complex pre-compilation of strings, as detailed in Stephen (1992)

established we rate the score for the completed alignment by use of the original string edit technique. The original Gale and Church algorithm is much cheaper and yields very similar results to those obtained from our adaptation, so we chose to continue to use it.

Adjusting alignment parameters: The version of the string edit distance measure which we found most useful is that which assigns a penalty of 100 to omissions and to partial mismatches, while giving a penalty of 195 to full mismatches. The values for the penalties were determined empirically: the algorithm itself dictates that if either mismatch penalty is greater than twice the omission penalty no mismatches will appear in the best solution, and the value of 100 for omissions is arbitrary. There are therefore two parameters to vary: namely the difference between an omission and a partial mismatch, and the difference between full and partial mismatches. We then systematically varied these two parameters, searching for the regions of space which maximized the number of exact matches between words. The closer the penalty for full mismatches is to its theoretical limit of twice the omission penalty the better, and small variations in the penalty for partial mismatches make little difference, while large variations degrade performance.

Other string edit measures: We have carried out small scale explorations of more sophisticated string edit measures, notably those in which the penalties are sensitive to the membership of word-classes other than those defined by the tagger. The most interesting of these is one in which some initial estimate of the penalties is adjusted to take account of the patterns of word-word alignment which occur in the full set of pairwise alignments between members of the standard set. In principle this ought to pick out words which are synonyms (like the technique of synonym chasing outlined above it is a special case of the IBM translation work (Brown et al. 1988)), but once again we are limited by the small size of the available corpus.

Compound edit distance

The final experiment which we discuss is an application of a new version of the compound edit distance technique. (The original implementation of this technique was used in the pilot study (Thompson 1991). In this we search a large space of possibilities defined by the standard set, looking for a synthesized translation which is a good match to the target.

Preliminary alignment: The first step is to obtain candidate clause to clause matches by carring out all pairwise alignments using and Gale and Church's algorithm (Gale and Church 1993). For the texts with which we dealt this already gave very good results, in spite of the fact that it is sensitive to nothing more than the length in characters of each clause.

Choosing the best alignment: If there are N translations this preliminary step leaves us with a set of $N - 1$ possible alignments for each clause. The Gale-Church algorithm produces not only $1 - 1$ alignments (in which a single clause is aligned with a single counterpart) but also $1 - 0$,$1 - 2$,$2 - 1$ and $2 - 2$ alignments. The set of alignments involving translation k constitute a finite state transducer whose arcs are labelled with ordered pairs. One element of the ordered pair is a clauses (or a pair of clauses which have been concatenated by the Gale-Church algorithm) from translation k, while the other element is a sequence of clauses from some other member of the standard set. The best match to translation k is found by using dynamic programming to find the cheapest path which traverses the finite-state transducer while generating the clauses of translation k. It is possible to use a different scoring function for this part of the alignment process than was used for the preliminary alignment of clauses. The variation of this technique which gave the best results was to use the Gale-Church metric (based on clause length) to perform the preliminary alignment, then to choose the closest match using the variation of the string edit distance discussed in the previous section.

An example alignment

A sample alignment is given in figure . This shows the alignment of part of translation 15 with the corresponding part of the synthesized standard generated by fitting of the other 15 translations to this target. The original text is in figure 4 and figure is the synthesized standard which was fitted to the above:

As is fairly typical with these texts the synthesized translation is almost indistinguishable from the original, and even when one knows that a synthesized version is being used it is not obvious which is which

Part of the full version of this alignment, with part of speech information, is shown in figure In this alignment four clauses of each version have been used. In the first part labelled (1) the only differences are the use of `information services market` rather than `information market`, and the use of `the development of` in place of `developing`. In (2) two clauses from translation 15 (the target) have been combined to produce a good match to a similar clause from translation 1, while in (3) the reverse

The group regularly publishes bulletins looking at specific legal questions and reporting on its discussions, thus it would be of benefit to the information services market if the results of the group's work were published more widely, since these results, which have so far been promising, play an essential role in developing the information services market. The number of representatives should be increased, thereby allowing the group to contribute more effectively towards eliminating the legal obstacles which are holding back the development of the market. The development of projects has remained practically uncommenced, . . .

