Measuring Confidence Intervals for the Machine Translation Evaluation Metrics

Ying Zhang   Stephan Vogel
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15217, U.S.A.
{joy+,vogel+}@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. 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. We study the effect of test set size and number of reference translations on the confidence intervals for these MT evaluation metrics.

1. 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. For MT evaluation this has been extended to multiple reference translations (Nießen et al. 2000), and allowing for differences in word order (Leusch et al. 2003). In (Papineni et al, 2002) the BLEU metric was proposed, which averages the precision for unigram, bigram and up to 4-grams and applies a length penalty for translations too short. A variant of BLEU has been developed by NIST, using the information gain of the n-grams. Additional modifications to BLEU-type metrics have been proposed to improve the correlation with human evaluation scores (Melamed 2003, Pepescu-Belis 2003).

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. In order to cover translation variations multiple human references are used, typically 4 or more. This makes building a test suite expensive. Therefore, the BLEU/NIST scores are usually based on one test suite. Thus, we have to ask ourselves a question: "Is this score reliable?" Or in other words, what is the confidence interval for a specific metric, a particular translation system, and a given test set. Fortunately, statistical testing theory has developed an appropriate tool to deal with this kind of situation, the so-called bootstrapping method.

After a short introduction into the MT evaluation metrics we will describe this bootstrapping approach. We will then study in detail the effect of test set size and the number of reference translations on the width of the confidence interval. In the case study presented in this paper we will use results from the TIDES MT evaluation 2002, esp. from the so-called large data track Chinese-English translation systems.

2. MT Evaluation Metrics

2.1. IBM BLEU metric

The BLEU metric is based on the *modified n-gram precision*, which counts how many n-grams of the candidate translation match with n-grams of the reference translation:

$$p_n = \frac{\sum\limits_{Sent\in\{Hyp\}} \sum\limits_{n-gram\in Sent} Count_{matched}(n-gram)}{\sum\limits_{Sent\in\{Hyp\}} \sum\limits_{n-gram\in Sent} Count(n-gram)} \qquad (Eq.\ 1)$$

To compute p_n, one first counts the maximum number of times an n-gram occurs in any single reference translation. Next, one clips the total count of each candidate n-gram by its maximum reference count, adds these clipped counts up, and divides by the total (unclipped) number of candidate words.

To bring in the factor of "recall", BLEU uses a "brevity penalty" to penalize candidates shorter than their reference translations. For a candidate translation with length c, its brevity penalty is defined as:

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{(1-r/c)} & if \quad c \leq r \end{cases} \qquad (Eq.\ 2)$$

where r is the "best match length" among all reference translations.

The final BLEU score is then the geometric average of the modified n-gram precision multiplied by the brevity penalty: $BLEU = BP \bullet \exp(\sum\limits_{n=1}^{N} w_n \log p_n)$ (Eq. 3)

Usually, $N=4$ and $w_n=1/N$.

2.2. Modified BLEU metric

The BLEU metric focuses heavily on long n-grams. A low score on 4-grams will result in an overall low score, even if unigram precision is high. This is due to the fact that the geometric mean of the n-gram precision scores is used. We proposed a modified version to the original BLEU metric called the "modified BLEU" (M-BLEU for short). In M-BLEU, a more balanced contribution from the different n-grams is achieved using the arithmetic means of the n-gram precisions. $M-BLEU = BP \bullet \sum\limits_{n=1}^{N} w_n p_n$ (Eq. 4). The calculation of the modified n-gram precision and the brevity penalty is the same as in BLEU.

2.3. NIST Mteval metric

The motivation of NIST MTeval scoring metric (NIST score in short) is to weight more heavily those N-grams that are more informative. This would, in addition, help to combat possible gaming of the scoring algorithm, since those N-grams that are most likely to (co-)occur would add less to the score than less likely N-grams. With the information gain

$$Info(w_1....w_n) = \log_2\left(\frac{the \ \# \ of \ occurrences \ of \ w_1...w_{n-1}}{the \ \# \ of \ occurrences \ of \ w_1...w_n}\right) \qquad (Eq.\ 5)$$

we get:

$$Averaged \quad Modified \quad n-gram \quad \Pr ecision = \sum_{n=1}^{N} \left\{ \frac{\sum_{all\_w_1...w_n\_that\_co-occur} Info(w_1...w_n)}{\sum_{all\_w_1...w_n\_in\_hyp} (1)} \right\} \qquad \text{(Eq. 6)}$$

$$BP = \exp\{\beta \log^2[\min(\frac{L_{hyp}}{L_{ref}},1)]\} \qquad \text{(Eq. 7)}$$

The brevity penalty is calculated as Eq. 7, where β is chosen to make the penalty=0.5 when L_{hyp} = 2/3 * L_{ref}. β =-4.22.

