A New Method for the Study of Correlations between MT Evaluation Metrics

Paula Estrella
ISSCO/TIM/ETI
University of Geneva
40, bd. du Pont-d'Arve
1211 Geneva, Switzerland
paula.estrella@issco.unige.ch

Andrei Popescu-Belis
ISSCO/TIM/ETI
University of Geneva
40, bd. du Pont-d'Arve
1211 Geneva, Switzerland
andrei.popescu-belis@issco.unige.ch

Maghi King
ISSCO/TIM/ETI
University of Geneva
40, bd. du Pont-d'Arve
1211 Geneva, Switzerland
Maghi.King@gmail.com

Abstract

This paper aims at providing a reliable method for measuring the correlations between different scores of evaluation metrics applied to machine translated texts. A series of examples from recent MT evaluation experiments are first discussed, including results and data from the recent French MT evaluation campaign, CESTA, which is used here. To compute correlation, a set of 1,500 samples for each system and each evaluation metric are created using bootstrapping. Correlations between metrics, both automatic and applied by human judges, are then computed over these samples. The results confirm the previously observed correlations between some automatic metrics, but also indicate a lack of correlation between human and automatic metrics on the CESTA data, which raises a number of questions regarding their validity. In addition, the roles of the corpus size and of the selection procedure for bootstrapping (low vs. high scores) are also examined.

1 Introduction

One of the design principles of automatic MT evaluation metrics is that their scores must "correlate" with a reliable measure of translation quality, generally estimated by human judges. Indeed, the claim that an automatic scoring procedure applied to MT output can provide an accurate view of translation quality must be substantiated by a proof that the scores do reflect genuine quality, as perceived by human users of a translation. For instance, the proponents of BLEU or WNM (Babych and Hartley, 2004; Papineni et al., 2001) have compared the scores produced by their metrics – which compare n-grams of MT-generated sentences with one or more reference translations produced by humans – with adequacy and fluency scores assigned by human judges.

It is not, of course, that all metrics of translation quality *must* be correlated. Although adequacy (i.e. fidelity or "semantic correctness") and fluency (acceptability as a valid sample of the target language) do seem correlated to some extent (White, 2001), one can easily imagine MT output with high fluency but low adequacy. However, an automatic MT evaluation metric should at least correlate with one quality characteristic on which human judges would reliably agree, which can be some aspect of intrinsic quality, or a utility-based measure with respect to a given task.

Given the low cost of automatic metrics, they have been widely used in recent experiments, three of which are discussed in Section 5. However, the results obtained on the correlation between metrics that were used are difficult to compare, and therefore the reliability of automatic metrics is hard to assess.

In this article, we propose a method to measure the correlation between two MT evaluation metrics based on bootstrapping (Section 3) and apply it to data from the recent French MT evaluation campaign, CESTA

(Section 4). Our experiments (Section 5) analyze the correlation between metrics and show that correlation is lower than expected for automatic *vs*. human metrics. The experiments also show that correlation varies with sample size, as well as with the subset of sentences that is considered (low vs. high quality). Samples from the two CESTA runs indicate however that correlations do not vary significantly with a different translation domain.

2 Correlation between MT Evaluation Metrics in Previous Experiments

Many authors report on the correlation between human and automated metrics: some working at the sentence level (Kulesza and Shieber, 2004; Russo-Lassner et al., 2005), and some at the corpus level (Doddington, 2002; Papineni, 2002), in a variety of approaches and setups. Recent experiments, for instance, report that the correlation of the well-known BLEU metric with metrics applied by humans is not always as high as previously reported (Callison-Burch et al., 2006). In this section, we analyze three recent contributions that illustrate clearly the variety of methodologies used to compute correlations between metrics.

2.1 An Experiment with the Europarl Corpus

Koehn and Monz (2006) describe the competition organized during the Statistical MT Workshop at NAACL 2006. Its main goal was to establish baseline performance of MT evaluation for specific training scenarios. The test corpus consisted of sentences from the Europarl corpus (Koehn, 2005) and from editorials of the Project Syndicate website, and contained a total of 3,064 sentences. The translation directions were SP↔EN, FR↔EN, DE↔EN and there were 14 participating systems.