Figure 4: A fragment of translated text

. . . The group regularly issues bulletins analysing specific legal questions and explaining its discussions. It would be useful to the information market observatory to improve the circulation of the results of the group's work, since these results, which have so far been promising, play an essential role in the development of the information market. The group should be increased in size to enable it to contribute more effectively to the elimination of the obstacles holding back the development of the market. However, the development of projects has hardly even begun,. . . . . .

Figure 5: A synthesized text

```
...

(1)

play=VB an=AT essential=JJ role=NN in=IN
developing=VBG the=AT information=NN services=VBZ
market=NN

play=VB an=AT essential=JJ role=NN in=IN the=AT
development=NN of=IN the=AT information=NN market=NN


(2)

the=AT number=NN of=IN representatives=NNS should=MD
be=BE increased=VBN ,=CM thereby=RB allowing=VBG
the=AT group=NN to=TO contribute=VB more=QL
effectively=RB towards=IN eliminating=VBG the=AT legal=JJ
obstacles=NNS which=WDT are=BER holding=VBG back=RB
the=AT development=NN of=IN the=AT market=NN

the=AT group=NN should=MD be=BE increased=VBN in=IN
size=NN to=TO enable=VB it=PPO to=TO contribute=VB
more=QL effectively=RB to=IN the=AT elimination=NN
of=IN the=AT obstacles=NNS holding=VBG back=RB the=AT
development=NN of=IN the=AT market=NN


(3)

the=AT development=NN of=IN projects=NNS has=HVZ
remained=VBN practically=RB uncommenced=VBN ,=CM

however=RB ,=CM the=AT development=NN of=IN
projects=NNS has=HVZ hardly=QL even=RB begun=VBN ,=CM

...
```

Figure 6: The best alignment of translation 15 with the standard set

happened, with two clauses from translation 10 are combined to achieve a good match for the target.

Results on the texts used in the pilot study:

When this technique is applied to the texts which we used in the original pilot study, it yields equivalent results to those found in the pilot. Once a single outlier had been removed the pilot and main studies yielded a correlation of 0.6. Inspection of the alignments produced by the two methods revealed the following:

- It was rare for the two versions of the metric to produce the same alignments.
- When they differed the differences could often be ascribed to apparently unimportant differences in tokenization between the two versions. (The pilot experiment used a specially developed tokenizer, while the final version makes use of the tokenizer which comes with the Xerox tagger).
- A further set of differences results from the fact that the final version prefers to align words which carry the same part of speech tags, a refinement to which the version in the pilot experiments was insensitive.

It seems that the important differences are those arising the availability of a tagger in the second experiment. This additional information allows the compound metric to provide a still closer emulation of the human judgements than (already satisfactory) outcome which was achieved in the pilot. It seems clear that our current implementation of the compound-edit distance technique is a worthwhile improvement on what was used in the pilot study.

Results with the Heriot-Watt texts

However, when the same technique is applied to the longer texts in the Heriot-Watt corpus, it does not perform so satisfactorily. Our hypothesis that the combination of a standard set and compound string edit distance will be useful for evaluating translations seems not to be borne out for these texts . Two matters are at issue:

1. The ability of the compound string edit distance to generate a space of possible translations from which we can extract a synthesized standard which is reasonably close to the target.

2. The meaning of the distance remaining between the target and the closest standard synthesized from the corpus.

The first matter is resolved – at least for the translations with which we were dealing it is indeed possible to generate synthesized standards of consistently high quality. The second matter is less clear. We thought the residual distance might be strongly (negatively) correlated with human judgements of translation quality. We found this to be the case in the pilot study, but this result was not reproduced with the texts from the Heriot-Watt corpus. We think, on the basis of comparisons using the pilot study data, that our techniques are at least as effective as those in the pilot study, so we are forced to the conclusion that something about the corpus or the human judgements makes an important difference to performance.

