Correlating Automated and Human Assessments of Machine Translation Quality

Deborah Coughlin

Microsoft Research
Redmond Washington, USA
deborahc@microsoft.com

Abstract

We describe a large-scale investigation of the correlation between human judgments of machine translation quality and the automated metrics that are increasingly used to drive progress in the field. We compare the results of 124 human evaluations of machine translated sentences to the scores generated by two automatic evaluation metrics (BLEU and NIST). When datasets are held constant or file size is sufficiently large, BLEU and NIST scores closely parallel human judgments. Surprisingly, this was true even though these scores were calculated using just one human reference. We suggest that when human evaluators are forced to make decisions without sufficient context or domain expertise, they fall back on strategies that are not unlike determining n-gram precision.

1 Introduction

Developing a machine translation (MT) system involves a fine balancing act. Each improvement must be reviewed to rule out unexpected interactions with prior development efforts. Traditionally, human evaluation has been the only means of providing the necessary feedback to keep development moving forward. But human evaluation has serious drawbacks: in addition to relying on subjective judgments, it is both time-consuming and costly. This in turn means that the scale of these evaluations tends to be so small -- usually no more than a few hundred sentences examined by a small number of raters – that it can be difficult to draw firm conclusions about system quality.

In recent years, these problems have spurred the development of several automatic evaluation metrics. The speed, convenience, and economic advantages of these metrics have led to their quick adoption by the field. A number of techniques have been proposed as a means of evaluating MT sentence quality, including using named entity translations as an indicator of overall quality (Heirschman et al, 2000), using decision trees to classify machine- vs. human-generated sentences (Corston-Oliver et al, 2002), and using existing manually evaluated translations to extrapolate the quality of new translations (Nieβen et al, 2002). The dominant approach, though, involves computing the closeness of a machine translated sentence to one or several reference sentences (Papineni et al, 2001; Doddington, 2002b; Nieβen et al, 2000). Papineni's BLEU (Bilingual Evaluation Understudy) and Doddington's related NIST metric are two in common use today.

Since practical considerations have forced the field to rely on automated metrics, it is crucial to determine how well these metrics compare to human judgments. The remainder of this paper explores the relationship between a large number of carefully-collected human judgments of MT sentence quality and the automated measures of BLEU and NIST. We conclude that for many purposes, automated metrics can indeed replace human raters. Of particular interest is our finding these metrics can be highly reliable even when only one reference translation is available.

1.1 Human Evaluation

The data used in this investigation was collected over a period of 2 years. Evaluations were conducted by the Butler Hill Group, an independent vendor agency, in part to ensure that none of the human raters involved had any role in building the MT system they were rating. The data reported here reflect 124 evaluations involving multiple language pairs.

14	English =>German (EG)	4	German => English (GE)
36	English =>Spanish (ES)	38	Spanish => English (SE)
20	French => English (FE)	8	Hansards French => English (QE)
4	French => Spanish (FS)		

Each of the 124 evaluations resulted in an absolute quality judgment of an MT system's output. These judgments were done in pairs, allowing us to contrast the performance of our MT system (MSR-MT) with a comparison system (either an older version of the MSR-MT system or a commercial system[1]) (Richardson, 2001). Each pair of evaluations was conducted on a fresh set of randomly selected sentences, which were blind to the system developers. In most of what follows we will treat the 124 evaluations as separate data points, but in Section 2.1.1 we consider the difference between absolute judgments of quality for two MT systems scored on the same dataset.

Raters judged the acceptability of translations on a scale of 1 to 4:

4 = **Ideal**: Not necessarily a perfect translation, but grammatically correct, and with all information accurately transferred.
3 = **Acceptable**: Not perfect (stylistically or grammatically odd), but definitely comprehensible, AND with accurate transfer of all important information.
2 = **Possibly Acceptable**: Possibly comprehensible (given enough context and/or time to work it out); some information transferred accurately.
1 = **Unacceptable**: Absolutely not comprehensible and/or little or no information transferred accurately.

