BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, and **Wei-Jing Zhu**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

Abstract

Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that can not be reused. We propose a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run. We present this method as an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations.[1]

1 Introduction

1.1 Rationale

Human evaluations of machine translation (MT) weigh many aspects of translation, including *adequacy*, *fidelity* , and *fluency* of the translation (Hovy, 1999; White and O'Connell, 1994). A comprehensive catalog of MT evaluation techniques and their rich literature is given by Reeder (2001). For the most part, these various human evaluation approaches are quite expensive (Hovy, 1999). Moreover, they can take *weeks* or *months* to finish. This is a big problem because developers of machine translation systems need to monitor the effect of *daily* changes to their systems in order to weed out bad ideas from good ideas. We believe that MT progress stems from evaluation and that there is a logjam of fruitful research ideas waiting to be released from the evaluation bottleneck. Developers would benefit from an inexpensive automatic evaluation that is quick, language-independent, and correlates highly with human evaluation. We propose such an evaluation method in this paper.

[1] So we call our method the bilingual evaluation understudy, BLEU.

1.2 Viewpoint

How does one measure translation performance? *The closer a machine translation is to a professional human translation, the better it is.* This is the central idea behind our proposal. To judge the quality of a machine translation, one measures its closeness to one or more reference human translations according to a numerical metric. Thus, our MT evaluation system requires two ingredients:

1. a numerical "translation closeness" metric
2. a corpus of good quality human reference translations

We fashion our closeness metric after the highly successful *word error rate* metric used by the speech recognition community, appropriately modified for multiple reference translations and allowing for legitimate differences in word choice and word order. The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family.

In Section 2, we describe the baseline metric in detail. In Section 3, we evaluate the performance of BLEU. In Section 4, we describe a human evaluation experiment. In Section 5, we compare our baseline metric performance with human evaluations.

2 The Baseline BLEU Metric

Typically, there are many "perfect" translations of a given source sentence. These translations may vary in word choice or in word order even when they use the same words. And yet humans can clearly distinguish a good translation from a bad one. For example, consider these two candidate translations of a Chinese source sentence:

Example 1.

Candidate 1: `It is a guide to action which ensures that the military always obeys the commands of the party.`

Candidate 2: `It is to insure the troops forever hearing the activity guidebook that party direct.`

Although they appear to be on the same subject, they differ markedly in quality. For comparison, we provide three reference human translations of the same sentence below.

Reference 1: `It is a guide to action that ensures that the military will forever heed Party commands.`

Reference 2: `It is the guiding principle which guarantees the military forces always being under the command of the Party.`

Reference 3: `It is the practical guide for the army always to heed the directions of the party.`

It is clear that the good translation, Candidate 1, shares many words and phrases with these three reference translations, while Candidate 2 does not. We will shortly quantify this notion of sharing in Section 2.1. But first observe that Candidate 1 shares `"It is a guide to action"` with Reference 1, `"which"` with Reference 2, `"ensures that the military"` with Reference 1, `"always"` with References 2 and 3, `"commands"` with Reference 1, and finally `"of the party"` with Reference 2 (all ignoring capitalization). In contrast, Candidate 2 exhibits far fewer matches, and their extent is less.

It is clear that a program can rank Candidate 1 higher than Candidate 2 simply by comparing *n*-gram matches between each candidate translation and the reference translations. Experiments over large collections of translations presented in Section 5 show that this ranking ability is a general phenomenon, and not an artifact of a few toy examples.

The primary programming task for a BLEU implementor is to compare *n*-grams of the candidate with the *n*-grams of the reference translation and count the number of matches. These matches are position-independent. The more the matches, the better the candidate translation is. For simplicity, we first focus on computing unigram matches.

2.1 Modified *n*-gram precision

The cornerstone of our metric is the familiar *precision* measure. To compute precision, one simply counts up the number of candidate translation words (unigrams) which occur in any reference translation and then divides by the total number of words in the candidate translation. Unfortunately, MT systems can overgenerate "reasonable" words, resulting in improbable, but high-precision, translations like that of example 2 below. Intuitively the problem is clear: a reference word should be considered exhausted after a matching candidate word is identified. We formalize this intuition as the *modified unigram precision.* To compute this, one first counts the maximum number of times a word occurs in any single reference translation. Next, one clips the total count of each candidate word by its maximum reference count,[2] adds these clipped counts up, and divides by the total (unclipped) number of candidate words.

Example 2.