Despite the motivation to put more weights on those n-grams that are more "informative", the NIST metric fails to do so especially for the high-order n-grams. Zhang et al. (2004) observed that 80% of the NIST score for a typical MT system came from the unigram matches. Some 5-gram matches were given no credits because their information value is 0.

A particular feature of the NIST metric is that the scores increase with test set size. The reason for this is that when the test set size increases, the number of different n-grams, and thereby the information gain for each n-gram also increases. This leads to problems when comparing NIST scores. For example a system with NIST score 10.3 over a test set of 100 documents is not necessarily better than a system with NIST score 8.9 over a test set of 80 documents.

2.4. Human judgment

Human assessments were carried out by LDC for the test set used in the 2002 TIDES MT evaluation. Similar to the DARPA-94 MT evaluation (White 94), the human assessment was a holistic scoring by 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.

3. Confidence Intervals based on Bootstrap Percentiles

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. Bootstrapping is a data-based statistical method for statistical inference, which can be used to measure the confidence interval (Efron and Tibshirani, 1993).

3.1. Algorithm

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 reference 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 $t_i=<s_i, r_{i1}, r_{i2},..,r_{iR}>$, 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 this process for $\mathcal{B}$ times, e.g. $\mathcal{B}$=2000, and we have $\mathcal{B}$ new test suites: T_1... $T_\mathcal{B}$. T_1 to $T_\mathcal{B}$ are artificial test suites (also called *bootstrap samples*) created by resampling T_0. Evaluate the MT systems on each of these $\mathcal{B}$ test suite using any MT evaluation metric, like WER, BLEU, M-BLEU, NIST, or even human evaluation scores. We will then have $\mathcal{B}$ scores. As one may expect, these scores have a rough normal distribution. Figure 1 shows an example of the BLEU score distribution over 20000 resampled test suites for an MT system. From these $\mathcal{B}$ scores, find the middle 95% of the scores (i.e. the 2.5th percentile: $score_{low}$ and the 97.5th percentile $score_{up}$). $[score_{low}, score_{up}]$ is the 95% confidence interval for the used evaluation metric for this MT system (Figure 2).
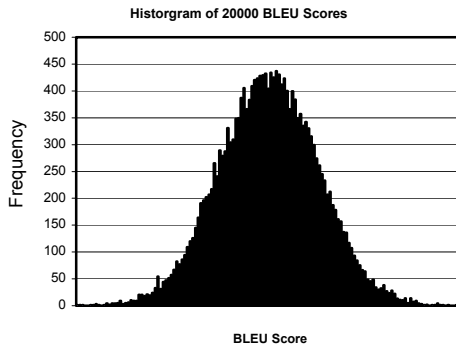

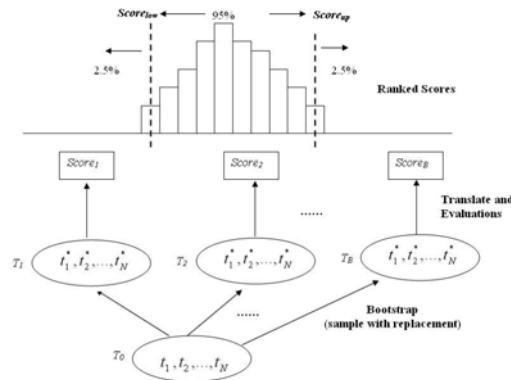Figure 1. The histogram of 20000 BLEU scores

Figure 2. Measuring the confidence intervals for a MT evaluation score

We evaluated 7 Chinese-English MT systems based on the June 2002 evaluation set. The testing data has 100 documents (878 sentences) and 4 human translations are used as references. We created 2000 bootstrapping samples for each system and report their median score and the 95% relative confidence intervals in Table 1. Relative confidence interval is defined as

$$\left[-\frac{Median - Score_{low}}{Median}\%, \quad +\frac{Score_{up} - Median}{Median}\% \right]$$

Results are given in Table 1. We see that we need different relative improvements for different metrics before we can claim to have made a statistically significant improvement in our machine translation system. It seems that the focus on longer n-grams in the BLEU metric make it less discriminative than the unigram centered NIST score. This is in line with the design principles of the NIST metric, to have high sensitivity when comparing different systems. M-BLEU stands in the middle. There seems also to be a tendency that better systems are more consistent in their translations, leading to smaller confidence intervals.

