Sensitivity of automated MT evaluation metrics on higher quality MT output: BLEU vs task-based evaluation methods

Bogdan Babych, Anthony Hartley

Centre for Translation Studies, University of Leeds, UK

E-mail: b.babych@leeds.ac.uk, a.hartley@leeds.ac.uk

Abstract

We report the results of an experiment to assess the ability of automated MT evaluation metrics to remain sensitive to variations in MT quality as the average quality of the compared systems goes up. We compare two groups of metrics: those which measure the proximity of MT output to some reference translation, and those which evaluate the performance of some automated process on degraded MT output. The experiment shows that proximity-based metrics (such as BLEU) loose sensitivity as the scores go up, but performance-based metrics (e.g., Named Entity recognition from MT output) remain sensitive across the scale. We suggest a model for explaining this result, which attributes the stable sensitivity of performance-based metrics to measuring the cumulative functional effect of different language levels, while proximity-based metrics measure structural matches at a lexical level only and therefore miss higher-level errors that are more typical for better MT systems. Development of new automated metrics should take into account the possible decline in sensitivity for higher-quality MT, which should be tested as part of meta-evaluation of the metrics.

1 Introduction

Automated MT evaluation metrics, such as BLEU or NIST, compute numerical scores that characterise certain aspects of machine translation quality. The accuracy of these metrics is usually verified by correlations between a range of automated scores and scores given by human judges, who typically assess two aspects of quality: Adequacy (how much information from the original is preserved in the output) or Fluency (how natural and comprehensible the translated texts sounds in the target language).

Automated metrics are used not only for assessing the achieved level of MT performance, but also for optimising system parameters during development. Therefore, improving the quality of automated metrics can lead directly to improvement in MT output.

Since human scores are difficult and expensive to obtain, it is possible to reliably validate automated MT evaluation metrics only under certain restrictions, for example: by linguistic variables (text type, genre, target language); by granularity of evaluated units (sentence, text, corpus); by system characteristics included in particular evaluation (different versions of the same system developed over time, or a collection of different MT systems having the same architecture, or a heterogeneous collection which includes both SMT and RBMT).

An automated metric may subsequently be used under very different conditions, for which its accuracy has not been tested. Users then need to make the assumption that the metric will be as reliable under these new conditions as under the tested conditions. However, this assumption often does not hold. For example, for test sets which include both RBMT and SMT systems the correlation of BLEU/NIST type scores is lower, since these metrics overestimate the Adequacy of SMT output (Callison-Burch et al., 2006). Other features are also dependent on experimental conditions: e.g., the regression figures which predict human scores using automated scores (the slope and the intercept of the fitted line) depend on the combination of text type and target language (Babych et al., 2005). Projecting automated scores onto human scores is important if we are interested in the acceptability of MT output at or above some known threshold given by human scores.

Therefore, it is essential to know the limits of each automated MT evaluation metric and the conditions beyond which high correlation with human judgements cannot be assured. Awareness of such limits leads to more careful application of automated methods, to more conscious selection of metrics for specific tasks, and eventually to a deeper understanding of MT evaluation and of translation itself.

This paper explores one of these limits: sensitivity of automated metrics in different quality ranges. The problem is whether all types of automated MT evaluation metric maintain a high correlation with human judgements – that is, their *sensitivity* in distinguishing between better and worse translations – in the case where the average quality of the evaluated systems goes up. This problem is important because, with time, the general quality of MT will gradually yet substantially improve (Thurmair, 2007). Consequently, if a certain class of automated evaluation metrics is less sensitive for texts produced by higher quality systems, the value of such metrics will degrade with time. More importantly, the usefulness of these metrics for optimising systems parameters will hit inherent limits, i.e., at some point the developers will no longer be able to "continue to use BLEU to further improve (their) models and systems" (Marcu et al., 2006). After reaching a certain quality level the systems may cease to be reliably guided by this metric, since human judgments about any further improvements in MT quality and the metric's view may differ substantially.

Our paper examines two MT evaluation metrics, representing two different types of evaluation method: the BLEU metric exemplifies distance-based evaluation, while a method based on Named-Entity recognition in degraded MT output exemplifies task-based evaluation.

2 Distance-based and task-based MT evaluation models

As suggested in (Popescu-Belis, 2007), the majority of MT evaluation systems can be grouped around two central principles used to score MT output: *distance-based* metrics compute some sort of distance between MT output and a gold-standard human translation (e.g., edit distance, N-gram distance), while *task-based* metrics measure performance of an automated process or system on degraded MT output, assuming that the degree of this degradation is proportional to any decline in the system's performance.