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Modified Unigram Precision $= 2/7$.[3]

In Example 1, Candidate 1 achieves a modified unigram precision of $17/18$; whereas Candidate 2 achieves a modified unigram precision of $8/14$. Similarly, the modified unigram precision in Example 2 is $2/7$, even though its standard unigram precision is $7/7$.

[2] $Count_{clip} = \min(Count, Max\_Ref\_Count)$. In other words, one truncates each word's count, if necessary, to not exceed the largest count observed in any single reference for that word.

[3] As a guide to the eye, we have underlined the important words for computing modified precision.

Modified *n*-gram precision is computed similarly for any *n*: all candidate *n*-gram counts and their corresponding maximum reference counts are collected. The candidate counts are clipped by their corresponding reference maximum value, summed, and divided by the total number of candidate *n*-grams. In Example 1, Candidate 1 achieves a modified bigram precision of 10/17, whereas the lower quality Candidate 2 achieves a modified bigram precision of 1/13. In Example 2, the (implausible) candidate achieves a modified bigram precision of 0. This sort of modified *n*-gram precision scoring captures two aspects of translation: adequacy and fluency. A translation using the same words (1-grams) as in the references tends to satisfy *adequacy.* The longer *n*-gram matches account for *fluency.* [4]

2.1.1 Modified *n*-gram precision on blocks of text

How do we compute modified *n*-gram precision on a multi-sentence test set? Although one typically evaluates MT systems on a corpus of entire documents, our basic unit of evaluation is the sentence. A source sentence may translate to many target sentences, in which case we abuse terminology and refer to the corresponding target sentences as a "sentence." We first compute the *n*-gram matches sentence by sentence. Next, we add the clipped *n*-gram counts for all the candidate sentences and divide by the number of candidate *n*-grams in the test corpus to compute a modified precision score, p_n, for the entire test corpus.

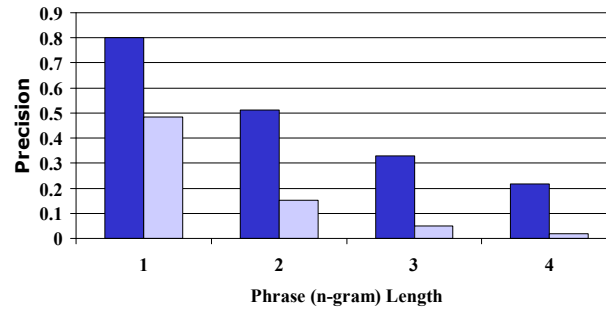
$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

[4] BLEU only needs to match human judgment when averaged over a test corpus; scores on individual sentences will often vary from human judgments. For example, a system which produces the fluent phrase "East Asian economy" is penalized heavily on the longer *n*-gram precisions if all the references happen to read "economy of East Asia." The key to BLEU's success is that all systems are treated similarly and multiple human translators with different styles are used, so this effect cancels out in comparisons between systems.

2.1.2 Ranking systems using only modified *n*-gram precision

To verify that modified *n*-gram precision distinguishes between very good translations and bad translations, we computed the modified precision numbers on the output of a (good) human translator and a standard (poor) machine translation system using 4 reference translations for each of 127 source sentences. The average precision results are shown in Figure 1.

Figure 1: Distinguishing Human from Machine

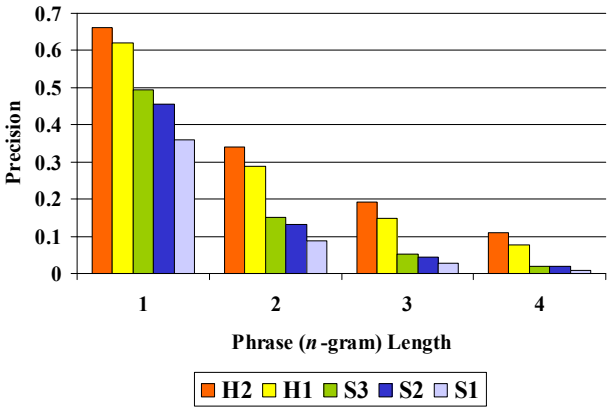


The strong signal differentiating human (high precision) from machine (low precision) is striking. The difference becomes stronger as we go from unigram precision to 4-gram precision. It appears that any single *n*-gram precision score can distinguish between a *good* translation and a *bad* translation. To be useful, however, the metric must also reliably distinguish between translations that do not differ so greatly in quality. Furthermore, it must distinguish between two human translations of differing quality. This latter requirement ensures the continued validity of the metric as MT approaches human translation quality.