Raters were given no instructions beyond the scale above regarding how they should make their judgments. Leaving these decisions up to the rater meant we were unable to separate fluency and adequacy judgments as others have done (Doddington, 2002b; Papineni et al, 2001; Reeder, 2001, etc). It also freed us from having to determine the importance of one characteristic over another when deciding what acceptable quality was. Raters balanced these different characteristics as they saw fit, yet determined "acceptable" with high inter-rater agreement.

Reference sentences were the product of domain-expert human translators working from corresponding source sentences in context. Using reference sentences as the gold standard allowed evaluations to be entirely monolingual. Except for the 8 Hansards French => English evaluations, all evaluations were conducted on highly technical corpora (Microsoft computer manuals and support documents). Evaluators were not domain experts and did not have access to context beyond the sentence.

Average sentence length for the computer domain (computed on the English data) was 17.5; it was 21 for the Hansards data.

Between 4 and 7 raters (predominately 6-7) were used for each evaluation, and results reported here are based on averaging the scores of all evaluators. Human evaluators bring with them individual differences; differences in intelligence, reading ability, professional training, etc. It is hoped that taking the mean of multiple evaluator scores reduces the effect of these differences.

Rating translation quality is both tedious and repetitive. It is presumed that evaluators have different tolerances for these working conditions. There is evidence to suggest that evaluations are somewhat compromised by evaluator speed or inattention. In a small percentage of cases (for example, 1% in ES), individual evaluators gave different scores to identical sentences (presented side by side) or did not give the top score (4) to sentences that were identical to the reference.

Though the average inter-rater agreement (correlation) is strong (EG: 0.763; QE: 0.830), individual raters occasionally display a low inter-rater correlation. The correlation coefficient for one EG rater correlation was 0.59. Correlations between human assessments and BLEU/NIST scores reported below include the average of all evaluators. Evaluators with low inter-rater agreement were not excluded.

In addition to variations introduced by individual raters, the choice of corpus also impacts quality judgments, even within the same domain. In the computer domain, for instance, translations of manuals consistently received lower scores than product support translations, even when the system was held constant.

1.2 Automatic Evaluation

IBM's BLEU is a modified n-gram precision measure. It uses a weighted geometric average of n-gram matches between test sentences and reference sentences, which is modified to penalize overgeneration of correct word forms. Also included is a multiplicative brevity penalty that penalizes test sentences found to be shorter than the reference sentences; this is computed at the corpus level. The resulting score is a numeric metric intended to indicate the closeness of a set of test sen-

[1] One of Sail Labs, Systran, BabelFish, or Lernout & Hauspie. In each case, an attempt was made to pick the strongest comparison system.

tences to a corresponding set of reference sentences.

The n-gram co-occurrence metric (NIST) developed by the National Institute of Standards and Technology is based on the BLEU metric. For a complete description and comparison to BLEU, see Doddington (2002b). One primary difference between the two is that NIST uses an arithmetic n-gram average while BLEU relies on a geometric average. We show that this difference is significant when dealing with very low-quality translations (discussed below: Section 2.2). The two algorithms also differ in how they calculate their respective brevity penalties.

We discovered that when using NIST scores in evaluations, corpus size must be kept constant because NIST scores increase logarithmically with corpus size. We verified this by running NIST over files of varying size which contained test sentences identical to the reference sentences. BLEU, on the other hand, returned the top score in all these tests, suggesting that it is not affected by corpus size. To compensate, comparisons with NIST are limited to evaluations of 250 sentences (the most common size). Comparisons with BLEU include all evaluations for the reported language pairs.

Before delving into the detailed results of our comparison, we must note one important difference between our evaluation methodology and the NIST/BLEU philosophy. In all of the human evaluations reported here, raters were presented with just a single reference sentence. This reliance on a single reference sentence runs counter to the spirit of BLEU/NIST, which assume that there can be many correct translations for any one source sentences. The ability to weigh multiple "correct" answers simultaneously in calculating a quality score is a clear strength of these algorithms.

Yet it is in the interests of the field to look for correlations even when there is just one reference sentence available. Like much of the bilingual data available, ours is in a highly specialized domain, so acquiring multiple translations would require equally specialized (and expensive) translators. We have also chosen to use a fresh set of sentences for each evaluation so that developers can have access to 500 or more sentences pre-categorized by quality score. This approach would become prohibitively expensive even with an automated evaluation metric, if that metric requires generating multiple reference translations for each new set of evaluation sentences.

