Interpreting BLEU/NIST Scores: How Much Improvement Do We Need to Have a Better System?

Ying Zhang **Stephan Vogel** **Alex Waibel**
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213-3891 U.S.A.
{joy+, vogel+, ahw+}@cs.cmu.edu

Abstract

Automatic evaluation metrics for Machine Translation (MT) systems, such as BLEU and the related NIST metric, are becoming increasingly important in MT. Yet, their behaviors are not fully understood. In this paper, we analyze some flaws in the BLEU/NIST metrics. With a better understanding of these problems, we can better interpret the reported BLEU/NIST scores. In addition, this paper reports a novel method of calculating the confidence intervals for BLEU/NIST scores using bootstrapping. With this method, we can determine whether two MT systems are significantly different from each other.

Introduction

Automatic evaluation for Machine Translation (MT) systems has become prominent with the development of data driven MT. The essential idea comes from 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 central idea of automatic evaluation 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. Based on these principles, the IBM MT research group proposed the BLEU metric. BLEU averages the precision for unigram, bigram and up to 4-grams and applies a length penalty if the generated sentence is shorter than the best matching (in length) reference translation (Papineni et al, 2001). A variant of BLEU has been adopted by NIST for its MT effort. The NIST metric is derived from the BLEU evaluation criterion but differs in one fundamental aspect: instead of n-gram precision the information gain from each n-gram is taken into account. The idea behind this is to give more credit if a system gets an n-gram match that is difficult, but to give less credit for an n-gram match which is easy (NIST 2002).

Limitations of BLEU/NIST

Both BLEU and NIST metrics are based on the idea of modified n-gram precision. Typical questions one encounters during an attempt to analyze the BLEU/NIST scores could be "For the overall score, how much does the unigram match contribute? How much does the bigram match contribute? Etc." In other words, how much credit does an MT system get for generating the correct words and how much additional credit for putting them in the correct order. We ran a number of analyses and investigated this question. Table 1 shows a typical example of an MT system evaluated by the NIST metric (considering up to 5-grams). In this table, each row lists the matching information for the unigram, bigram and up to 5-grams. For a certain *n*, "# In Trans" is the number of the *n*-grams in the translation. "# of Match" is the total number of the *n*-grams in the translation that can also be found in the human references. "Info Gain" is the sum of the information gain for the *n*-grams and the "Avg Info Gain" is the averaged information gain for each matched *n*-gram (Info Gain / # of Match). The contribution from the *n*-grams is the "Prec. Score". "Prec. Score" is the weighted precision for the *n*-grams (Info Gain / # In Trans). For this MT system, the precision score is 7.188 and the length penalty is 1.0 (no penalty). Its 1-gram precision is 5.704, bigram precision 1.222, 3-gram 0.215, 4-gram 0.037 and 5-gram 0.009 respectively. As we can see, about 80% of the overall precision score comes from the uni-grams. Another 17% comes from the bi-grams; 5-grams contribute only 0.1%. Of course, the number of longer n-gram matches is smaller compared with shorter n-gram matches (18243 matches for uni-gram and only 544 matches for 5-grams). But this is only part of the effect. A closer look reveals that most of the 544 matching 5-grams do not contribute to the final score because their information gain was zero. In addition, even if there is a positive information gain, it is on average much smaller for longer n-grams than for shorter n-grams.

n	# In Trans	# of Match	Info Gain	Avg Info Gain	Prec. Score	% Prec. Contrib
1	28113	18243	160365.1	8.79	5.704	79.4
2	27235	7690	33280.3	4.33	1.222	17.0
3	26357	3145	5661.7	1.80	0.215	3.0
4	25479	1336	950.3	0.71	0.037	0.5
5	24601	544	228.4	0.53	0.009	0.1
Σ					7.188	100

Table 1. n-gram contributions to the NIST score.