To this end, we obtained a human translation by someone lacking native proficiency in both the source (Chinese) and the target language (English). For comparison, we acquired human translations of the same documents by a native English speaker. We also obtained machine translations by three commercial systems. These five "systems" — two humans and three machines — are scored against two reference professional human translations. The average modified *n*-gram precision results are shown in Figure 2.

Each of these *n*-gram statistics implies the same

Figure 2: Machine and Human Translations



ranking: H2 (Human-2) is better than H1 (Human-1), and there is a big drop in quality between H1 and S3 (Machine/System-3). S3 appears better than S2 which in turn appears better than S1. Remarkably, this is the *same rank order* assigned to these "systems" by human judges, as we discuss later. While there seems to be ample signal in any single *n*-gram precision, it is more robust to combine all these signals into a single number metric.

2.1.3 Combining the modified *n*-gram precisions

How should we combine the modified precisions for the various *n*-gram sizes? A weighted linear average of the modified precisions resulted in encouraging results for the 5 systems. However, as can be seen in Figure 2, the modified *n*-gram precision decays roughly exponentially with *n*: the modified unigram precision is much larger than the modified bigram precision which in turn is much bigger than the modified trigram precision. A reasonable averaging scheme must take this exponential decay into account; a weighted average of the logarithm of modified precisions satisifies this requirement.

BLEU uses the average logarithm with uniform weights, which is equivalent to using the geometric mean of the modified *n*-gram precisions.[5,6] Experimentally, we obtain the best correlation with monolingual human judgments using a maximum *n*-gram order of 4, although 3-grams and 5-grams give comparable results.

2.2 Sentence length

A candidate translation should be neither too long nor too short, and an evaluation metric should enforce this. To some extent, the *n*-gram precision already accomplishes this. *N*-gram precision penalizes spurious words in the candidate that do not appear in any of the reference translations. Additionally, modified precision is penalized if a word occurs more frequently in a candidate translation than its maximum reference count. This rewards using a word as many times as warranted and penalizes using a word more times than it occurs in any of the references. However, modified *n*-gram precision alone fails to enforce the proper translation length, as is illustrated in the short, absurd example below.

Example 3:

Candidate: `of the`

Reference 1: `It is a guide to action that ensures that the military will forever heed Party commands.`

Reference 2: `It is the guiding principle which guarantees the military forces always being under the command of the Party.`

Reference 3: `It is the practical guide for the army always to heed the directions of the party.`

Because this candidate is so short compared to the proper length, one expects to find inflated precisions: the modified unigram precision is 2/2, and the modified bigram precision is 1/1.

2.2.1 The trouble with recall

Traditionally, precision has been paired with recall to overcome such length-related problems. However, BLEU considers *multiple* reference translations, each of which may use a different word choice to translate the same source word. Furthermore, a good candidate translation will only use (recall) one of these possible choices, but not all. Indeed, recalling all choices leads to a bad translation. Here is an example.

[5]The geometric average is harsh if any of the modified precisions vanish, but this should be an extremely rare event in test corpora of reasonable size (for $N_{max} \leq 4$).

[6]Using the geometric average also yields slightly stronger correlation with human judgments than our best results using an arithmetic average.

Example 4:

Candidate 1: `I always invariably perpetually do.`

Candidate 2: `I always do.`

Reference 1: `I always do.`

Reference 2: `I invariably do.`

Reference 3: `I perpetually do.`

The first candidate recalls more words from the references, but is obviously a poorer translation than the second candidate. Thus, naïve recall computed over the set of all reference words is not a good measure. Admittedly, one could align the reference translations to discover synonymous words and compute recall on concepts rather than words. But, given that reference translations vary in length and differ in word order and syntax, such a computation is complicated.

2.2.2 Sentence brevity penalty

Candidate translations longer than their references are already penalized by the modified *n*-gram precision measure: there is no need to penalize them again. Consequently, we introduce a multiplicative *brevity penalty* factor. With this brevity penalty in place, a high-scoring candidate translation must now match the reference translations in length, in word choice, and in word order. Note that neither this brevity penalty nor the modified *n*-gram precision length effect directly considers the source length; instead, they consider the range of reference translation lengths in the target language.

We wish to make the brevity penalty 1.0 when the candidate's length is the same as any reference translation's length. For example, if there are three references with lengths 12, 15, and 17 words and the candidate translation is a terse 12 words, we want the brevity penalty to be 1. We call the closest reference sentence length the "*best match length*."