The BLEU metric was used for automatic evaluation, as the most commonly used metric in the MT community. To provide human quality judgments, the workshop participants had to assess 300–400 sentences each, in terms of adequacy and fluency, on a 5-point scale. Each evaluator was in fact simultaneously given 5 machine translations, one reference translation, and one source sentence, and was asked to perform a comparative evaluation of the machine translations. The scores for adequacy and fluency were then normalized and were finally converted into rankings, to increase robustness of the conclusions.

The similarity between the performances of the systems and the problems encountered in the human evaluation made it difficult to draw strong conclusions about the correlation of human and automatic metrics. Some evaluators explicitly pointed out how difficult it was to maintain consistency of judgment, especially when the sentences are longer than average. Evaluators also suggested extending the scale for adequacy scores, as this would improve the reliability of judgments.

2.2 Reliability and Size of Test Set

Coughlin (2003) reports results on the correlation between human assessments of MT quality and the BLEU and NIST metrics (Doddington, 2002) in a large scale evaluation, using data collected during two years. The judges were neither domain experts (in computer science), nor were they involved in the development of the participating systems. Having access only to high quality reference translations, they had to rate sentences in pairs, to compare two different systems. The innovative methodology of human evaluation was to rate the overall *acceptability* of the sentences – and not their adequacy or fluency – on a 4-point scale, without further instructions, thus generating only one human score per sentence.

The sentences were evaluated by 4–7 judges, leading to an average inter-rater agreement of 0.76 for EN→DE and 0.83 for FR→EN.

Contrary to the work described in the previous subsection, Coughlin (2003) found a very high correlation between the BLEU metric and the human judges, especially when test data sets comprise more than 500 sentences. For the NIST metric, on the contrary, correlation is lower for data sets that comprise more than 250 sentences. In general, Coughlin (2003) shows a high correlation between BLEU/NIST and human scores, for all language pairs and systems used, except for the FR→EN pair which had low negative correlation, for which they suggest that the Hansard domain might be more difficult to translate for the systems under evaluation.

2.3 Correlations in the CESTA Campaign

The French MT evaluation campaign, CESTA, also reported results on the meta-evaluation of automatic metrics, i.e. their comparison to the human scores of adequacy and fluency (Hamon et al., 2006). The data used for the evaluation is described in detail in Section 4, since it is also used in this paper. The main automatic metrics used in CESTA are BLEU, NIST, Weighted N-gram Metric (WNM) (Babych, 2004), mWER (Niessen et al., 2000), and mPER (Tillmann et al., 1997).

CESTA used human judges to assign adequacy and fluency scores on a 5-point scale with a protocol and interfaces that changed from the first to the second run. The rating scale in the first run explicitly listed the intermediate labels for the values, while for the second run the labels were removed. In addition, while in the first run the evaluation of adequacy and fluency was done at the same time, in the second run, the judges scored every segment separately for fluency and for adequacy. In both runs the final scores for each sentence are the average of two assessments.

When defined as the percentage of identical values from the 5-point scale, the inter-judge agreement is only 40% for fluency, and varies from 36% to 47% for adequacy in the first vs. second run (EN→FR). However, when defined as the percentage of scores that differ by at most one point between two judges (e.g. a segment rated 3 by one judge and 2 by the other would count as an agreement), inter-judge agreement increases significantly, to 84% for fluency and 78% for adequacy. Moreover, the CESTA campaign reports acceptable correlation between automatic metrics and adequacy/fluency, when computed over the five participating systems, that is, as the Pearson correlation of five pairs of values. For example, the correlation of NIST (or BLEU) with fluency is around 0.67 in the first run[1].

