Estimating the predictive power of n-gram MT evaluation metrics across languages and text types

Bogdan BABYCH
Centre for Translation Studies, University of Leeds
Leeds, UK, LS2 9JT
bogdan@comp.leeds.ac.uk

Anthony HARTLEY
Centre for Translation Studies, University of Leeds
Leeds, UK, LS2 9JT
a.hartley@leeds.ac.uk

Debbie ELLIOTT
School of Computing
University of Leeds
Leeds, UK, LS2 9JT
debe@comp.leeds.ac.uk

Abstract

The use of n-gram metrics to evaluate the output of MT systems is widespread. Typically, they are used in system development, where an increase in the score is taken to represent an improvement in the output of the system. However, purchasers of MT systems or services are more concerned to know how well a score predicts the acceptability of the output to a reader-user. Moreover, they usually want to know if these predictions will hold across a range of target languages and text types. We describe an experiment involving human and automated evaluations of four MT systems across two text types and 23 language directions. It establishes that the correlation between human and automated scores is high, but that the predictive power of these scores depends crucially on target language and text type.

1 Introduction

The day by day concern of MT system developers is that their system is progressing rather than regressing as they create new rules and adapt dictionaries, or train on new corpora, depending on the architecture of the system. The concern of purchasers of a system is that it is, at the time of purchase, fit for purpose, whatever that purpose may be. They want to feel confident that the quality of the output reaches or exceeds some threshold at which it can be deemed ‘acceptable’.

We report here on a series of evaluation experiments undertaken on behalf of one such prospective purchaser, a company proposing to market a novel Internet MT service and seeking to identify those MT engines capable of delivering output of a quality acceptable to its end clients.

The objectives of the evaluations were threefold:

- assess the relative quality of the output of a number of potential engines, that is, to rank the candidate systems for a given language direction;
- establish an acceptability threshold below which the quality of translation is deemed unfit for purpose;
- provide a benchmark for evaluating quickly and at low cost the performance of later versions of the MT systems evaluated here or of possible future candidate systems.

Initial experiments were conducted with six MT engines. Two engines were subsequently eliminated from the trials, leaving one statistical engine and three linguistic knowledge-based engines, two of which were to be tested before and after adapting their dictionaries to the input corpus. Thus, in the end a total of six systems were subjected to both human and automated evaluations. The systems remain anonymous in this paper, since the focus is on the methodology rather than the absolute results.

Six languages figured in the experiment – English, French, German, Italian, Portuguese, Spanish – and although each language was a source and a target for at least two pairs, not all possible combinations were represented. Output was generated from a total of 23 of the potential 30 directed language pairs.

The source texts consisted of a collection of emails and an EU whitepaper, text types given as representative of the end users’ translation inputs.

The quality attribute we focused on in the human evaluations was adequacy – the extent to which the information content of the original, source text is judged to be preserved in the translation produced by the MT system. This decision reflected the projected use of the service for gisting and transactional correspondence rather than for publication.

The n-gram metrics we used for the automated evaluations were BLEU (Papineni et al, 2002) developed at IBM and a weighted n-gram metric WNM (Babych, 2004).

The correlations between the human judgements and the scores produced by both automated metrics were found to be highly reliable, but not in themselves sufficient for an automated score to be extrapolated to a human score. This prediction depends on two parameters of the regression line, namely target language and text type.

2 Source and reference texts

Two types of document – taken as representative of the anticipated use of MT by the target users – were used to evaluate the performance of the MT systems.

- The first 3,200 words (approximately) of a European Commission whitepaper on safe Internet access: For the purposes of human evaluation, this was divided into 150 text segments. Each segment comprised a complete sentence or heading, with the exception of very long sentences, which were split.
- A set of 36 emails, 3,800 words in length (24 business-related and 12 personal): The emails varied in length between 31 and 210 words, the average being 107 words. They were divided into 228 segments.