One consideration remains: if we computed the brevity penalty sentence by sentence and averaged the penalties, then length deviations on short sentences would be punished harshly. Instead, we compute the brevity penalty over the entire corpus to allow some freedom at the sentence level. We first compute the test corpus' effective reference length, r, by summing the best match lengths for each candidate sentence in the corpus. We choose the brevity penalty to be a decaying exponential in r/c, where c is the total length of the candidate translation corpus.

2.3 BLEU details

We take the geometric mean of the test corpus' modified precision scores and then multiply the result by an exponential brevity penalty factor. Currently, case folding is the only text normalization performed before computing the precision.

We first compute the geometric average of the modified n-gram precisions, p_n, using n-grams up to length N and positive weights w_n summing to one.

Next, let c be the length of the candidate translation and r be the effective reference corpus length. We compute the brevity penalty BP,

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}.$$

Then,

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

The ranking behavior is more immediately apparent in the log domain,

$$\log \mathrm{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

In our baseline, we use $N = 4$ and uniform weights $w_n = 1/N$.

3 The BLEU Evaluation

The BLEU metric ranges from 0 to 1. Few translations will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1. It is important to note that the more reference translations per sentence there are, the higher the score is. Thus, one must be cautious making even "rough" comparisons on evaluations with different numbers of reference translations: on a test corpus of about 500 sentences (40 general news stories), a human translator scored 0.3468 against four references and scored 0.2571 against two references. Table 1 shows the BLEU scores of the 5 systems against two references on this test corpus.

The MT systems S2 and S3 are very close in this metric. Hence, several questions arise:

Table 1: BLEU on 500 sentences

S1	S2	S3	H1	H2
0.0527	0.0829	0.0930	0.1934	0.2571

Table 2: Paired t-statistics on 20 blocks

	S1	S2	S3	H1	H2
Mean	0.051	0.081	0.090	0.192	0.256
StdDev	0.017	0.025	0.020	0.030	0.039
t	—	6	3.4	24	11

- Is the difference in BLEU metric reliable?
- What is the variance of the BLEU score?
- If we were to pick another random set of 500 sentences, would we still judge S3 to be better than S2?

To answer these questions, we divided the test corpus into 20 blocks of 25 sentences each, and computed the BLEU metric on these blocks individually. We thus have 20 samples of the BLEU metric for each system. We computed the means, variances, and paired t-statistics which are displayed in Table 2. The t-statistic compares each system with its left neighbor in the table. For example, $t = 6$ for the pair S1 and S2.

Note that the numbers in Table 1 are the BLEU metric on an aggregate of 500 sentences, but the means in Table 2 are averages of the BLEU metric on aggregates of 25 sentences. As expected, these two sets of results are close for each system and differ only by small finite block size effects. Since a paired t-statistic of 1.7 or above is 95% significant, the differences between the systems' scores are statistically very significant. The reported variance on 25-sentence blocks serves as an upper bound to the variance of sizeable test sets like the 500 sentence corpus.

How many reference translations do we need? We simulated a single-reference test corpus by randomly selecting one of the 4 reference translations as the single reference for each of the 40 stories. In this way, we ensured a degree of stylistic variation. The systems maintain the same rank order as with multiple references. This outcome suggests that we may use a big test corpus with a single reference translation, provided that the translations are not all from the same translator.

4 The Human Evaluation

We had two groups of human judges. The first group, called the monolingual group, consisted of 10 native speakers of English. The second group, called the bilingual group, consisted of 10 native speakers of Chinese who had lived in the United States for the past several years. None of the human judges was a professional translator. The humans judged our 5 standard systems on a Chinese sentence subset extracted at random from our 500 sentence test corpus. We paired each source sentence with each of its 5 translations, for a total of 250 pairs of Chinese source and English translations. We prepared a web page with these translation pairs randomly ordered to disperse the five translations of each source sentence. All judges used this same webpage and saw the sentence pairs in the same order. They rated each translation from 1 (very bad) to 5 (very good). The monolingual group made their judgments based only on the translations' readability and fluency.

As must be expected, some judges were more liberal than others. And some sentences were easier to translate than others. To account for the intrinsic difference between judges and the sentences, we compared each judge's rating for a sentence across systems. We performed four pairwise t-test comparisons between adjacent systems as ordered by their aggregate average score.