Papineni et al (2001) discuss work on systematically testing the use of one reference sentence (multiple translators) per source sentence within the context of BLEU; see also Doddington (2002b) for similar work with NIST. This preliminary work suggests that even with just one reference sentence, BLEU/NIST can produce meaningful results.

Our reference sentence data is not entirely homogenous, which may give it some of the characteristics of a dataset involving multiple translations per source sentence. The original dataset, consisting of hundreds of thousands of translated sentences, was created by multiple translators employed by multiple vendor agencies. This data was randomly split into training/test sets, and each evaluation involved sentences randomly selected from the held-out test data. Given this methodology, we can assume that each evaluation included translations from many translators.

Using the dataset described above, we are unable to systematically vary the number of translators, as Papineni did in his preliminary work. Our contribution is to test whether his findings would generalize to far-larger datasets where multiple translators are assumed.

2 Human vs. BLEU/NIST

2.1 Data: Standard Evaluations (Random Sentences)

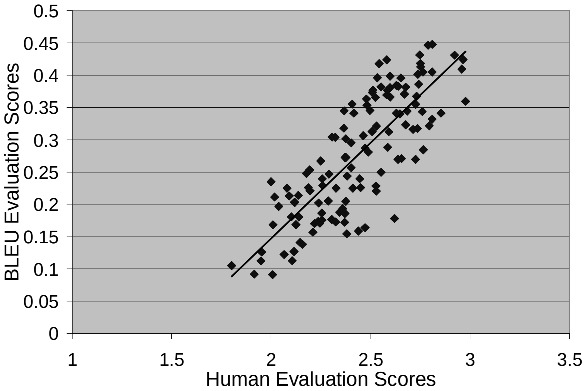

Figure 1: BLEU vs. Human; 124 250-400 sentence evaluations; 7 language pairs; 6 MT systems.

We begin our study by comparing BLEU and NIST to the results of 124 human evaluations of MT quality for 7 language pairs (Figure 1). The output from 6 different MT systems (MSR-MT and

4 commercial systems) is included in this figure. (Refer to section 1.1 to view the composition of the 124 evaluations included in this study.) Though we find a fairly even linear band of data points, the width of that band indicates that BLEU can make only the grossest distinctions when the dataset is limited to files of predominately 250 sentences each. The BLEU correlation coefficient[2] is 0.811. For NIST the chart is quite similar, displaying a wide linear band of data points for 104 evaluations of 250 sentences each. The NIST correlation coefficient is 0.796.

BLEU and NIST are highly correlated (0.993) when compared to each other (Figure 2).

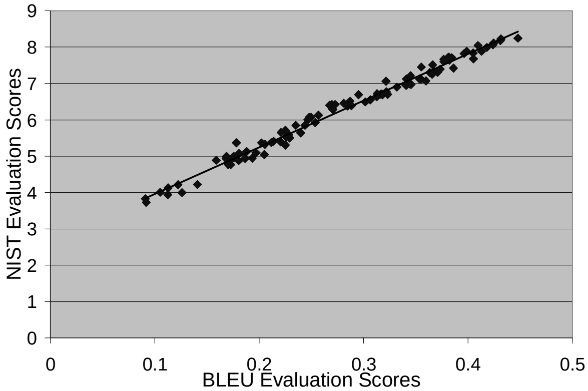

Figure 2: BLEU score vs. NIST score; 104 250 sentence evaluations.

Correlations for individual language pairs are no better than in the aggregate, as shown in Table 1.

Lang. Pair	# of evals	Human-BLEU correlation	# of evals	Human-NIST correlation
ALL	124	0.811	104	0.796
EG	14	0.906	10	0.973
ES	36	0.826	30	0.815
SE	38	0.755	34	0.756
FE	20	0.711	16	0.650
QE	8	-0.130	4	-0.500

Table 1: Number of evaluations; correlation coefficient for human assessments compared to BLEU/NIST.