The whitepaper existed in official EU versions in all the languages under consideration. All language versions were checked manually to remove from any one version those few segments that did not have a direct counterpart in all the other versions. Thus a strict parallelism was enforced across all language versions.

Professional translators generated translations of the emails from English into all language versions. These were checked for parallelism in the same way as the whitepaper.

Thus each text served two functions: source text for machine translation into all available target languages; reference text against which to check all translations into that target language.

2.1 Linguistic characteristics

The whitepaper presented a number of challenges to the MT systems. First, it contained many strings of nouns and names of organisations, policies and legislation which were not in the system dictionaries. Second, some sentences were very long and complex.

The emails posed a different set of problems: phrasal verbs with multiple interpretations; abbreviated words; colloquial or new words; occasional long sentences, acronyms; named entities.

2.2 Dictionary adaptation

For two engines, in the directions French>English and English>French, we created new dictionaries to account for missing and mis-translated words.

We first generated translations of all source texts using the default dictionaries. We then created a new user dictionary for each system, using the human-produced reference version of the text as a gold standard for the target language lexicon. Finally, we generated a second batch of translations with the user-defined dictionaries.

3 Human evaluations

The human judgments served as our gold standard quality benchmark.

3.1 Evaluators

With the collaboration of research partners in Europe, we engaged as evaluators 135 native speakers of the target languages, the majority of whom were postgraduate students and non-linguists: 45 English, 33 French, 18 Italian, 18 Spanish, 12 German, 9 Portuguese. The numbers were calculated to yield three judgments per segment.

3.2 Evaluation materials

All machine translation output was first collated according to the target language. Segments of output text from different systems and different source languages were then combined automatically to create one file per evaluator, containing all emails and the entire whitepaper document.

The resulting evaluator packs were sent electronically to the coordinators in the six countries. Coordinators were given precise instructions on how to conduct the evaluations. They were asked to explain the evaluator instructions (shown below) in the target language, and to tell students to work at their own pace and take a break whenever they needed to.

The emails were divided into 228 segments (often sentences or headings) and the whitepaper into 150 segments. Each segment was paired with the ‘gold standard’ human translation, referred to as a ‘reference text’.

Each evaluator judged all 378 segments in order, unaware that the candidate texts were translations or that they came from different sources. In this way, each judge would see (and intuitively compare) segments of varying quality instead of output from one system alone.

The segments were presented in the form of a table containing 378 rows with a scoring box adjacent to each segment. Judges worked in a computer cluster and entered their scores electronically. The time taken for students to complete the evaluation varied between 1.5 and 3 hours.

This is the adequacy task set to the evaluators:

For each numbered segment, read the reference text on the left very carefully. Then decide how much of the same information you can find in the candidate text on the right. You should NOT be concerned with grammatical errors or differences in the choice of words.

For each segment, enter your score in the box in the right hand column. Please DO NOT go back to a segment once you have made a judgement.

Give each segment of text a score of 5, 4, 3, 2 or 1 where:

5 = **All** of the content is present

1 = **None** of the content is present (OR the text completely contradicts the information given on the left hand side).

NB Please bear in mind that this is a running piece of text and that it has been segmented in this way only for the purposes of this experiment.

3.3 Results and acceptability threshold

The scores for each target language text were computed as the average of the scores per segment awarded by the three judges.

We set the threshold of acceptability at the human score of 3.5. This value was established experimentally.

First, human scores given by individual judges for individual segments were mapped into the scale of weightings as follows.

Human score		**Acceptability weighting**
5	→	+ 2
4	→	+ 1
3	→	– 1
2	→	– 2
1	→	– 4

This scale is weighted against segments that receive bad, poor or average scores. It penalises segments that preserve none, little or some of the content of the source text, while rewarding – but more modestly – segments that preserve most or all of the information. It is a severe rather than a lenient scale.

The resulting score was multiplied by the number of words in each evaluated segment, e.g., if a segment with 15 words received the score –2, the product is –30.

We then summed all the products for each evaluated system in each translation direction. The intuition is that the acceptability threshold corresponds to a zero sum.