3 Using Bootstrapping to Study the Correlation between Metrics

We propose here the use of bootstrapping to investigate the correlation between the scores of different metrics on a *per system* basis, and not only between the various systems participating in an evaluation. To calculate the correlation between two or more variables (metrics in this case), we need two or more samples of each variable: for example, in an evaluation campaign, the samples are the final scores obtained by each system, which are then correlated to explore relations between different metrics (cross-system correlation). Our approach consists of (artificially) generating several sample scores of the same system and calculating the correlations of two metrics over the set of samples, for that particular system. The advantages of this method are that we only need the output of one system and that the results obtained are specific to that system. The disadvantage is of course, that direct comparison with standard cross-system correlation is not possible, since we only consider one system at a time.

Therefore, this method can be used to estimate the correlation of metrics as the result of evaluating one system only, and can include of course any kind of metrics, human and automatic, in the analysis.

3.1 Bootstrapping Samples of Scores

Bootstrapping is a statistical technique that is used to study the distribution of a variable based on an existing set of values (Efron and Tibshirani, 1993). This is done by randomly resampling *with replacement* (i.e. allowing repetition of the values) from the full existing sample and computing the desired parameters of the distribution of the samples. The method has the practical advantage of being easy to implement and the theoretical advantage of not presupposing anything about the underlying distribution of the variable. A simple programming routine can thus calculate the estimators of the mean, variance, etc., of any random variable distribution.

Moreover, when the original sample is resampled a large number of times, the law of large numbers ensures that the observed probability approaches (almost certainly) the actual probability. Also, when N is sufficiently large, the sample scores are quite close to the normal distribution, as illustrated in Figure 1.

The bootstrapping algorithm can be summarized as follows:

[1] The CESTA final report provides the detailed scores: http://technolangue.net/IMG/pdf/Rapport_final_CESTA_v1.04.pdf.

1. Given a sample $X = (X_1, X_2, \ldots, X_n)$ from a population **P**, generate N random samples (noted X^*) of the same size by drawing n values from the sample, with replacement (each value having probability $1/N$).
2. The resulting population $\mathbf{P}^*$, noted $X^* = (X_1^*, \ldots, X_N^*)$, constitutes the N bootstrapped samples.
3. If the original estimator of a given population parameter was $\theta(X)$, with the bootstrapped samples we can calculate the same estimator as $\theta(X^*)$.

An important parameter for bootstrapping is N, the number of bootstrapped samples, i.e. the number of times the process is repeated. This number should be large enough to build a representative number of samples. It appears that, for instance, $N = 200$ leads to slightly biased estimations (Efron and Gong, 1983; Efron and Tibshirani, 1993; Koehn, 2004; Zhang et al., 2004, so N ~ 1,000 is preferred, for example N = 1,000) or even $N = 10{,}000$ (Bisani and Ney, 2004). Based on these examples, we decided to use $N = 1{,}500$ bootstrapped samples.

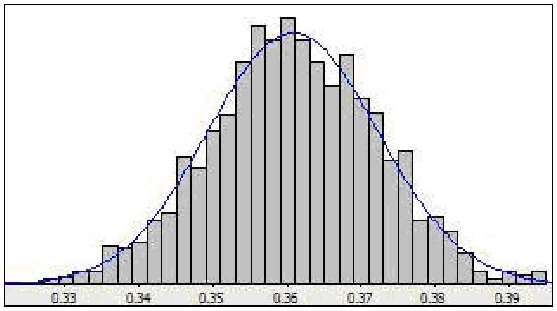

Figure 1. Example of histogram for the WER scores obtained with 1,500 bootstrapped samples (CESTA scores, first run, system S2)

3.2 Application to MT Evaluation Scores

In the MT field, bootstrapping has been mainly used to estimate confidence intervals for automatic metrics and to compute the statistical significance of comparative performance of different MT systems, e.g. using the BLEU (Koehn, 2004; Kumar and Byrne, 2004; Zhang et al., 2004) or WER metric (Bisani and Ney, 2004). Here, bootstrapping will be used to compute the correlation between metrics for MT. These correlations will be studied for each system, i.e. they are calculated on a *per system* basis as opposed to the common cross-system correlation.