Table 1. The relative confidence intervals for 7 MT systems with $\mathcal{B}$=2000

System	NIST		BLEU		M-BLEU	
	Median	Interval	Median	Interval	Median	Interval
A	7.191	[-1.69%, +1.69%]	0.184	[-4.35%, +4.41%]	0.125	[-2.48%, +2.48%]
B	6.194	[-2.68%, +2.63%]	0.165	[-5.44%, +5.32%]	0.113	[-3.18%, +2.91%]
C	6.954	[-1.83%, +1.87%]	0.180	[-4.55%, +4.66%]	0.120	[-2.75%, +2.75%]
D	6.527	[-1.74%, +1.78%]	0.145	[-4.98%, +5.25%]	0.108	[-2.51%, +2.60%]
E	4.941	[-2.23%, +2.10%]	0.076	[-6.48%, +6.88%]	0.072	[-2.90%, +2.62%]
F	7.487	[-1.82%, +1.75%]	0.240	[-3.99%, +3.74%]	0.147	[-2.65%, +2.45%]
G	7.165	[-1.72%, +1.66%]	0.184	[-4.67%, +4.35%]	0.124	[-2.58%, +2.42%]

3.2. Comparing Two MT Systems

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.

Create test suites T_0, T_1... $T_\mathcal{B}$, where T_1 to $T_\mathcal{B}$ are artificial test suites created by resampling T_0. System X scored x_0 on T_0 and system Y scored y_0. The discrepancy between system X and Y is $\delta_0 = x_0 - y_0$. Repeat this process on every $\mathcal{B}$ test suite and we have $\mathcal{B}$ discrepancy scores: δ_1, δ_2...$\delta_\mathcal{B}$. From these $\mathcal{B}$ discrepancy scores, find the middle 95% of the scores (i.e. the 2.5[th] percentile and the 97.5[th] percentile). That is the 95% confidence interval for the discrepancy between MT system X and Y. If the confidence interval does not overlap with zero, we can claim that the difference between system X and Y are *statistically significant*.

In Figure 3 we compared 7 Chinese-English MT systems according to their Human, BLEU, NIST and M-BLEU scores. In this figure, ">" 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.
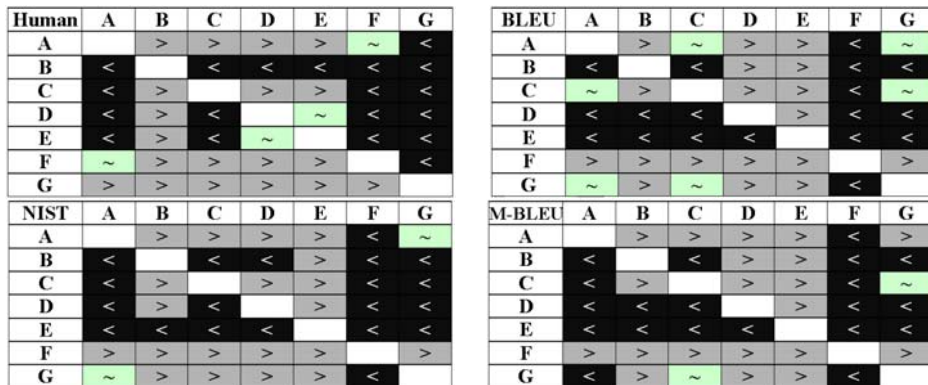
Human	A	B	C	D	E	F	G
A		>	>	>	>	~	<
B	<		<	<	<	<	<
C	<	>		>	>	<	<
D	<	>	<		~	<	<
E	<	>	<	~		<	<
F	~	>	>	>	>		<
G	>	>	>	>	>	>	