Thus, if the sum is greater than 0, the level of MT quality is ‘acceptable’ (since the majority of segments receive positive marks), otherwise the quality is ‘not acceptable’.

4 Set-up of the correlation experiment

In the first stage of the experiment we computed Pearson’s correlation coefficient *r* between automated N-gram metrics (BLEU and WNM) and the human evaluation scores. We also computed the two parameters of the regression line (the slope and the intercept), which allow us to predict human scores, given automated scores for some new system:

HumanSc = ***Slope*** * AutomatedSc + ***Intercept***

All coefficients were computed individually for each target language and for each evaluated text type.

The resulting figures are given in Table 1.

TL/Text Type	*r* corr BLEU/ WNM	Slope BLEU/ WNM	Intercept BLEU/ WNM
DE/em	0.7694 0.8824	1.3301 0.9973	-0.7663 -0.4373
DE/wp	0.9168 0.9487	0.0898 0.0915	-0.2275 -0.1583
EN/em	0.7699 0.8215	0.5996 0.6096	-0.2291 -0.1247
EN/wp	0.7086 0.6742	0.0961 0.0957	-0.1992 -0.1026
ES/em	0.3539 0.5674	0.0997 0.1224	0.1099 0.1599
ES/wp	0.8909 0.8487	0.2823 0.2355	-0.7484 -0.5198
FR/em	0.7732 0.8202	0.5043 0.4278	-0.1636 -0.0394
FR/wp	0.7182 0.7883	0.1351 0.1360	-0.2153 -0.1371
IT/em	0.7400 0.7345	0.2847 0.1965	-0.0573 0.0841
IT/wp	0.9064 0.8873	0.1687 0.1379	-0.3803 -0.1918
PT/em	0.7833 0.7660	0.7996 0.5787	-0.3568 -0.1390
PT/wp	0.9020 0.9174	0.3042 0.2774	-0.9233 -0.7709

Table 1. Correlation and regression coefficients

It can be seen from the table that there are differences in terms of absolute values for correlation, slope and intercept across languages and text types.

In the second stage of the experiment we addressed the problem whether the differences in values of these coefficients are statistically significant or whether they can be attributed to chance and random error that may be due to the relatively small size of the evaluated text.

In order to answer this question we contrasted the computed coefficients with a gold standard MT evaluation benchmark – the DARPA 94 MT evaluation corpus (White et al., 1994). We used the French-into-English part of the corpus, which contains 100 news texts, each text being approximately 360 words long. For a corpus of this size, high correlation figures are reported for both BLEU and WNM with human evaluation scores (Babych and Hartley, 2004).

In the current experiment we divided the DARPA corpus into 10 chunks; each chunk contained 10 texts and was about the same size as our new source texts – approximately 3,600 words.

We generated BLEU and WNM scores for each text in the corpus using a single human reference. Since two independent human translations are available for each text in the DARPA corpus, the scores for each text were generated twice – using both the 'expert' and the 'reference' human translations. We then computed the average human and automated scores for the 10 texts in each of the 10 chunks. These scores became the basis for making comparisons with the corresponding scores in our new texts.

The comparison was done in the following way: first we examined the variation of the correlation and regression parameters across chunks in the DARPA corpus; second, we established whether the same parameters in our new texts were within the limits of such variation or whether they stood significantly beyond the outer limits of such variation 'noise'. If so, they could be said to carry some 'signal' about the evaluated target language or the text type.

We computed the same set of parameters for each chunk: the *r* correlation coefficient, and the slope and the intercept of the regression line.

We assessed the variation of these parameters in the DARPA corpus by computing the average and standard deviation figures for the 10 chunks. These figures are presented in Table 2.