The row marked "ALL" includes language pairs not listed; for GE and FS only a small number of evaluations were performed. Though in the case of EG the correlation is fairly strong (see also Figure 3), these findings suggest that BLEU/NIST are of limited use as a substitute for human assessment in tracking incremental improvements when the test set is relatively small and not held constant across systems.

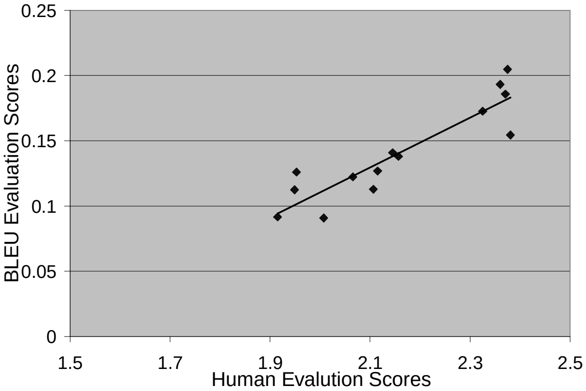

Figure 3: EG Human scores vs. BLEU scores; 14 evaluations

2.1.1 Relative Assessments

Given that BLEU and NIST do not closely correlate with human assessments when datasets are predominately 250 sentences in size, it is remarkable that when viewing the paired tests (current system vs. comparison) on identical sentences, BLEU and NIST parallel human assessments. Of the 62 paired tests examined, BLEU agreed with human assessments of which system was superior in 59 cases, or 95% of the time. This was true in several tests where the difference was not statistically significant. In 2 of the 3 tests in which BLEU disagreed with the human assessments, the difference between the human scores for the test system and the comparison was less than 0.01 (0.008 and 0.003). In all three cases, at least one evaluator was not in agreement with the overall preferred system. The 0.95 confidence intervals for the human evaluation scores ranged from +/- 0.27 to 0.32.

NIST disagreed with these same three evaluation pairs, as well as one additional one. This last case is surprising, since the difference between the two absolute scores for this pair of evaluations was dramatic (0.361). The human score for the test system was 2.01 and comparison system score was 2.37. The 0.95 confidence interval was between +/- 0.29 and 0.33. Yet NIST gave both systems similar scores (0.07 difference), slightly preferring the test system system. BLEU and NIST performed similarly except in this one instance.

[2] Pearson product moment correlation has been used throughout.

Not only did BLEU and NIST agree with human evaluators with regard to the preferred systems, but they also correlate well with regard to the magnitude of this preference. We calculated the difference between the human evaluation scores of the paired systems and plotted them against the difference between BLEU/NIST scores for the paired systems. The results are reported in Table 2.

We conclude from these numbers that BLEU, and to a lesser extend NIST, can accurately determine the relative standing of two systems measured against the same corpus. In addition, these automated metrics provide a rough but reliable indication of the magnitude of the difference between the two systems. This is a significant and exciting finding.

Lang. Pair	# paired evals	Human diffs BLEU diffs	# paired evals	Human diffs NIST diffs
ALL	62	0.853	52	0.851
EG	7	0.776	5	0.846
ES	18	0.935	15	0.908
SE	19	0.948	17	0.930
FE	10	0.842	8	0.819
QE	4	0.619		

Table 2: Correlation coefficient resulting from comparing the difference between human scores on two systems (current and comparison) tested on the same sentences to the difference between BLEU/NIST scores on the same set of evaluations.

2.1.2 Hansards Domain

The weakest correlations in our dataset are for the Hansards French => English (QE) evaluations. The correlation coefficient for direct comparisons of human evaluation scores to BLEU scores, shown in Table 1, actually indicates a negative slope. When correlating the difference between the preferred system score and its paired system's score for human and BLEU evaluations, there is a moderate correlation (0.619), but this correlation is substantially lower than those in the computer domain. We can speculate on the cause of this phenomenon, but the fact that only 4 pairs of QE evaluations were available makes it difficult to draw firm conclusions. Our working hypothesis is that our computer domain data represent a much tighter domain than parliamentary debates, where topics range from domestic issues to military concerns, and where rhetorical style ranges from formally written speeches to rambling rants. There is a much more limited range of acceptable translations for our computer domain data, where editors strive to achieve a consistent tone, style, and set of lexical choices. This may also explain why a single reference sentence in this computer corpus provides enough information to BLEU and NIST to allow a high correlation with human judgments.