Since correlation concerns two sets of scores, we need to apply the metrics simultaneously to the same bootstrapped samples to keep consistency in the scores. Put in simpler words, we apply two (or more) different metrics to the same random sample per iteration of the bootstrapping process. A *random sample* is a set of segments randomly selected from the corpus and of the same size of the corpus used in the evaluation.

Described in pseudo code, the routine computing correlation is particularly simple: M is the number of segments to be considered, N is the numbers of iterations, `sample[m]` is the m-th element of the random sample and `sample*` is the complete bootstrapped sample:

```
for(n=0; n<N; n++){
    for(m=0; m<M; m++){
       sample[m] = selectRandSeg();
    }
    scoresA[n]=calcMetricA(sample*);
    scoresB[n]=calcMetricB(sample*);
}
   calcCorrelation(scoresA, scoresB);
```

4 Evaluation Resources: Data, Systems and Metrics

For the experiments presented here, we used the resources of the EN→FR translation task in the CESTA MT evaluation campaign (Hamon et al., 2006). In all cases, the results of the participating systems are anonymized, therefore the systems will simply be referred to by the codes S1 to S5 in no particular order.

One of the goals of the first run was to validate the use of automatic evaluation metrics with French as a target language, by comparing the results of well-known automatic metrics with fluency and adequacy scores assigned by human judges. The test data for the first run consisted of 15 documents from the Official Journal of the European Communities (JOC, 1993) with a total of 790 segments and an average of 25 words per segment. The documents contain transcribed questions and answers in a parliamentary context, and since no particular domain was targeted when putting together the corpus, the CESTA campaign considered this as *general domain* data. Five systems participated in the EN→FR first run, both commercial and research ones.

For the second run, the goal was to improve the evaluation protocols used in the first run and to observe the impact of system adaptation to a particular domain. Therefore, the *medical domain* was chosen, using data collected from the *Santé Canada* website, with a total of 288 segments and an average of 22 words per segment. Almost the same systems participated in the second run.

In addition to the automatic metrics used in the CESTA campaign, we included in our experiment precision and recall from the General Text Matcher (Turian et al., 2003).

5 Experimental Study of Correlation

Although we performed the study using all the systems participating in the CESTA campaign, we will only present here the results of two systems, namely S2 and S5, chosen among the best. In Section 5.1, we compute correlations between metrics on two test sets of dissimilar size, in Section 5.2 we study the correlations for segments of very high and very low adequacy scores and, finally, in Section 5.3 we present the results of the correlations for a test set of a different domain.

5.1 Correlation Values and the Influence of the Size of Test Data

In the first experiment, we compared correlation between metrics, when calculated on a test set of 5 documents and on a larger set of 15 documents from the general domain corpus. We hypothesize that if a strong correlation exists between two score sets, it should be stable, i.e. it should be similar or even higher, when using a larger test set.

Tables 1 to 4 show the Pearson R coefficients for all the metrics applied in this study, separately for systems S5 (Tables 1 and 2) and S2 (Tables 3 and 4). The correlation figures were computed on 5 documents in Tables 1 and 3, and respectively on 15 documents in Tables 2 and 4. Negative values generally occur when the metrics vary in the opposite direction, e.g. higher scores of the first one correspond (correctly) to lower scores of the second one.

As we expected, there is a relatively high correlation between metrics of the same type (except for adequacy and fluency for S5) regardless of the size of the test data set: for instance, the following correlations between metrics appear to be quite high: WER *vs*. PER > 0.81, BLEU *vs*. NIST > 0.72, PREC *vs*. REC > 0.76. However, the figures show also that automatic metrics correlate better with other automatic metrics than with adequacy or fluency; for both systems, the NIST metric presents the lowest coefficients.