The BLEU score is the geometric mean of the n-gram precisions. Therefore it is harder to dissect the score and ask for the contributions from the n-grams of different length. The following constructed example demonstrates the effect of n-gram matches for BLEU. Assume that the n-gram precisions from system *A* are: 1.00, 0.21, 0.11, 0.06, whereas system *B* has score 0.35, 0.32, 0.28 and 0.26. System *A* has the right words, but not always in the right order, where system *B* gets about a third of the sentence right, but fails on the remainder of the sentence. The overall BLEU score for system *A* is 0.19 and 0.29 for *B*, indicating that the second system generated a much "better" translation. It is difficult to believe that a translation which has only about a third of its words matching with the reference translation should be better than a translation which has most of the words correct.

Similar behavior in the BLEU score has been observed in real experiments.

Our analysis shows that the BLEU and the NIST metrics display opposite behavior with respect to the questions: how much credit is given for correct lexical choice and how much additional credit is given for correct word order? NIST hardly gives any credit for correct word order, whereas BLEU gives too much credit for getting some 3- and 4-grams right, overriding the contribution from unigrams.

N-gram metrics are essentially document similarity measures rather than true translation quality measures (Popescu-Belis, 2003). Much care must be taken in using n-gram measures in formal evaluations of machine translation quality, though they are still valuable as part of the interactive development cycle.

Besides these inherent imperfections in both evaluation metrics, we are always faced questions like: how reliable are the scores? What are the confidence intervals? How significant is the difference if one system has a higher BLEU/NIST score than another? In the next section, we will describe a bootstrapping approach to measure the statistical significance for BLEU/NIST scores.

Confidence Intervals for BLEU/NIST Scores

Both BLEU/NIST metrics require a test suite to evaluate the MT systems. A test suite consists of two parts: testing sentences in the source language and multiple human reference translations in the target language. To have enough coverage in the source language, a test suite usually has hundreds of sentences. For example, in the NIST June 2002 MT evaluation suite, there are 878 sentences for Chinese-English systems and 728 sentences for Arabic-English. In order to cover the translation variations, a test suite needs multiple human references, typically 4 or more. With these two factors, building a test suite is not cheap. In fact, since the introduction of BLEU, the MT community has had only a few test suites with multiple human references. The BLEU/NIST scores are usually based on one test suite. Thus, when we have a BLEU/NIST for one MT system, we have to ask ourselves a question: "Is this score precise?"

How precise is the BLEU/NIST score?

In statistical tests, we often use confidence interval to measure the precision of an estimated value. The interval represents the range of values, consistent with the data, which is believed to encompass the "true" value with high probability (usually 95%). The confidence interval is expressed in the same units as the estimate. Wider intervals indicate lower precision; narrow intervals, greater precision. The estimated range is calculated from a given set of sample data.

Since building test suites is expensive, it is not practical to create a set of testing suites to generate a set of sample BLEU/NIST scores. Instead, we use the well-known *bootstrapping* technique to measure the confidence interval for BLEU/NIST.[1]

Bootstrapping is a statistical method of getting the confidence interval (Efron and Tibshirani, 1986, 1993). The term *bootstrapping* refers to the old story about people lifting themselves off the ground by pulling on the backs of their own boots. A similar seemingly impossible thing occurs when we resample the data to get confidence intervals. The essential idea here is to test the system using a variety of testing suites created by resampling. Here is how it works.

Suppose we have a test suite T_0 to test several Machine Translation systems translating from Chinese to English. There are N Chinese testing segments in the suite and for each testing segment we have R human translations. A segment is typically a sentence, but it can also be a paragraph or a document. Let's represent the *i-th* segment of T_0 as an n-tuple $<s_i, r_{i1}, r_{i2},..,r_{iR}>$, where s_i is the *i-th* Chinese segment to be translated and r_{i1} to r_{iR} are the R human translations (references) for segment s_i. Create a new test suite T_1 with N segments by sampling with replacement from T_0. Since we sample with replacement, a segment in T_0 may occur zero, once or more than once in T_1. Repeat these process M times, e.g. M=1999, and we have $M+1$ test suites: T_0, T_1... T_M, where T_1 to T_M are artificial test suites created by resampling T_0.