BLEU	A	B	C	D	E	F	G
A		>	~	>	>	<	~
B	<		<	>	>	<	<
C	~	>		>	>	<	~
D	<	<	<		>	<	<
E	<	<	<	<		<	<
F	>	>	>	>	>		>
G	~	>	~	>	>	<	

NIST	A	B	C	D	E	F	G
A		>	>	>	>	<	~
B	<		<	<	>	<	<
C	<	>		>	>	<	<
D	<	>	<		>	<	<
E	<	<	<	<		<	<
F	>	>	>	>	>		>
G	~	>	>	>	>	<	

M-BLEU	A	B	C	D	E	F	G
A		>	>	>	>	<	>
B	<		<	>	>	<	<
C	<	>		>	>	<	~
D	<	<	<		>	<	<
E	<	<	<	<		<	<
F	>	>	>	>	>		>
G	<	>	~	>	>	<	

Figure 3. Comparison among 7 Chinese-English MT systems by BLEU

3.3. Implementation

To calculate the confidence intervals using bootstrapping, we need to translate and evaluate the MT systems on each of the $\mathcal{B}$ test suites. $\mathcal{B}$ needs to be large, say, 1,000 or even 10,000, to guarantee reliable results. 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 $\mathcal{B}$ 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[1].

4. Discussions

Equipped with the bootstrapping method, we can now study in detail the effect of test set size and the number of reference translations on the width of the confidence interval.

4.1. How much testing data is needed?

For the TIDES MT evaluations the test set contain typically 100 documents, where a document has about 7~9 sentences on average. Do we really need nearly 1,000 test sentences to make the evaluation meaningful? Figures 4 to 7 show the NIST, BLEU, M-BLEU, and human evaluation score (sum of fluency and adequacy scores) for 7 different translation systems when the test set size varies from 10 to 100 documents, corresponding to about 10-100% of the entire test set. Each time 10 random documents were added to the existing test set in this ablation study.
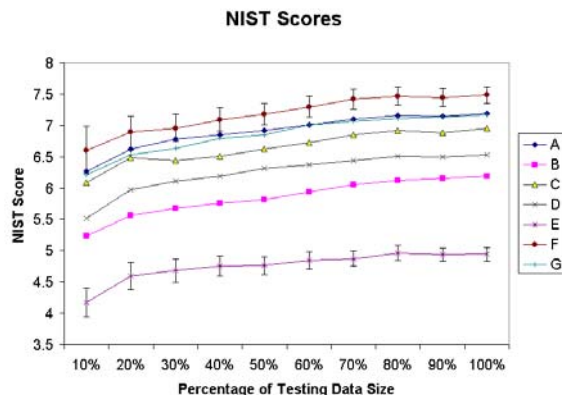

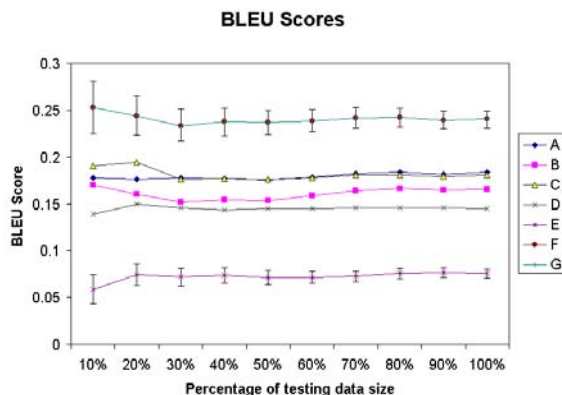Figure 4. NIST Scores for 7 MT systems over different size of testing data