2.2 Data: Grouped by Unigram Score

We suspect that our standard evaluation size (250 sentences) is a limiting factor in determining BLEU's ability to correlate with our human evaluations (Section 2.1). We also did not want to limit our view of BLEU to evaluation sets that have scores concentrating in the 2 to 3 range. We looked, therefore, for a metric that would allow us to group each language pair's complete set of evaluated sentences into test sets likely to represent a full range of quality scores. We could not choose BLEU itself because BLEU has components that are calculated at the corpus level, making it impossible to calculate meaningful BLEU scores for each sentence[3]. We chose to group sentences by their unigram scores.

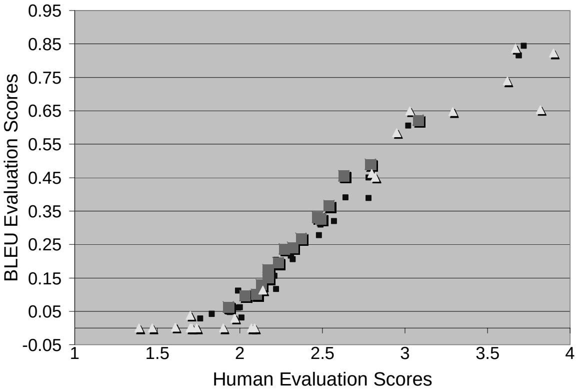

Figure 4: All language pairs combined; BLEU and human scores compared; test files created using average unigram score. Large squares represent file size over 500 sentences; small squares, 100-500 sentences; triangles, less than 100 sentences.

We then compared the correlation between human assessments and BLEU scores on these sentences grouped by unigram score. For all language pairs combined, the results are quite linear, with a correlation coefficient of 0.972 (Figure 4). The scatter plots for individual language pairs have a similar, strongly linear shape and have correlation coefficients ranging from 0.967 to 0.995.

[3] A three word sentence that is identical to the reference sentence would receive a BLEU score of zero.

BLEU gives a score of 0.0 to several files at the lower end. File sizes for those sentence files receiving a 0.0 BLEU score are quite small (2 to 42 sentences). For FS, the files with low unigram scores (0.0, 0.1, 0.2) received a BLEU score of 0.0. Human scores are above 1.0 (the lowest possible human score) in these cases. Though we cannot compare NIST scores because the file sizes are not the same, NIST gives each of these files a small but positive score.

This inability to distinguish between very low scoring corpora suggests a weakness in the BLEU metric. Because BLEU uses the geometric mean of n-gram scores at its foundation, it does a poor job of scoring files whose sentences are so poorly (or freely) translated that they share no trigrams or 4-grams with the reference sentences. Since none of the MT systems we tested produced data of this sort, this limitation in the scoring algorithm will not concern us further.

For datasets of over 500 sentences (Figure 4, squares), the correlation coefficient is extremely high: 0.993. This suggests that BLEU correlates strongly with human assessments when the file size is larger than 500 sentences. These datasets contained from 541 to 1872 sentences.

The strength of the positive linear correlation of human and BLEU scores indicates that BLEU can effectively distinguish sentence sets of different quality. This adds further weight to the finding (Section 2.1.1) that BLEU can reliably gauge the relative quality of two systems tested on the same corpus.

2.3 Data: Grouped by Average Human Score

For this next view on the data we grouped sentences in each language pair by their average human score, rounding to the nearest tenth. (Recall that human evaluators assigned an integer from 1 (poor) to 4 (ideal) to each translation.) This provided data covering the full range of human scores (1.0 to 4.0).

We computed the BLEU score for each of these files. Figure 5 shows the results of the language pair with the largest corpus of sentences (ES: 8390 sentences). Notice the highly linear middle section with a correlation coefficient of 0.987. At the extremes, the correlation is less obvious, presumably reflecting the small (<100 sentences) file sizes in these regions. In ES, for instance, the file corresponding to an average human score of 1.0 had just 20 sentences; at 1.2 there were 78 sentences, at 1.3, 48 sentences, and at 3.8, 26 sentences.