Why does the metric perform less well than expected?

At this stage we do not know which of several differences are the ones mainly responsible for the difference in outcome, but we can list the major differences and explain why and in what way one might expect them to contribute to the change in effectiveness of our techniques.

Text length: The most obvious single difference is the much increased text length in the main study. For reasons which we discuss below, there were two versions of each of the texts used in the main study, one long and one shorter. Even the shortest of the texts used in the main study were approximately three times longer than those used in the pilot.

In the pilot study the human raters had access to the whole of the text, and this complete text was then used by the distance metric. In the main study, due to the short time available in the sessions with the human raters, it was not practical to present the full text. Instead an extract was used, chosen by the experimenter to cover as much as possible of the variation between translations. It might be argued that this handicaps the human raters by reducing the amount of information which they can use, but had we provided the full text we would have run the risk of so overloading the raters that they would have been unable to produce judgements which were anything better than random.

There was no analogous reason for restricting the amount of text to which the distance metrics had access. Nevertheless, we also applied the edit distance metric to the exact texts which the translators saw. This measure gave different results from what is obtained by applying the same metric to the full text, but was neither better nor worse at predicting human performance. It seems unlikely that this factor is a major contributor to the observed difference in performance.

However, the very length of the texts of the Heriot-Watt corpus makes it more likely that raters will base their evaluations on discourse–related criteria, whereas the short texts used in the pilot probably led the raters to concentrate on more local (primarily lexical) criteria. It seems likely that the compound edit distance metric is able to emulate the latter style of evaluation much more effectively than anything requiring sensitivity to the high-level structure of the text. We think this is a large part of the explanation for the dip in performance when moving to the longer texts.

Size of the standard set: In the Heriot-Watt corpus we are dealing with small numbers of translations. It also seems that we are dealing with weak effects, where the correlations between the human judgements and the automatic metrics are low. It may be that the effects which were detected in the pilot study are also present in the main study, but that the smaller size of the standard set makes it impossible to get a statistical demonstration of these effects. A follow up study involving a larger number of versions would clearly be desirable. One opportunity for such a study would involve the use of the texts provided by Charney and Carlson (1994).

Dynamic range: A second major difference between the pilot and the main study is the quality of the texts used. Whereas the pilot used texts coming from a mixed community ranging from experienced professional translators to people claiming next to no knowledge of the source language, the main study drew its translation samples from a relatively homogeneous group of students, nearly all of whom were producing translations of very high quality. Moreover, these translations were submitted for assessment, and may therefore have received greater care and attention than they would have received at the hands of even an experienced translator.

Indeed, once a few outliers (both good and bad) had been removed, the raters found few differences in the quality of the translations in the Heriot-Watt corpus. We were reassured to discover that translators can reliably tell the difference between human translations and post-edited machine translations, and that they achieve a large measure of agreement in identifying the very best translations. For the rest there was relatively little agreement.

One would not expect our automatic metrics to perform well on translations among which even human informants can find very little difference. Even if they had, it is unclear what significance we would have wished to ascribe to this. Recall that our original goal was to develop a metric which would allow us to measure incremental changes in the quality of automatically generated text. Since such text is currently much inferior to human generated text, it is doubtful whether criteria developed for making fine discriminations between different high-quality human translations will have much bearing on the grosser differences between differing versions of an automatic system.

Results of the project

Software

The software which was used in the project is mainly written in CMU Common Lisp as a module of the objected oriented architecture for text retrieval designed by Cutting, Pedersen and Halvorsen (1991) . This architecture was chosen because it is used in the Xerox tagger, and because it makes available generally useful abstractions for corpus handling. The code which provides the infrastructure for the architecture is freely available for research purposes (It is part of the Xerox part of speech tagger described in Cutting, Pedersen and Halvorsen (1991)). CMU Common Lisp is also freely available. Our code should be portable to any Common Lisp implementation, but has only been tested with CMU Common Lisp and (to a limited extent) with Macintosh Common Lisp.