Figure 5. BLEU Scores for 7 MT systems over different size of testing data

---

[1] The toolkit can be downloaded at
http://projectile.is.cs.cmu.edu/research/public/tools/bootStrap/tutorial.htm
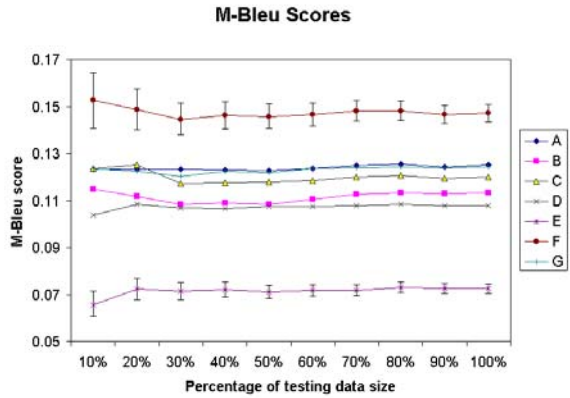
Figure 6. M-BLEU Scores for 7 MT systems over different size of testing data
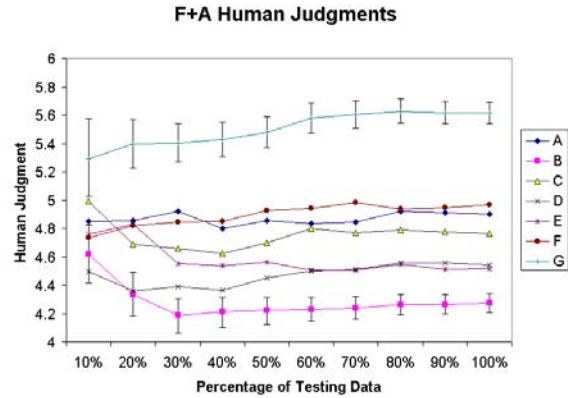

Figure 7. F+A Human Judgment Scores for 7 MT systems over different size of testing data

In Table 2, we repeated the ablation experiments for 100 times and calculated the averaged scores and corresponding confidence intervals. By doing this, the differences among documents are isolated from the effects of test set size.

Table 2. The relative confidence interval vs. the size of testing data for MT system A

Testing Data Size	Avg. NIST Score Interval	Avg. BLEU Score Interval	Avg. M-BLEU Score Interval	Avg. Human Score Interval
10%	[-4.96%, +4.93%]	[-13.40%, +13.66%]	[-7.36%, +7.72%]	[-4.65%, +4.77%]
20%	[-3.61%, +3.57%]	[-9.58%, +9.69%]	[-5.32%, +5.45%]	[-3.37%, +3.42%]
50%	[-2.35%, +2.35%]	[-6.06%, +6.14%]	[-3.39%, +3.46%]	[-2.18%, +2.21%]
80%	[-1.88%, +1.88%]	[-4.81%, +4.83%]	[-2.69%, +2.73%]	[-1.74%, +1.74%]
100%	[-1.69%, +1.69%]	[-4.35%, +4.41%]	[-2.48%, +2.48%]	[-1.61%, +1.51%]

For NIST scores (Figure 4) we see a steady increase with growing test set size, as expected. But the distance between the scores of the different systems remains stable when using 40% or more of the test set. Similarly, the BLEU and M-BLEU scores (Figures 5 and 6) stay within close bounds when using 30% or more of the test data. This would suggest that we need much less test data to evaluate the quality of the different MT systems. However, the confidence intervals change with the size of the test set (to make the figures clear, we show only the confidence intervals for the best and the worst systems). Therefore, we also need to ask how much data is needed to be able to confidently say that one MT system is better than another. From above tables and charts, we see that the relative confidence interval becomes narrower as the size of the testing data increases. A rough rule of thumb is: doubling the testing data size narrows the confidence interval by 30%. The confidence intervals for human evaluation scores are smaller than for any of the automatic metrics, even though the scores themselves vary more with test data size. One explanation could be that the translation quality is different for different documents, which is reflected in the human evaluation scores, less so captured with the automatic evaluation.

4.2. How many reference translations are needed?

Translations can vary widely. This is why many MT evaluations nowadays use several reference translations. Figure 8 and Table 3 show the effect of increasing the number of reference translations on the confidence intervals for the different metrics. The general observation is that the relative confidence interval becomes narrower with more reference translations, which is desired.   That is to say, more reference translations make the evaluation scores more discriminative.
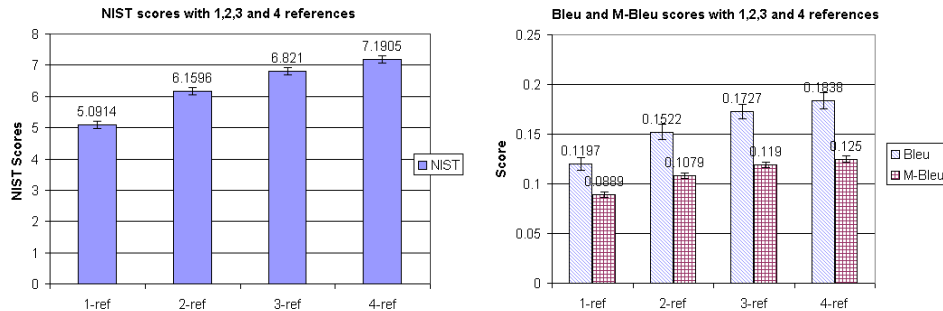