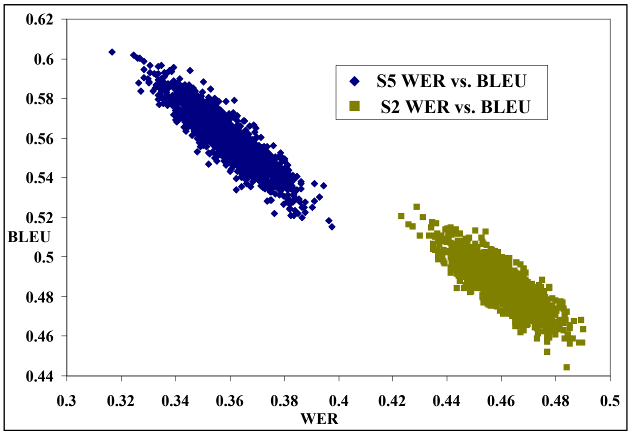

Figure 2. Scatter plot of WER vs. BLEU bootstrapped scores using 5 documents

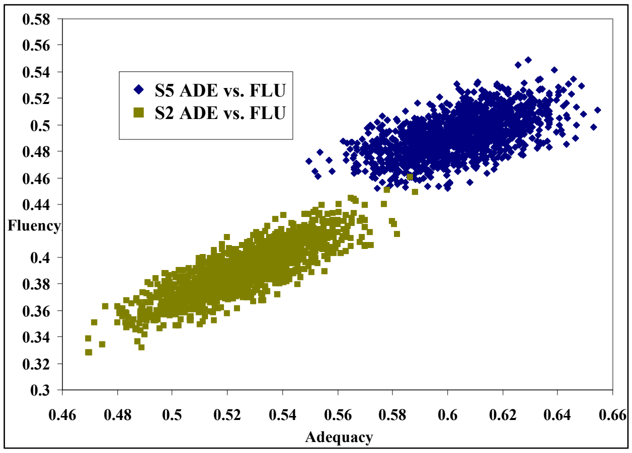

Figure 3. Scatter plot of adequacy vs. fluency bootstrapped scores using 5 documents

Regarding the change in the size of the test data, the correlations (excluding adequacy *vs*. fluency) for S2 systematically increase when using 15 documents with respect to 5. However, this is less clear for S5: the correlation of NIST with all other metrics increases, BLEU *vs*. WER/PER remains stable, but the correlations between automatic metrics and the human ones decrease, quite considerably in some cases, e.g. BLEU *vs*. fluency. This is probably due to the particular documents selected, since scores vary more on small test sets, as shown in (Estrella et al., 2007).

A graphical representation of the scores appears in Figures 2 to 5, which plot two scores for each of the 1,500 bootstrapped samples, for systems S2 (light/green) and S5 (dark/blue). Figure 2 illustrates two metrics that are highly correlated, BLEU and WER: the clouds of dots are organized along a line, which has negative slope as lower WER corresponds to higher BLEU (and to better performance, in principle). The correlation coefficients for the samples in Figure 2 are respectively -0.83 and -0.89.

A similar, albeit lower, correlation appears in Figure 3 for the two human metrics, adequacy vs. fluency. Again, the clouds of dots are organized along lines, this time with positive slopes. The correlation coefficients are respectively 0.84 and 0.58 for S2 and S5, the lower value for S5 being quite visibly reflected in the more scattered pattern of blue dots (less linear and more rounded shape).

S5	WER	PER	BLEU	NIST	ADE	FLU	PREC	REC
WER	1	0.93	-0.90	-0.69	-0.42	-0.43	-0.72	-0.56
PER		1	-0.89	-0.76	-0.40	-0.41	-0.84	-0.68
BLEU			1	0.83	0.39	0.44	0.82	0.71
NIST				1	0.26	0.27	0.87	0.68
ADE					1	0.58	0.34	0.39
FLU						1	0.34	0.37
PREC							1	0.79
REC								1