Evaluate the MT systems on each of these $M+1$ test suite using either the BLEU or the NIST metric. For each MT system, we will then have $M+1$ BLEU/NIST scores. As one may expect, these scores have a normal distribution. Figure 1 shows an example of the BLEU score distribution over 2000 resampled test suites for an MT system. From these $M+1$ scores, find the middle 95% of the scores (i.e. the 2.5th percentile and the 97.5th percentile). That is the 95% confidence interval for the BLEU score of this MT system.

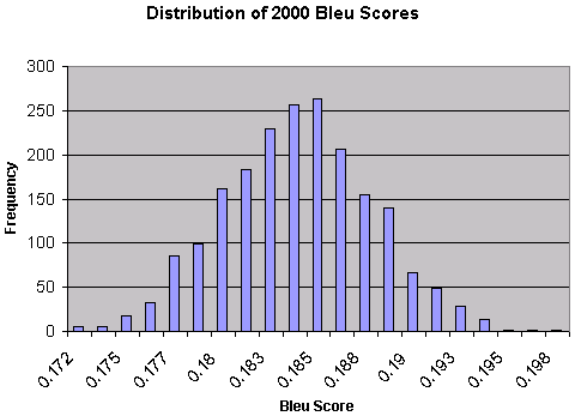

Figure 1. Distribution of BLEU scores over 2000 test suites

For the MT system shown in Figure 1, we find that the mean BLEU score is 0.184 and the 95% confidence interval is [0.176, 0.192].

[1] The idea of using bootstrapping for confidence intervals was originally suggested by Franz Och.

From the *M+1* scores, we can calculate the mean μ and the standard deviation σ. Define the *Relative Standard Deviation* (*RSD*) as: $RSD = (100*\sigma/\mu)\%$. *RSD* correlates with the width of the confidence interval scaled according to μ. It indicates the precision of the estimated score.

Listed in Table 2 and 3 are 7 Chinese-English MT systems evaluated on 878 testing sentences using 2,000 test suites. It is interesting to notice that systems with higher BLEU/NIST scores also have higher precision (lower RSD values).

System	Mean	Interval	RSD
A	0.184	[0.176, 0.192]	2.17%
B	0.165	[0.157, 0.174]	2.80%
C	0.180	[0.172, 0.189]	2.46%
D	0.144	[0.137, 0.152]	2.64%
E	0.072	[0.068, 0.078]	3.47%
F	0.241	[0.232, 0.250]	1.95%
G	0.182	[0.173, 0.189]	2.30%

Table 2. Examples of BLEU scores with confidence intervals and Relative Standard Deviation (RSD)

System	Mean	Interval	RSD
A	7.188	[7.059, 7.313]	0.84%
B	6.191	[6.023, 6.358]	1.46%
C	6.935	[6.812, 7.064]	0.94%
D	6.524	[6.410, 6.638]	0.90%
E	4.939	[4.832, 5.045]	1.07%
F	7.468	[7.335, 7.595]	0.91%
G	7.153	[7.037, 7.268]	0.86%

Table 3. Examples of NIST scores with confidence intervals and Relative Standard Deviation (RSD)

Is one MT system really better than another?

Since the year 2002, BLEU/NIST metrics have been used as the official evaluation metrics in the TIDES MT project evaluation. The MT community has been using BLEU/NIST to compare the performances of different MT systems, and to report the improvement of MT systems. The question yet to be answered is how significant are the differences between the systems? For example, MT system *A* scored (BLEU) 0.184 on a certain test suite and another MT system *G* scored 0.182 on the same test suite. 0.184 is certainly bigger than 0.182, but is this difference statistically significant? In another example, MT system *D* scored 6.524 (NIST) on a certain test suite. We developed some new alignment algorithm and added it to system *D*, the new system *D'* scored 6.935 on the same test suite. The score 6.935 is higher than 6.524 by 6%, but can we claim that this new alignment algorithm has significantly improved the translation quality?