Figure 8. NIST, BLEU and M-BLEU scores for MT system A using 1, 2, 3 and 4 references

Table 3. NIST, BLEU and M-BLEU scores for MT system A using 1,2,3 and 4 references

	NIST		BLEU		BLEU	
	Median	Intervals	Median	Intervals	Median	Intervals
1-ref	5.091	[-2.09%, +2.16%]	0.120	[-5.18%, +5.60%]	0.089	[-2.70%, +3.04%]
2-ref	6.160	[-1.89%, +2.01%]	0.152	[-4.73%, +4.99%]	0.108	[-2.59%, +2.69%]
3-ref	6.821	[-1.78%, +1.70%]	0.173	[-4.40%, +4.28%]	0.119	[-2.52%, +2.44%]
4-ref	7.191	[-1.69%, +1.69%]	0.184	[-4.35%, +4.41%]	0.125	[-2.48%, +2.48%]

Increasing the testing data size as well as using more reference translations increases the precision of the evaluation metrics, i.e. narrows down the confidence interval. Table 4 shows the relative confidence intervals of NIST scores for system A, using 1, 2, 3 and 4 references over different size of testing data. An interesting observation is that 100% testing data with 1 reference is equivalent to 80~90% of testing data with 2 references, or 70~80% of testing data with 3 references, or 60~70% of testing data with four references. In other words, adding an additional reference translation will compensate the effects of removing 10~15% of the testing data on the relative confidence interval. Therefore, it seems more cost effective to have more test sentences but fewer reference translations. This observation also holds for BLEU and M-BLEU scores.

Table 4. The effects of the reference number and testing data size

System	1-ref	2-ref	3-ref	4-ref
50%	[-2.91%, +3.01%]	[-2.65%, +2.55%]	[-2.47%, +2.63%]	[-2.36%, +2.36%]
60%	[-2.64%, +2.66%]	[-2.36%, +2.38%]	[-2.29%, +2.37%]	[-2.16%, +2.12%]
70%	[-2.57%, +2.49%]	[-2.25%, +2.24%]	[-2.16%, +2.08%]	[-1.90%, +2.09%]
80%	[-2.44%, +2.33%]	[-2.13%, +2.08%]	[-1.96%, +2.08%]	[-1.86%, +1.92%]
90%	[-2.32%, +2.21%]	[-2.02%, +1.92%]	[-1.80%, +1.84%]	[-1.85%, +1.85%]
100%	[-2.09%, +2.16%]	[-1.89%, +2.01%]	[-1.78%, +1.70%]	[-1.69%, +1.69%]

4.3. How many bootstrap samples are needed?

Finally, we studied the stability of the bootstrapping approach. That is, how many bootstrap samples do we need to have reliable confidence intervals? From Table 5 we can see that there are only small changes when going beyond $\mathcal{B}$=2,000 samples and with $\mathcal{B}$=20,000 we are close to convergence.

Table 5. Confidence intervals for system A with different bootstrapping sample size

$\mathcal{B}$	NIST		BLEU		M-BLEU	
	Median	Interval	Median	Interval	Median	Interval
1,000	7.191	[-1.59%, +1.61%]	0.184	[-4.02%, +4.40%]	0.125	[-2.24%, +2.32%]
2,000	7.188	[-1.65%, +1.74%]	0.184	[-4.19%, +4.46%]	0.125	[-2.32%, +2.48%]
5,000	7.192	[-1.65%, +1.70%]	0.184	[-4.24%, +4.40%]	0.125	[-2.32%, +2.48%]
10,000	7.190	[-1.61%, +1.70%]	0.184	[-4.24%, +4.35%]	0.125	[-2.32%, +2.48%]
20,000	7.192	[-1.69%, +1.68%]	0.184	[-4.29%, +4.29%]	0.125	[-2.40%, +2.48%]
50,000	7.191	[-1.68%, +1.69%]	0.184	[-4.29%, +4.29%]	0.125	[-2.40%, +2.40%]
100,000	7.191	[-1.68%, +1.68%]	0.184	[-4.29%, +4.29%]	0.125	[-2.40%, +2.48%]