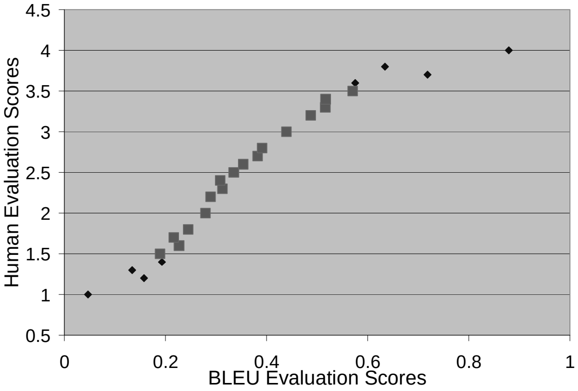

Figure 5: ES language pair; sentences grouped by average human score and plotted against the BLEU score for that grouping.

Figure 6 contains the output for the language pair with the smallest number of sentences (FS: 989 sentences).

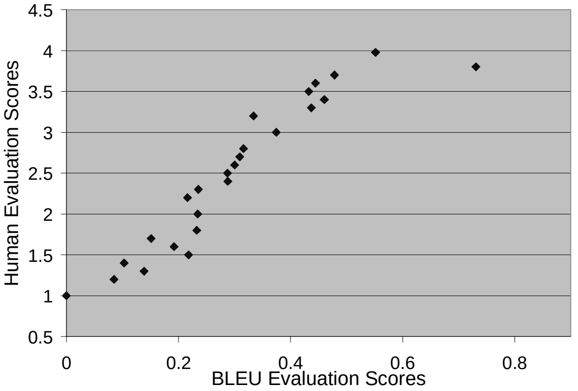

Figure 6: FS language pair; sentences grouped by average human score and plotted against the BLEU score for that grouping.

The charts for all seven language pairs examined in this study have a similar shape, though with file sizes as small as they are for FS, the line is less sharply defined. In FS, only two files contain more than 100 sentences, and the bottom five and top six files have fewer than 25 sentences each (ranging from 6 to 22).

We did not compute the NIST scores because file sizes are uneven and to even them out would result in very small files or a reduced range.

In all but the FS language pair, the file corresponding to sentences with a human score of 4.0 receives a very high BLEU score relative to the other human sentence averages. A review of the contents of those 4.0 files (including FS) reveals that a high number of sentences are identical to the reference (69% for ES; 81% for FS). In general, human raters appeared very reluctant to give a perfect 4.0 rating to translations that did not exactly match the reference. In the wide-domain QE data, where many different translations for a given source might be expected to be acceptable, a full 81% of the sentences that averaged a human score of 4.0 were identical to the reference.

In light of these results, we speculate that human raters follow a BLEU-like matching strategy, looking for shared material between the reference and translated sentences. Given the conditions under which raters performed the task, this claim seems plausible. Consider that our human evaluators were not experts in the highly technical computer domain and presumably unfamiliar with Canadian politics. Also consider that the task is tedious and speed is encouraged. Evaluators have only one reference sentence, with no larger context[4], to compare to the MT output. Under these conditions, it makes sense to rely on a simple strategy of comparing string-level n-grams, favoring longer n-grams when a test sentence deviates from the reference.

3 Conclusion

Automatic evaluation metrics have the potential to lower costs and speed development of MT systems. We have examined BLEU, and to a lesser extent NIST, comparing them to our extensive collection of human evaluations in order to determine if confidence in these measures is well placed.

To begin with, we showed that BLEU and NIST exhibit very similar behavior. As long as datasets are of equal size, we saw little to suggest we should strongly prefer one metric over the other.

When comparing two systems run on the same test sentences, we determined that BLEU and NIST reliably agree with human relative assessments, correlating rather strongly in many cases with the magnitude of the 'win'. Evaluations done in the Hansards domain did not correlate as strongly, suggesting that in wider domains BLEU and NIST may need more than one reference translation or a large test set in order to produce results that correlate reliably with human assessments.