In a way similar to measuring the confidence intervals for an MT system's BLEU/NIST score, we can use bootstrapping to measure the confidence intervals for the discrepancy between the two MT systems. Here is how it works.

Create test suites $T_0, T_1 \ldots T_M$, where T_1 to T_M are artificial test suites created by resampling T_0. System *A* scored a_0 on T_0 and system *B* scored b_0. The discrepancy between system *A* and *B* is $\delta_0 = a_0 - b_0$. Repeat this process on every *M+1* test suite and we have *M+1* discrepancy scores: $\delta_0, \delta_1 \ldots \delta_M$. From these *M+1* discrepancy scores, find the middle 95% of the scores (i.e. the 2.5th percentile and the 97.5th percentile). That is the 95% confidence interval for the discrepancy between MT system *A* and *B*. If the confidence interval does not overlap with zero, we can claim that the difference between system *A* and *B* are *statistically significant*.

Table 4 and 5 are examples of significant differences among seven MT systems as discussed in Table 2 and 3. In these two tables, ">" means system *X* is significantly "better" than system *Y*, where as "<" means that system X is significantly "worse" than Y. If the discrepancy between *X* and *Y* is not significant, i.e. the confidence interval overlaps with zero, we use "~" to represent that the two systems are not significantly different.

Sys Y: / Sys X:	A	B	C	D	E	F	G
A (0.184)		>	~	>	>	<	~
B (0.165)	<		<	>	>	<	<
C (0.180)	~	>		>	>	<	~
D (0.144)	<	<	<		>	<	<
E (0.072)	<	<	<	<		<	<
F (0.241)	>	>	>	>	>		>
G (0.182)	~	>	~	>	>	<	

Table 4. Comparison among 7 Chinese-English MT systems by BLEU

Sys Y: / Sys X:	A	B	C	D	E	F	G
A (7.188)		>	>	>	>	<	~
B (6.191)	<		<	<	>	<	<
C (6.935)	~	>		>	>	<	<
D (6.524)	<	>	<		>	<	<
E (4.939)	<	<	<	<		<	<
F (7.468)	>	>	>	>	>		>
G (7.153)	~	>	>	>	>	<	

Table 5. Comparison among 7 Chinese-English MT systems by NIST

From these results we can see that the NIST metric has more discriminative power than the BLEU metric. For example, BLEU cannot distinguish between system *A* (BLEU=0.184, NIST=7.188) and system *C* (BLEU=0.180, NIST=6.935), yet they are significantly different when evaluated using the NIST metric.

To compare with the automatic evaluation metrics, we also measured the significance of the discrepancies among the human judgments (Table 6). The human assessments were carried out by LDC on July 2002. Similar to the DARPA-94 MT evaluation (White 94), the human assessment was a holistic scoring by committees of human evaluators on the basis of the somewhat vaguely specified parameters of *fluency* and *adequacy*. Human evaluators were asked to assign the *fluency* and *adequacy* scores for each sentence generated by MT systems. The scores range from 1 to 5, where 1 stands for "worst" and 5 for "best". Each sentence was evaluated by at least two evaluators and we use the averaged value as

the human judgment for that sentence. Averaged among all the translation sentences, the sum of the *fluency* and *adequacy* is the human judgment for that MT system.