A prototypical example of a distance-based metric is BLEU, but the following assumption behind it characterises all other metrics in this group: "…the closer the machine translation is to a professional human translation, the better it is" (Papineni et al., 2002). Distance-based metrics are now most widely used; however, their central problem is that legitimate variation in the gold standard: mismatches between MT output and a human translation will be treated as MT errors rather than as legitimate alternatives. As a result, these metrics have a relatively low correlation with human judgments at the level of smaller segments (sentences and individual texts), so they are not very useful for automating error analysis in MT. For example, (Babych et al. 2007b) show that BLEU converges with human scores only after the evaluated corpus reaches some 7,000 words in size.

Task-based metrics can work without a human reference translation. Their principle was initially suggested for human evaluation: "…can someone using the translation carry out the instructions as well as someone using the original?" (Hutchins and Somers, 1992: 163), and has been successfully applied in automated metrics. One of the first examples of automated task-based metrics is the X-score suggested in (Rajman and Hartley, 2001). It is computed by running the Xerox shallow dependency parser XELDA on MT output. Note that behind task-based methods there is an implicit assumption related to the redundancy of natural languages: MT errors more frequently destroy the contextual conditions which trigger rule application in down-stream automated systems, but they rarely create spurious conditions for recognising such phenomena. For example, the X-score rewards the ability of MT to preserve sentence-level dependencies, like the relation between a predicate of a subordinate clause and an antecedent of its pronominal subject in the main clause: *a **hearing** that **lasted** more than two hours*. On the other hand, some simple local relations are frequently over-generated by MT, e.g. adverb-adjective dependency: ***brightly colored** doors*. In this case the score should reward a system for its ability to avoid such over-generation. However, the type of phenomena which can characterise MT quality for a given language pair can only be established experimentally.

(Babych and Hartley 2004b) proposed an evaluation method based on Named Entity (NE) recognition in MT output, using the ANNIE open-source NE recognition system available in GATE (Cunningham et al. 2002). The idea that certain types of NEs, such as Organisation Names, rely for their identification on complex linguistic contexts which can be easily destroyed by imperfect MT, as illustrated in Example (1).

(Ex. 1) French original: ... *le chef de la diplomatie égyptienne*

(1.1) Human translation: *the <Title>Chief</Title> of the <Organization>**Egyptian Diplomatic Corps** </Organization>*

(1.2) MT output: *the <JobTitle> chief </JobTitle> of the **Egyptian diplomacy***

The relevant context in 1.2 is destroyed by MT and cannot trigger annotation of the Organisation Name. In general, the number of extracted Organisation Names correlates with human judgments about the Adequacy of the MT system. Interestingly only the annotation of this type of NE produced by ANNIE has such discriminative power.

Hybrid methods combine both models; for example, within a performance-based model we can measure the performance of the MT system itself using a distance metric (e.g., BLEU) on texts with varying difficulty. Here a *difficulty slope* parameter is computed which relates performance of a tested system against some reference system (Babych et al. 2007b). This method was suggested for comparing different pivot MT architectures against a direct translation route (used as a reference); it shows how systems cope with increasing difficulty of segments or texts.

3 Sensitivity of automated evaluation metrics

It is useful to distinguish two dimensions of MT quality: (A) there are stronger and weaker systems; (B) there are easier and more difficult texts or segments. This distinction shows that any *absolute* interpretation of automated evaluation scores (e.g., BLEU) is impractical, because these scores need to be weighted by an independent measure of text difficulty for MT, which has yet to be discovered (Babych et al. 2004a). The automated scores make sense only in comparison to each other, and their absolute values cannot be supposed meaningful.

However, a desired feature of automated metrics (on dimension B) would be to distinguish correctly the quality of different sections of a corpus translated by the same MT system. We define ***sensitivity*** as the ability of a metric to predict human scores for different sections of an evaluation corpus, such that if some sections receive higher human scores, the metric also consistently rates them higher. For any given metric an important question is whether dimensions A and B are independent or whether sensitivity (B) depends on overall system quality (A), in other words, whether sensitivity changes in different areas of the quality scale. For any automated metric it is desirable to minimise any such dependency.

Varying sensitivity is a possible limitation on the usefulness of automated metrics: if sensitivity declines in a certain area of the scale, then automated scores become less meaningful and less reliable, both for comparing easier and more difficult segments and for distinguishing between better and worse systems.