Conclusions

We have provided 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.

The technique which has been most extensively tested in this project is the compound edit distance measure originally presented in the pilot study. Although we have a good implementation of this metric we have been unable to demonstrate that it is useful as a measure of the quality of the kind of texts which are provided by translation students. Given that both versions of the edit distance metric perform adequately on the short and relatively numerous texts used in the pilot study, it seems that the primary reasons for this are the relatively low dynamic range of the Heriot-Watt translations and the small number of them which we were able to collect. It would clearly be of interest to try our techniques with larger standard sets.

For the other methods it may also be that the use of larger corpora of translations will help. But there is no strong evidence of this, since preliminary investigations reveal that taking away single translations does not produce a concomitant decrease in quality. At least in the region we were able to explore it does not appear that the performance of any of the metrics is a fast-increasing function of the number of versions used. It remains possible that adding even a small number of translations will have a disproportionate effect. In all cases text length is lower than is really required. This seems unlikely to change, since the classroom situation is the only plausible source of appropriate texts, and it turns out that the rate at which even translation students generate parallel text is smaller than we had anticipated.

References

(1966) Language and Machines. Computers in Translation and Linguistics. A Report by the Automatic Language Processing Advisory Committee.

Brew, C. and H. S. Thompson (1994) Automatic Evaluation of Computer Generated Text: A Progress Report on the Texteval Project. In C. Weinstein, ed., *Human Language Technology: Proceedings of the Workshop*, ARPA/ISTO.

Brown, P., J. Cocke, S. D. Pietra, V. E. Pietra, F. Jelinek, R. Mercer and P. Roossin (1988) A statistical approach to language translation. In *Proceedings of the 12th International Conference on Computational Linguistics*, pp. 71–76.

Charney, D. and R. Carlson (1994) Learning to write in a genre: What student writers take from model texts. Unpublished ms., March 1994, submitted to *Reading and Teaching of English*.

Coxon, A. P. M. (1982) *The User's Guide to Multidimensional Scaling: With Special Reference to the MDS(X) Library of Computer Programs*. London: Heinemann Educational.

Cutting, D., J. Pedersen and P-K. Halvorsen (1991) An object-oriented architecture for text retrieval. In *RIAO 91*.

Dunning, T. (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* **19**(1), 61–74.

Finch, S. P. (1993) *Finding Structure in Language*. PhD thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh.

Gale, W. A. and K. W. Church (1993) A program for aligning sentences in bilingual corpora. *Computational Linguistics* **19**(1), 75–102.

Hatch, E. and A. Lazaraton (1991) *The Research Manual: Design and Statistics for Applied Linguistics*. New York: Newbury House.

Hatim, B. and I. Mason (1991) *Discourse and the Translator*. London: Longman.

Hughes, J. and E. Atwell (1994) The automated evaluation of inferred word classifications. In A. Cohn, ed., *Proceedings of the Eleventh European Conference on Artificial Intelligence*, pp. 535–539.

Jardine, N. and R. Sibson (1971) *Mathematical Taxonomy*. London: Wiley.

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

Lodge, M. (1981) *Magnitude Scaling: Quantitative Measurement of Opinions*. Beverley Hills, Cal. and London: Sage Publications.

Stephen, G. A. (1992) String search. Technical Report TR-92-gas-01, School of Electronic Engineering Science, University College of North Wales, Bangor, Wales.

Thompson, H. S. (1991) Automatic evaluation of translation quality: Outline of methodology and report on pilot experiment. In K. Falkedal, ed., *Proceedings of the Evaluators' Forum*, ISSCO.

Wagner, R. A. and M. J. Fischer (1974) The string-to-string correction problem. *Journal of the ACM* **21**(1), 168–73.