TL/Text Type	*r* corr BLEU/ WNM	Slope BLEU/ WNM	Intercept BLEU/ WNM
EN/news/ AVERAGE	0.6709 0.7666	0.4611 0.404	-0.096 -0.0009
EN/news/ STDEV	0.1873 0.1799	0.2479 0.203	0.1676 0.1376

Table 2. Average and StDev: DARPA

It can be seen from the table that, on average, WNM scores have a higher correlation with adequacy than do BLEU scores (*r* = 0.767 vs. 0.671), which confirms previous results obtained on the complete DARPA corpus and on other texts (Babych and Hartley, 2004a; Babych and Hartley, 2004b). However, since the size of the evaluated texts is smaller, the standard deviation figures are also high (about 25%–30% of the mean) and, again, slightly higher for BLEU.

On the one hand, such a high level of variation 'noise' on smaller corpora makes any predictions about MT evaluation scores more risky; on the other hand, for the purposes of the current experiment we are not interested in specific predictions per se; rather we want to know if the accuracy of such predictions depends on the target language or text type. For this purpose having a smaller corpus with 'noisier' variation is even beneficial, because only the parameters that carry the strongest 'signal' will stand out from the noise.

For each of the correlation and regression parameters in our new texts we computed the z-score (the standard score which tells how far the tested score is from the expected average in terms of standard deviations):

$$z = \frac{TestedSc - ExpectedMean}{STDEV}$$

ExpectedMean and *STDEV* are taken from Table 2, while *TestedSc* comes from Table 1.

We assume that variations in the DARPA scores fit a Gaussian distribution, so 95% of the points are within the limit of 1.96 standard deviations from the mean, and 99% are within the limit of 2.576 standard deviations. Therefore, if the z-score for a particular parameter is outside the range ±2.576, we can be 99% confident that the difference between the tested parameter and the corresponding parameter in the DARPA corpus can be attributed to some features in the target language and the text type, and did not happen by chance, e.g., was not influenced by the size of the evaluated text.

5 Results of the experiment

The z-scores for each of the tested correlation and regression parameters are presented in Table 3.

TL/Text Type	*z* – *r* Corr BLEU/ WNM	*z* – Slope BLEU/ WNM	*z* – I'cept BLEU/ WNM
DE/em	0.5259 0.6434	**3.5060** **2.9228**	**-3.9990** **-3.1717**
DE/wp	1.3132 1.0120	-1.4980 -1.5400	-0.7850 -1.1440

TL/Text Type	*z* – *r* Corr BLEU/ WNM	*z* – Slope BLEU/ WNM	*z* – I'cept BLEU/ WNM
EN/em	0.5284 0.3051	0.5587 1.0127	-0.794 -0.8998
EN/wp	0.2015 -0.5130	-1.472 -1.519	-0.6160 -0.7390
ES/em	-1.6930 -1.1072	-1.4580 -1.3871	1.2276 1.1687
ES/wp	1.1749 0.4559	-0.7210 -0.8300	**-3.8920** **-3.7710**
FR/em	0.5462 0.2976	0.1745 0.1173	-0.4040 -0.2800
FR/wp	0.2523 0.1205	-1.3150 -1.3200	-0.7120 -0.9900
IT/em	0.36910 -0.1788	-0.7120 -1.0220	0.2308 0.6177
IT/wp	1.2575 0.6707	-1.1800 -1.3110	-1.6960 -1.3880
PT/em	0.6003 -0.0037	1.3657 0.8606	-1.5560 -1.0040
PT/wp	1.2338 0.8378	-0.6330 -0.6240	**-4.9350** **-5.5960**

Table 3. z-scores for correlation/regression

As an illustration, the differences between z-score moduli for WNM intercept are visualised in Table 4. The horizontal line shows the value corresponding to a confidence level of 99.9% (z=3.09). Visualisation of the z-scores for BLEU intercept scores will look similar.