Table 1. Correlation matrix for S5 using 5 documents

S5	WER	PER	BLEU	NIST	ADE	FLU	PREC	REC
WER	1	0.92	-0.90	-0.75	-0.28	-0.32	-0.74	-0.55
PER		1	-0.89	-0.79	-0.25	-0.29	-0.84	-0.65
BLEU			1	0.86	0.25	0.29	0.83	0.66
NIST				1	0.16	0.16	0.86	0.64
ADE					1	0.63	0.25	0.30
FLU						1	0.24	0.26
PREC							1	0.78
REC								1

Table 2. Correlation matrix for S5 using 15 documents

S2	WER	PER	BLEU	NIST	ADE	FLU	PREC	REC
WER	1	0.81	-0.83	-0.52	-0.48	-0.46	-0.61	-0.41
PER		1	-0.73	-0.60	-0.43	-0.42	-0.75	-0.54
BLEU			1	0.72	0.43	0.41	0.74	0.61
NIST				1	0.13	0.13	0.84	0.58
ADE					1	0.84	0.27	0.32
FLU						1	0.26	0.30
PREC							1	0.76
REC								1

Table 3. Correlation matrix for S2 using 5 documents

S2	WER	PER	BLEU	NIST	ADE	FLU	PREC	REC
WER	1	0.83	-0.85	-0.59	-0.49	-0.49	-0.64	-0.50
PER		1	-0.81	-0.69	-0.44	-0.43	-0.79	-0.61
BLEU			1	0.79	0.43	0.43	0.78	0.65
NIST				1	0.23	0.20	0.86	0.61
ADE					1	0.79	0.30	0.35
FLU						1	0.28	0.33
PREC							1	0.77
REC								1

Table 4. Correlation matrix for S2 using 15 documents

5.2 Correlation for High and Low Quality Translations

The findings from the previous section can be due to many factors; for example, using a corpus containing segments of diverse translation difficulty or using the average of two judgments for adequacy or fluency might give less informative results, since the final scores are calculated on the entire test set. Or it might be, as pointed out by Coughlin (2003), that humans could be influenced by the reference translation they see during the evaluation and therefore evaluate systems depending more on the algorithm they use (statistical or rule-based) than on their intrinsic quality.

To further investigate the correlations described in Sections 5.1, we carried out another experiment, focusing on the highest and lowest scores assigned by adequacy judgments. The goal is to explore the agreement among some metrics when the adequacy scores are very high and very low. An *a priori* hypothesis is that low quality translations might be more difficult to evaluate (leading to a larger variation of scores) than high quality translations. According to this hypothesis, the correlation between metrics applied on almost perfect segments should be stronger than that of metrics applied on low quality segments. We consider "quality" in terms of the score provided by human judges of adequacy, fluency or the average of both; for the purpose of this experiment we take adequacy as the measure of quality, but results using fluency or the average do not change dramatically.

Each segment of the CESTA data was evaluated for adequacy and for fluency by two judges, and the final scores for each metric are the average between the two assessments. These scores were then normalized and converted from a 5-point scale to a value between 0 and 1. To find only the segments with high adequacy score, we extracted, from the 15 documents of the first run, those segments with an average adequacy score above 0.825. For the low quality test set, we extracted the segments with an average adequacy below 0.125. We tried to keep the size constant, so we had around 130 segments in both new test sets, given that S5 had the least number of segments below 0.125. These empirical cut-off limits should also account for high inter-judge agreement, since a high/low score can only be reached if both assessors assigned similar high/low scores for the same segment.

To simplify the experiment, we only applied the WER and PER metrics to the corresponding outputs of S2 and S5. Tables 5 and 6 show the resulting R coefficients, the lower part of the tables corresponding to S2 and the upper part to S5 (for compactness reasons).