4 Experiment set-up: dependency between sensitivity and quality

At the first stage the task was to compute sensitivity figures covering different areas on the Adequacy scale, so we used a range of systems with different human scores

for Adequacy. There were four MT systems (one SMT, three RBMT) and one human translation from the DARPA-94 corpus (White et al., 1994). The corpus consists of MT output for 100 texts (i.e. "sections") with corresponding human scores for Adequacy. Sensitivity was approximated as the correlation between automated scores and human scores for the same text; high correlation means high sensitivity.

We applied two types of MT evaluation metric: the distance-based metric BLEU, and a task-based method of NE recognition in MT output.

At the second stage we observed the dependency between the metrics' sensitivity for each MT system and average system quality. This was done by correlating corpus-level figures for sensitivity from stage 1 and average human scores (again, over the whole corpus) for Adequacy. Here high correlation is not desirable: it means that sensitivity is dependent on quality and, in the case of high negative correlation, that the sensitivity of a metric declines as quality goes up. Low correlation, on the contrary, is a welcome feature, since it shows that a metric's sensitivity is homogeneous across the whole quality range.

Figure 2 suggests a compact representation of our experiment set-up. The formula describes the order of the stages, the computations performed and data used. Arguments for the operations are given in brackets, or in enumerator and denominator in the formula. Arguments with an initial upper-case letter are independent variables, while arguments with an initial lower-case letter are fixed parameters.

$$rCorrel(System)\left[\frac{humanScore(ade)}{rCorrel(Text)\left[\frac{humanScore(ade)}{bleuScore(n4r1) \vee neGate(organisation)}\right]}\right]$$

Figure 2. Experiment set-up

5 Results of the experiment

Table 1 compares the BLEU and GATE NE recognition metrics with respect to two parameters: row 1 presents the correlation between automated and human scores for each metric, a standard way to evaluate automated metrics; row 2 presents the proposed sensitivity scores, as described in the previous section.

	BLEU/ade	GATE/ade
system-correl	0.95	0.87
sensitivity-correl	**-0.76**	**-0.22**

Table 1. Metrics' correlation and sensitivity

It can be seen from the table that BLEU outperforms GATE in terms of system correlation. This may be due to the fact that there are many fewer Organisation Names in the evaluation corpus than N-gram matches, so BLEU benefits from having more data.

However, on the second parameter – sensitivity correlation – the GATE NE recognition method is much better: for BLEU there is a high negative correlation between sensitivity and quality, whereas the GATE metric's sensitivity does not degrade when MT output quality improves.

Figures 3 and 4 show data points for the sensitivity of both metrics for all four MT systems (the rightmost data point in both figures being the value for the human translation).

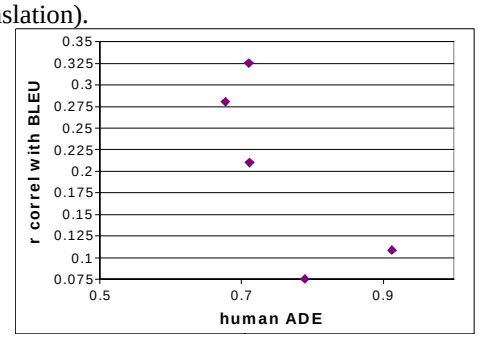

Figure 3. Sensitivity points for BLEU

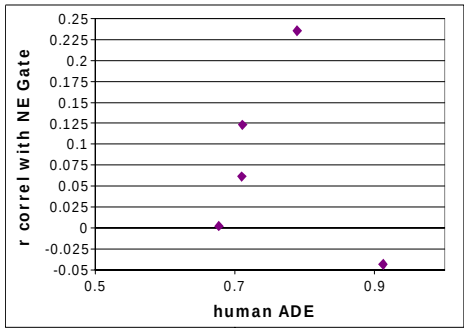

Figure 4. Sensitivity points for ANNIE

Figure 3 demonstrates the general tendency of BLEU scores to loose their sensitivity towards the higher end of the scale: for better MT systems BLEU becomes less sensitive at the text level. Figure 4 shows that there is no such tendency for task-based evaluation via NE recognition.

6 Discussion

A possible interpretation of this fact relies on the differing nature of distance-based and task-based evaluation methodologies. Distance-based metrics, such as BLEU, use *structural models* for MT quality: they focus on features coming from one particular level of language structure (in the case of BLEU – the lexical level) and therefore are not sensitive to errors at higher levels. However, in higher-quality MT systems these lexical issues are usually resolved, so the distribution of errors is shifted towards the textual level, e.g., textual cohesion and coherence, long-distance agreement in syntactic structures. The presence of such higher-level errors distinguishes the worse-translated from the better-translated segments, but BLEU is not sensitive to such types of errors beyond the lexical level.