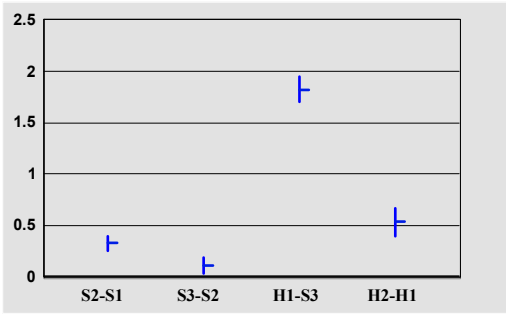
4.1 Monolingual group pairwise judgments

Figure 3 shows the mean difference between the scores of two consecutive systems and the 95% confidence interval about the mean. We see that S2 is quite a bit better than S1 (by a mean opinion score difference of 0.326 on the 5-point scale), while S3 is judged a little better (by 0.114). Both differences are significant at the 95% level.[7] The human H1 is much better than the best system, though a bit worse than human H2. This is not surprising given that H1 is not a native speaker of either Chinese or English,

[7] The 95% confidence interval comes from t-test, assuming that the data comes from a T-distribution with N degrees of freedom. N varied from 350 to 470 as some judges have skipped some sentences in their evaluation. Thus, the distribution is close to Gaussian.

whereas H2 is a native English speaker. Again, the difference between the human translators is significant beyond the 95% level.

Figure 3: Monolingual Judgments - pairwise differential comparison

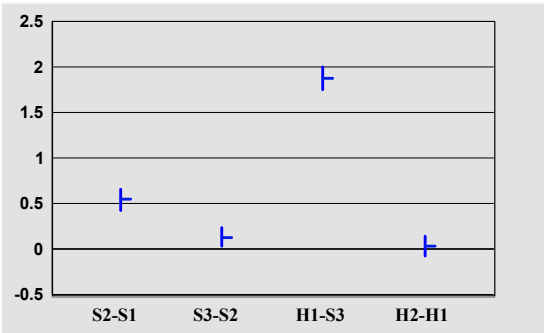


+95%	0.400	0.194	1.945	0.670
-95%	0.252	0.034	1.705	0.400
monolingual	0.326	0.114	1.825	0.535

4.2 Bilingual group pairwise judgments

Figure 4 shows the same results for the bilingual group. They also find that S3 is slightly better than S2 (at 95% confidence) though they judge that the human translations are much closer (indistinguishable at 95% confidence), suggesting that the bilinguals tended to focus more on adequacy than on fluency.

Figure 4: Bilingual Judgments - pairwise differential comparison



+95%	0.667	0.238	2.007	0.145
-95%	0.435	0.042	1.759	-0.069
bilingual	0.551	0.140	1.883	0.038

5 BLEU vs The Human Evaluation

Figure 5 shows a linear regression of the monolingual group scores as a function of the BLEU score over two reference translations for the 5 systems. The high correlation coefficient of 0.99 indicates that BLEU tracks human judgment well. Particularly interesting is how well BLEU distinguishes between S2 and S3 which are quite close. Figure 6 shows the comparable regression results for the bilingual group. The correlation coefficient is 0.96.

Figure 5: BLEU predicts Monolingual Judgments

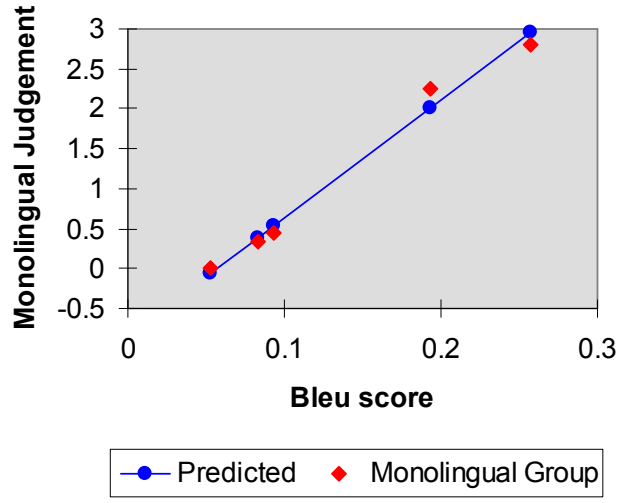
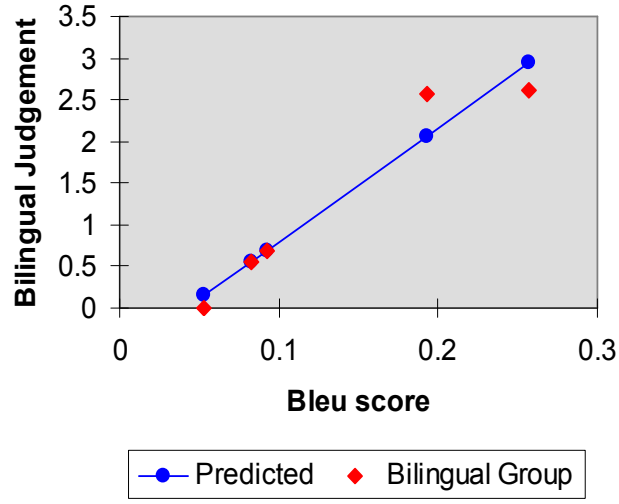