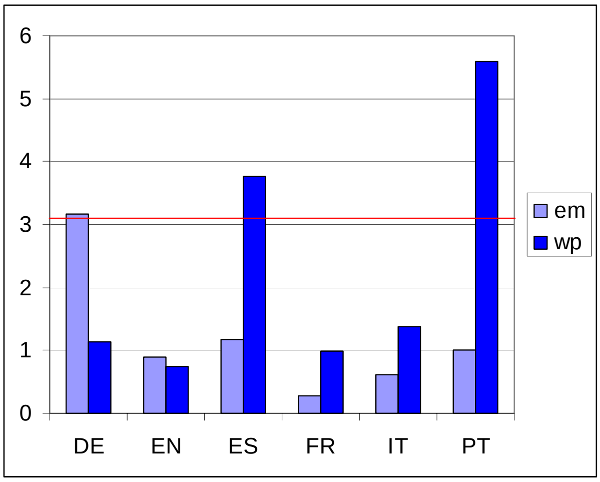

Table 4. Moduli of z-scores (WNM Intercept)

It can be seen from the Tables 3 and 4 that, for most parameters across the target languages and text types, the z-scores are smaller that 1.96; that is to say, the differences in such parameters can be attributed to variation that is typical for an evaluation corpus of a size of around 3,600 words. However, several parameters have z-scores higher than 2.576 (even higher than the next convenient 'confidence threshold' of 99.9% – 3.09). For these parameters the null-hypothesis must be rejected: their values are demonstrably influenced by the target language and text type.

First, note that for the Pearson's correlation coefficient *r* the z-scores for all target languages and text types are contained within the limits of the variation present in the French-English part of the DARPA corpus. The null-hypothesis for the *r* coefficient always holds, which confirms that for all evaluated target languages and text types the n-gram MT evaluation metrics can be used reliably, provided the user is interested only in correlation between the human scores and automated scores, e.g., for internal development purposes. Correlation is not influenced by the these 'external' factors, so higher automated n-gram scores will always indicate better quality in the eyes of human evaluators for all evaluated target languages and text types.

However, having reliable correlation figures is not sufficient to support predictions about human scores on the basis of automated scores (or predictions about the level of 'acceptability' of the output of a particular MT system for end-users). The additional parameters needed to make these predictions (such as the slope and the intercept of the regression line) are not stable across the target languages and text types, and are influenced by these 'real-world', evaluation-external factors.

Note that for target language German for emails, the z-scores for both parameters of the regression line stand out from the variation 'noise' for both n-gram metrics. The regression line for German emails is much steeper – a higher slope – and is moved down the *y* axis (the axis of human scores) – a lower intercept. This means that 'better quality' for human evaluators here needs a smaller number of n-gram matches, and that the improvement in human scores requires a much greater increase in the number of n-gram matches than is the case for the news texts in the French-English part of the DARPA corpus.

This does not hold for the whitepaper texts translated into German: here all the differences in the slope and the intercept of the regression line are within the variation limits of the French-English DARPA corpus.

Also note that for the whitepaper texts in the target languages Spanish and Portuguese, the intercept parameter of the regression line is also much lower than expected: here higher 'human' quality again relies on smaller number of n-gram matches. But the slope of the regression line is within the variation limits for both Spanish and Portuguese.

A surprising fact about these results is that regression parameters can be changed by the target language (possibly influenced by some language-specific features) or by text type, which from the point of view of MT evaluation may behave as a different language (or sub-language). The mechanism whereby such a language/sub-language influences the regression parameters is not clear, but it can be suggested that typological features (rather than genealogical factors) play an important role, since genealogically related languages (such as English and German or French/Italian and Spanish/Portuguese) often show differences in the parameters. An important factor could be the degree to which the target language is 'analytic' (relies on the use of free functional morphemes and syntactic means to express concepts) or 'synthetic' (more often uses fused functional morphemes and word formation for concepts). The difference in the degree of analytism may explain the differences in the parameters for French and Spanish whitepaper texts.

It should be also noted that within a particular language 'typological distance' between sub-languages (or text-types) could be different: it is intuitively plausible that the colloquial style of emails in German is very different from the style of legal documents, such as the whitepaper – in terms of lexicon and syntax – and that such a distance is possibly greater than between English or French emails and the whitepaper texts in those languages (cf. Kittredge, 1982). This could provide a clue as to why there is a difference in regression parameters across text types in German, but there is no such difference in English, French or Italian.