When we grouped sentences by unigram score, achieving both larger datasets and datasets covering a wider range of human scores, we found that BLEU and human assessment scores correlate strongly, positively and linearly. This is a very encouraging result suggesting that especially when dataset are large (>500), correlations with BLEU will be strong.

One *a priori* reason to prefer human over automated MT evaluators is that humans are intelligent: they can recognize paraphrase relationships between two equally good translations, while "dumb" automated metrics are limited to simple string comparisons to one or more reference translations. One of the most interesting conclusions of our study, though, is that the superior linguistic skills of human raters are not exploited by MT evaluation tasks that involve quickly comparing a machine-translated sentence to a human-translated reference. Instead, human raters faced with the task of making quality judgments on technical and non-technical, out-of-context prose appear to rely on superficial, string-based criteria. In other words, they behave like expensive, slow versions of BLEU.

Human evaluators of MT quality can perform some tasks that are currently beyond the reach of automated metrics, such as identifying with high accuracy whether individual sentences are well translated or not. But in trying to assess translation quality at the corpus level, our results indicate that BLEU and NIST are highly reliable alternatives to human evaluation. Furthermore, in highly technical domains like ours, a single reference translation is sufficient to produce high-quality results. This is a particularly pleasing finding, since datasets that include multiple high-quality reference translations are comparatively rare.

4 Future Directions

We have very preliminary evidence that statistical (n-gram-based) MT systems receive higher BLEU scores than commercial/non-n-gram-based MT systems even when their human ratings are similar. The DARPA 2002 MT Workshop (Doddington,

[4] No attempt was made to measure coherence, in part because of the lack of any context beyond the sentence. Sentences were viewed in isolation. Our system and (we presume) all of the comparison systems included in this study lack discourse-level processing capabilities.

2002a) also found a difference in how NIST scores these different approaches. They attributed this difference to a difference in capitalization handling. Though our own approach to MT is a hybrid of rule-based and statistical strategies, it does not rely on n-gram matching. We likewise presume that the commercial comparison systems included in this study are also not fundamentally statistical. We welcome partners in quantifying what may be a systematic difference in how n-gram-based metrics score statistical vs. rule-based MT systems.

Other directions we would like to pursue include verifying these results with non-Indo-European languages and using our extensive human-evaluated multilingual corpora to examine other automatic MT metrics. Finally, we would like to explore richer assessment models: it is important to leverage human raters' unique abilities, rather than forcing them to behave as if they were simple string comparison algorithms.

5 Acknowledgements

I wish to thank my colleagues at Microsoft Research NLP: Mike Carlson for ideas, advice and inspiration, Bill Dolan for substantial edits and counsel, Chris Quirk for patient tutorials in statistics, and the entire team for contributing to the MT effort which spurred the need for evaluation. I would like to also acknowledge the fine work from the Butler Hill Group: Mo Corston-Oliver for extensive work normalizing the evaluation results and the many evaluators who provided the quality scores that are at the foundation of this paper.

6 Bibliographical References

Butler Hill Group. http://butlerhill.com/.

Corston-Oliver, S., M. Gamon, and C. Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of Annual Meeting of the Association for Computational Linguistic (ACL01)*.

Doddington, G. 2002a. Automatic Evaluation of Language Translations using N-gram Co-Occurrence Statistics. *DARPA 2002 Presentation*.

Doddington, G. 2002b. Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In *Proceeding of the Second International Conference on Human Language Technology*. San Diego, CA, pp. 138-145.

Hirschman, L., F. Reeder, J. Burger, and k. Miller. 2000. Name Translation as a Machine Translation Evaluation Task. In *Proceedings of the Workshop on Machine Translation, LREC-2000.*

Nieβen, S., F. J. Och, G. Leusch, H. Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of LREC*, Athens, Greece.

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. 2001. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA.

Reeder, F., 2001. In One Hundred Words or Less. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain.

Richardson, S., W. Dolan, A. Menezes, and M. Corston-Oliver. 2001. Overcoming the customization bottleneck using example-based MT. In *Proceedings, Workshop on Data-driven Machine Translation, 39th Annual Meeting and 10th Conference of the European Chapter, Association for Computational Linguistics*. Toulouse, France, pp. 9-16.