Sys Y: / Sys X:	A	B	C	D	E	F	G
A (4.90)		>	>	>	>	~	<
B (4.27)	<		<	<	<	<	<
C (4.77)	<	>		>	>	<	<
D (4.55)	<	>	<		~	<	<
E (4.52)	<	>	<	~		<	<
F (4.97)	~	>	>	>	>		<
G (5.62)	>	>	>	>	>	>	

Table 6. Comparison among 7 Chinese-English MT systems by human judgment

It is obvious that neither the BLEU nor the NIST metric has the same effect as the human judgement. In fact, for these 7 MT systems, *correlation(human scores, BLEU) = 0.38*, and *correlation(human scores, NIST) = 0.73*. Also, notice that system *F*(BLEU=0.241, NIST=7.468) is *significantly better* than system *G*(BLEU=0.182, NIST=7.153) according to the BLEU/NIST, yet human ranked them in the opposite way: system *F*(human score = 4.97) is *significantly worse* than system *G*(human score = 5.62). One possible reason for this incoherence is that system *G* is a rule-based MT system. Unlike the statistical MT system *F*, *G* generates more fluent translations. Human evaluators give higher scores to fluent and natural translations yet automatic evaluation metrics do not gain much from longer matched n-grams as we have discussed in the second section. Another incoherence is system *E*'s score. *E*'s BLEU/NIST scores are *significantly worse* than *B* and *D*'s, but its human score is at least as good as *B* and *D*'s human score. When we investigated *E*'s translations, we found that *E*'s translation has some errors in its format, e.g. in the SGML tags. Both BLEU and NIST require the translations to be tagged correctly but human evaluators ignore errors in the taggings. We believe this could be the reason for the lower BLEU/NIST score for system *E*.

Implementation Issues

To calculate the confidence intervals using bootstrapping, we need to translate and evaluate the MT systems on each of the *M+1* test suites. *M* needs to be large, say, 1,000 or even 10,000, to guarantee reliable results. Translating 1000 test suites may take a very long time for some MT systems. But for most MT systems, the translation for a segment is independent of the previous segments in the test suite. In other words, the translation of segment *s* should always be the same no matter which test suite it is part of. In that sense, we do not need to translate *M+1* test suites. Instead, we only need to resample the translations of T_0 and their corresponding human references. We developed an efficient method for bootstrapping. After translating T_0, all the n-gram matching information for segments in T_0 are collected and stored in an array. To simulate the translation results of the artificial test suites, we need only resample the information from this array and calculate the BLEU/NIST scores from the segment's scores[2].

Conclusion

In conclusion, we analyzed two currently used approaches to automatic MT evaluation, the BLEU and the NIST metrics. Our analysis show that both metrics have some flaws; one needs to understand these problems to correctly interpret the reported BLEU/NIST scores. We also described a method to calculate the statistical significance for BLEU/NIST using bootstrapping. These tests will give us the confidence interval for the BLEU/NIST scores. They can also measure if two MT systems are significantly different from each other.

References

Culy, Christopher & Riehemann, Susanne Z. (2003). The Limits of N-Gram Translation Evaluation Metrics. In Proceedings of the Ninth Machine Translation Summit. New Orleans, Louisiana, USA.

Efron, B. and R. Tibshirani (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, Statistical Science 1, p. 54-77.

Efron, Bradley and Rob, Tibshirani (1993). An Introduction to Bootstrap. Chapman & Hall, New York.

NIST Report (2002) Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

Papineni, Kishore & Roukos, Salim et al. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics.

Pepescu-Belis, Andrei (2003). An Experiment in Comparative Evaluation: Humans vs. Computers. In Proceedings of the Ninth Machine Translation Summit. New Orleans, Louisiana, USA.

Van Slype, Georges. (1979). Critical Study of Methods for Evaluating the Quality of Machine Translation, European Commission / Directorate for General Scientific and Technical Information Management (DG XIII), BR 19142. http://issco-www.unige.ch/projects/isle/van-slype.pdf

White, J. S., T. A. O'Connell, & F. E. O'Mara. (1994). Advanced Research Projects Agency Machine Translation Program: 3Q94. Proceedings of the November 1994 Meeting.

[2] The script is available for downloading at: http://projectile.is.cs.cmu.edu/research/public/tools/bootStrap/tutorial.htm