S5 **S2**	WER	PER	FLU	ADE
WER		0.93	-0.17	-0.25
PER	0.71		-0.13	-0.28
FLU	-0.14	-0.11		*-0.13*
ADE	-0.09	-0.14	0.16	

Table 5. Correlations on the low-adequacy data set: S2 lower-left half, S5 upper-right

S5 **S2**	WER	PER	FLU	ADE
WER		0.94	-0.17	-0.32
PER	0.93		-0.27	*-0.10*
FLU	-0.43	-0.39		0.42
ADE	-0.36	-0.30	0.41	

Table 6. Correlation on the high-adequacy data set: S2 lower-left half, S5 upper-right

The correlations clearly increase in absolute value from low-adequacy to high-adequacy segments, as hypothesized, but are still much weaker than expected for high-adequacy segments. Two special cases with extremely low correlation values are marked in italics, namely fluency *vs.* adequacy in Table 5 and PER *vs.* adequacy in Table 6, respectively. In the first case, we manually inspected the results of the bootstrapping procedure, and observed that adequacy scores were much lower than the fluency scores. Figures 4 and 5 provide a graphical representation of these two cases.

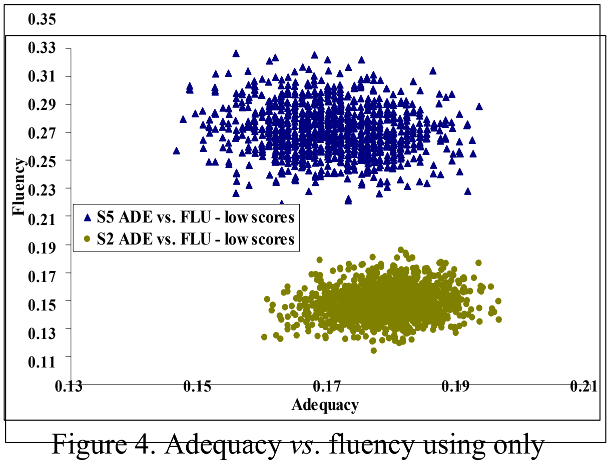

Figure 4. Adequacy *vs.* fluency using only segments with low adequacy scores

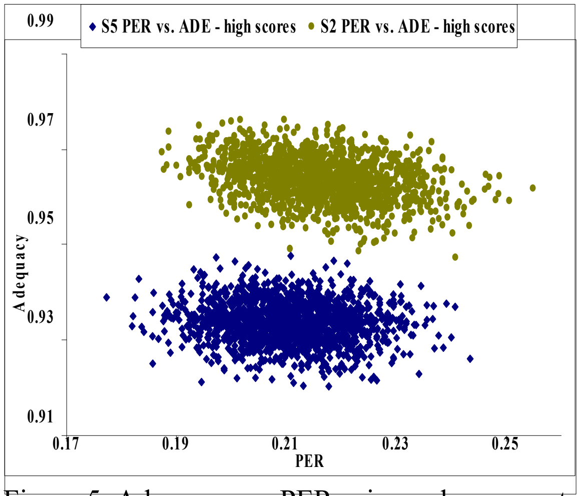

Figure 5. Adequacy *vs.* PER using only segments with high adequacy scores

For the PER *vs.* adequacy correlation, we found out that S2 has more segments scoring less than 0.125 (90 segments *vs.* 57 for S5) but has also more segments scoring 1 (121 segments vs. 93 for S5). This explains the scatter plot in Figure 5 but contradicts the expected results, since S5 was ranked among the best in the CESTA campaign. In overall scores this situation could be changed because the scores are averaged out. In practice, we believe that the difference between coefficients of -0.10 and -0.13 does not have a big impact, since one system provides clearly better translations than the other.

5.3 Correlations on a Different Domain

The last experiment consists of comparing the correlations obtained for test sets in a different domain than the previous one. For the second run of the CESTA campaign, the participants had the opportunity to train or adapt their systems to a particular domain (medical) using a special corpus for that purpose. Given that systems were trained for that specific domain, performance should have increased, as well as correlations between some metrics. Using the test corpus created for the second run of CESTA (288 segments), the results are comparable, in terms of size, to those obtained in Section 5.1 for 5 documents (270 segments).