On the other hand, task-based evaluation methods use *functional models* for MT quality, taking an external view on the structure and interaction of specific levels. They assess how the text or specific contexts within the text perform some function that is external to the structure. Therefore, task-based evaluation can potentially capture degradation at any structural level which contributes to this function. In particular, task-based metrics should better capture legitimate variation, since they do not make explicit assumptions about particular combinations of structural features that perform external textual functions.

This interpretation can be illustrated by Figure 5, which suggests that the major advantage of task-based MT evaluation metrics is that the adoption of a more appropriate functional perspective on linguistic phenomena preserves sensitivity to errors at all levels and across the whole range of evaluation scores.

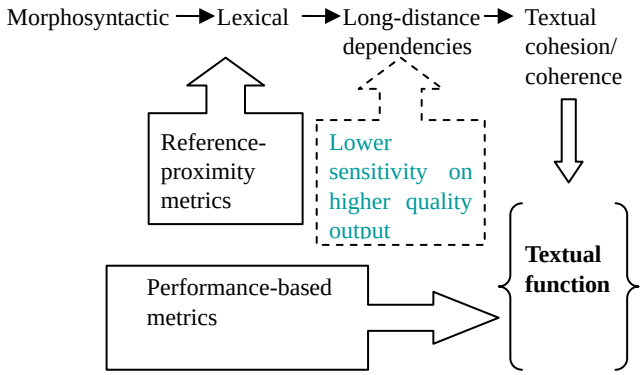

Figure 5. Interpretation for sensitivity loss of distance-based metrics

7 Conclusions and future work

Degradation in sensitivity of automated evaluation metrics for higher-quality MT systems is a major limitation of MT evaluation technology: it can seriously influence the reliability of predictions for human scores. As a distance-based metric, BLEU shows signs of such a degradation in sensitivity. On the other hand, functional models which work at the textual level minimize the dependency of a metric's sensitivity on MT system quality. Future work will include developing more adequate functional models for task-based evaluation, for example, models of non-local information in the text, such as textual cohesion and coherence.

Acknowledgements

This work was partially supported by the Leverhulme Trust.

References

Babych, Bogdan and Hartley, Anthony. (2004). Comparative Evaluation of Automatic Named Entity Recognition from Machine Translation Output. In *IJCNLP Workshop on Named Entity Recognition for Natural Language Processing Applications*.

Babych, Bogdan and Elliott, Debbie and Hartley, Anthony. (2004). Extending MT evaluation tools with translation complexity metrics. In *CoLing 2004: 20th International Conference on Computational Linguistics.*

Babych, Bogdan and Hartley, Anthony and Elliott, Debbie. (2005) Estimating the predictive power of n-gram MT evaluation metrics across language and text types. In *MT Summit X*.

Babych, Bogdan and Hartley, Anthony and Sharoff, Serge. (2007b). Translating from under-resourced languages: comparing direct transfer against pivot translation. In *MT Summit XI, Copenhagen, Denmark.*

Callison-Burch, Chris and Osborne, Miles and Koehn, Philippe. Re-evaluation the Role of Bleu in Machine Translation Research. (2006). In *11th Conference of the European Chapter of the Association for Computational Linguistics) (EACL06)*

Cunningham, H. and Maynard, D and Bontcheva, K. and Tablan V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia. 2002*.

Hutchins,W.John and Somers, Harold. (1992). An introduction to machine translation. London: Academic Press.

Marcu, Daniel and Wang, Wei and Echihabi, Abdessamad and Knight, Kevin. (2006). SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.*

Papineni Kishore and Roukos Salim and Ward Todd and Zhu WeiJing. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL 2002: the 40th Annual Meeting of the Association for Computational Linguistics.*

Popescu-Belis, Andrei. (2007). The place of automatic evaluation metrics in external quality models for machine translation. *In MT Summit XI Workshop: Automatic procedures in MT evaluation, Copenhagen.*

Rajman Martin and Hartley Anthony. (2001). Automatically predicting MT systems ranking compatible with fluency adequacy and Informativeness scores. In*: 4th ISLE Workshop on MT Evaluation, MT Summit VIII*

Thurmair, Gregor. (2007). Automatic evaluation in MT system production. In *MT Summit XI Workshop: Automatic procedures in MT evaluation, Copenhagen.*

White J. and O'Connell T. and O'Mara F. (1994). The ARPA MT evaluation methodologies: evolution lessons and future approaches. In *1st Conference of the Association for Machine Translation in the Americas.*