However, the most important and interesting result of our experiment is the very fact that the regression parameters do vary across text types and target languages (TLs), so they cannot be re-used for previously untested combinations of TLs/text-types. This means that knowing the regression line parameters for a certain combination of these evaluation-external factors is not helpful for predicting human evaluation scores or the acceptability of an MT system for some other combination. In order to predict these values, one needs to carry out (relatively expensive) human evaluations for every TL/text-type combination for which there is a demand to predict human evaluation scores from automated n-gram-based scores.

There is still an open question whether the TL and the text-type are the only factors which influence the parameters of the regression line. If this is the case, 'calibration' of human scores needs to be done only once for each TL/text-type combination by computing the parameters of slope and intercept on a larger corpus. Furthermore, these parameters can be re-used for the reliable prediction of human evaluation scores within the same TL/text-type combinations.

However, other external factors may also influence the regression parameters, e.g., the architecture of the evaluated MT system (statistical, example-based, rule-based, etc.) or the source language. Further experiments are needed to estimate if they have any effect on the prediction of human scores.

6 Conclusion and future work

We carried out a large-scale MT evaluation experiment for a number of languages and text types which had not been the subject of automated MT evaluation. The experiment involved generating human scores for adequacy and two sets of automated evaluation scores (BLEU and WNM), computing correlation and regression parameters between human and automated scores, and predicting the acceptability of the output of particular MT systems for different target languages and text types. We established experimentally the threshold of acceptability of the output at 3.5 on the 5 point scale used by the human judges, and mapped this threshold to the automated scores for each combination of text type and target language.

The analysis of this data involved measuring the difference between the correlation/regression parameters in our newly evaluated texts and in the gold standard DARPA 94 MT evaluation corpus. The principal findings are that the correlation figures for all target languages and text types are always reliably within the expected variation limits, so it can be expected that the correlation between human and automated n-gram metrics for all the evaluated target languages and sub-languages will be equally high. So the metrics can be reliably used for internal system development for all evaluated target languages.

However, end users of MT systems often need to estimate the level of acceptability of a particular MT system on the basis of automated MT evaluation scores, i.e., to predict human evaluation scores for the system on the basis of the automated scores. This task requires estimating regression parameters – the slope and the intercept of the regression line. Our results suggest that, unlike the correlation coefficient, these regression parameters may be specific to some languages and text types. Consequently, human evaluation scores for each new TL/text-type combination will still be necessary for making reliable predictions about human evaluation scores for new texts and MT systems. Absolute values of BLEU and WNM (which eventually come down to the number of n-gram matches) are sensitive to such evaluation-external factors and therefore their predictive power is ‘local’ to a particular language or a sub-language (text type). In the general case, the number of n-gram matches cannot give a ‘universal’ prediction of human quality perceptions.

Future work will involve accounting for the influence of other possible factors on the regression parameters (e.g., source language) and extending the number of evaluated target languages.

References

B. Babych. 2004. Weighted N-gram model for evaluating Machine Translation output. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*. M. Lee, ed., University of Birmingham, January 2004. pp. 15-22.

Babych B., Hartley A. 2004a. Extending the BLEU MT Evaluation Method with Frequency Weightings. In *ACL 2004 Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, July 2004. pp. 622-629.

Babych B., Hartley A. 2004b. Modelling legitimate translation variation for automatic evaluation of MT quality. In *LREC 2004 Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, May 2004. pp. 833-836.

R. Kittredge. 1982. Variation and homogeneity of sublanguages. In R. Kittredge and J. Lehrberger (eds). *Sublanguages: studies of language in restricted semantic domains.* deGruyter. pp. 107-137.

K. Papineni, S. Roukos, T. Ward, W-J Zhu. 2002 BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 311-318.

J. White, T. O’Connell and F. O’Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons and future approaches. In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*. Columbia, MD, October 1994. pp. 193-205.