For an MT system and a fixed $\mathcal{B}$, each bootstrapping test yields a different confidence interval because the bootstrap samples are generated randomly. We conducted 5,000 bootstrapping tests for system A using e.g. $\mathcal{B}$=500. This process resulted in a population of 5,000 confidence intervals. Standard deviation was then calculated for this population as a meta "confidence interval" for the relative confidence intervals (Table 6). For example, when $\mathcal{B}$=500, the lower bound of the NIST confidence interval has mean –1.70%, and STDEV=0.11 percentage point. In other words, for 95% of chances, the lower bound of relative confidence interval falls into the range of [-1.92%, -1.48%]. We can see that when $\mathcal{B}$ is greater than 2,000, the relative confidence intervals are pretty reliable.

Table 6. STDEV of the relative confidence intervals for system A with different $\mathcal{B}$

$\mathcal{B}$	NIST		BLEU		M-BLEU	
	STDEV	Interval	STDEV	Interval	STDEV	Interval
100	0.0024	[-1.77%, +1.60%]	0.0060	[-4.50%, +4.11%]	0.0034	[-2.51%, +2.31%]
500	0.0011	[-1.70%, +1.66%]	0.0027	[-4.33%, +4.27%]	0.0015	[-2.42%, +2.40%]
1,000	0.0007	[-1.68%, +1.68%]	0.0019	[-4.29%, +4.32%]	0.0012	[-2.40%, +2.43%]
2,000	0.0005	[-1.68%, +1.68%]	0.0013	[-4.29%, +4.32%]	0.0008	[-2.40%, +2.43%]
10,000	0.0002	[-1.68%, +1.68%]	0.0006	[-4.28%, +4.32%]	0.0004	[-2.39%, +2.44%]

5. Future Work

The current study was based on one test set which has been used in the TIDES translation evaluation. We believe that the observations will be similar on other test sets. In the future we plan to apply the bootstrap approach to other evaluation metrics and to rank them according to their confidence intervals. In addition, we will extend our current study to other test situations:

- Domain specific translation systems, notably speech translation systems, where the vocabulary is typically much smaller;

- Comparing with other types of MT systems (transfer, interlingua), as n-gram type metrics seem to favor SMT and EBMT systems.
- Different translation pairs.

6. References

M. Bisani and H. Ney : 2004, 'Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation', In *Proc. of ICASP*, Montreal, Canada, Vol. 1, pp. 409-412.

Chernick, Michael R: 1999, *Bootstrap Methods,* Wiley Series in Probability and Statistics, Applied Probability and Statistics Section, John Wiley & Sons, Inc.

Culy, Christopher & Riehemann, Susanne Z.: 2003, 'The Limits of N-Gram Translation Evaluation Metrics', In *Proc. of the 9th MT-Summit*. New Orleans, LA, USA.

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

G. Leusch, N. Ueffing, H. Ney : 2003, 'A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation', In *Proc.* 9th *MT Summit*, New Orleans, LA.

I Dan Melamed, Ryan Green and Joseph P. Turian : 2003, 'Precision and Recall of Machine Translation', In *Proc. of NAACL/HLT 2003*, Edmonton, Canada.

King M., Popescu-Belis A. & Hovy E. : 2003, 'FEMTI: creating and using a framework for MT evaluation', In *Proc. of 9th Machine Translation Summit*, New Orleans, LA, USA.

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

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', In *Proc. of the 20th ACL*.

Pepescu-Belis, Andrei : 2003, 'An Experiment in Comparative Evaluation: Humans vs. Computers', In *Proc. of the 9th MT-Summit*. New Orleans, LA 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.

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.

Ying Zhang, Stephan Vogel, Alex Waibel : 2004, 'Interpreting BLEU/NIST scores: How much improvement do we need to have a better system?,' In *Proc. of LREC 2004*, Lisbon, Portugal.