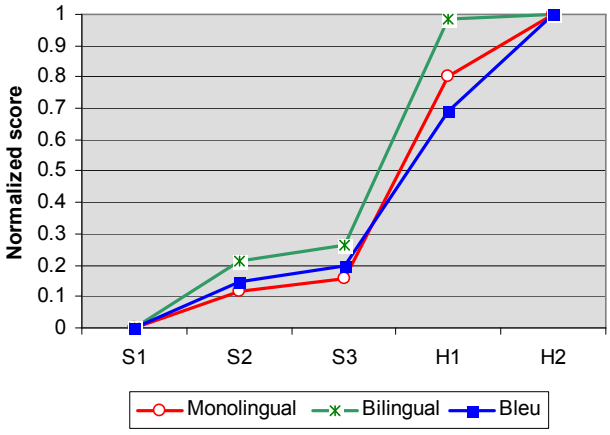
Figure 6: BLEU predicts Bilingual Judgments



We now take the worst system as a reference point and compare the BLEU scores with the human judg-

ment scores of the remaining systems relative to the worst system. We took the BLEU, monolingual group, and bilingual group scores for the 5 systems and linearly normalized them by their corresponding range (the maximum and minimum score across the 5 systems). The normalized scores are shown in Figure 7. This figure illustrates the high correlation between the BLEU score and the monolingual group. Of particular interest is the accuracy of BLEU's estimate of the small difference between S2 and S3 and the larger difference between S3 and H1. The figure also highlights the relatively large gap between MT systems and human translators.[8] In addition, we surmise that the bilingual group was very forgiving in judging H1 relative to H2 because the monolingual group found a rather large difference in the fluency of their translations.

Figure 7: BLEU vs Bilingual and Monolingual Judgments



6 Conclusion

We believe that BLEU will accelerate the MT R&D cycle by allowing researchers to rapidly home in on effective modeling ideas. Our belief is reinforced by a recent statistical analysis of BLEU's correlation with human judgment for translation into English from four quite different languages (Arabic, Chinese, French, Spanish) representing 3 different language families (Papineni et al., 2002)! BLEU's strength is that it correlates highly with human judgments by averaging out individual sentence judgment errors over a test corpus rather than attempting to divine the exact human judgment for every sentence: *quantity leads to quality.*

Finally, since MT and summarization can both be viewed as natural language generation from a textual context, we believe BLEU could be adapted to evaluating summarization or similar NLG tasks.

Acknowledgments This work was partially supported by the Defense Advanced Research Projects Agency and monitored by SPAWAR under contract No. N66001-99-2-8916. The views and findings contained in this material are those of the authors and do not necessarily reflect the position of policy of the Government and no official endorsement should be inferred.

We gratefully acknowledge comments about the geometric mean by John Makhoul of BBN and discussions with George Doddington of NIST. We especially wish to thank our colleagues who served in the monolingual and bilingual judge pools for their perseverance in judging the output of Chinese-English MT systems.

References

E.H. Hovy. 1999. Toward finely differentiated evaluation metrics for machine translation. In *Proceedings of the Eagles Workshop on Standards and Evaluation*, Pisa, Italy.

Kishore Papineni, Salim Roukos, Todd Ward, John Henderson, and Florence Reeder. 2002. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In *Proceedings of Human Language Technology 2002*, San Diego, CA. To appear.

Florence Reeder. 2001. Additional mt-eval references. Technical report, International Standards for Language Engineering, Evaluation Working Group. http://issco-www.unige.ch/projects/isle/taxonomy2/

J.S. White and T. O'Connell. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 193–205, Columbia, Maryland.

[8] Crossing this chasm for Chinese-English translation appears to be a significant challenge for the current state-of-the-art systems.