Results for S2 an S5 are reported respectively in Tables 7 and 8. For the human metrics, results are not directly comparable to those of the previous sections due to a change in the evaluation protocols from the first run of the campaign to the next. Unfortunately, it appears that correlation coefficients remain quite low, despite the adaptation. In Table 7 we observe a significant increase in correlation coefficients between automatic metrics and adequacy for S2; this difference between S5 and S2 might indicate a failure of S5 to fully acquire the relevant vocabulary for the new domain. Following the hypothesis of the previous section and recalling that S2 was ranked below S5 in the CESTA campaign, it appears that assessment of low quality segments leads to more variation of scores, thus resulting in low correlation coefficients.

S2	WER	PER	BLEU	NIST	ADE	FLU	PREC	REC
WER	1	0.98	-0.87	-0.72	-0.72	-0.27	-0.69	-0.77
PER		1	-0.81	-0.69	-0.70	-0.26	-0.67	-0.83
BLEU			1	0.84	0.68	0.36	0.77	0.47
NIST				1	0.51	0.24	0.68	0.40
ADE					1	0.27	0.50	0.52
FLU						1	0.27	0.15
PREC							1	0.35
REC								1

Table 7. Correlation matrix for S2 using corpus from health domain

S5	WER	PER	BLEU	NIST	ADE	FLU	PREC	REC
WER	1	0.87	-0.82	-0.67	-0.20	-0.28	-0.66	-0.29
PER		1	-0.80	-0.75	-0.18	-0.20	-0.78	-0.44
BLEU			1	0.80	0.17	0.21	0.74	0.48
NIST				1	0.21	0.21	0.85	0.63
ADE					1	0.34	0.18	0.13
FLU						1	0.15	0.12
PREC							1	0.64
REC								1

Table 8. Correlation matrix for S5 using corpus from health domain

6 Conclusion and Future Work

The method presented in this paper allows the computation of correlation between two metrics on a single system, using bootstrapping to create a large set of samples of variable qualities.

Observations clearly indicate that some related automatic metrics, such as BLEU and NIST, or BLEU and WER, are better correlated than automatic vs. human metrics. However, even for related metrics, the correlation is not necessarily very high.

It is quite surprising that, using this method, correlations between human and automatic metrics are much lower than figures obtained by other methods and published as arguments for the reliability of automatic metrics.

At this stage, it is not yet clear, which is the main factor that explains such a low correlation, and whether these figures do indicate a significant *lack of correlation* on the CESTA scores that we examined. For instance, these figures could be related to low inter-rater agreement between the two judges of adequacy and fluency, which is not compensated by the use of the average values or to the fact that these automatic metrics are not suitable for the evaluation of morphologically richer languages, such as French.

Future work in this direction will examine how human scores used in our experiments are distributed among systems. Of course, adding new human judgments of the same MT output could help to increase our confidence in adequacy and fluency, but this operation is quite costly. We also plan to repeat some of the experiments with other automatic metrics, which claim to improve some of the metrics used here and to improve correlation with human scores.

References

Babych, B. 2004. Weighted N-gram model for evaluating Machine Translation output. In *CLUK 2004*. Birmingham, UK.

Babych, B., and T. Hartley 2004. Extending the BLEU MT Evaluation Method with Frequency Weightings. In *ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics: ACL 2004)*, 621-628. Barcelona, Spain.

Bisani, M., and H. Ney 2004. Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation. In *IEEE International Conference on*

Acoustics, Speech, and Signal Processing, 409-412. Montreal, Canada.

Callison-Burch, C., M. Osborne, and P. Koehn 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, 249-256. Trento, Italy.

Doddington, G. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *HLT 2002 (Second Conference on Human Language Technology)*, 128-132. San Diego, CA.

Efron, B., and G. Gong 1983. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician* 37(1): 36-48.

Efron, B., and R. Tibshirani 1993. *An Introduction to the Bootstrap*: Chapman and Hall.

Estrella, P., O. Hamon, and A. Popescu-Belis 2007. How Much Data is Needed for Reliable MT Evaluation? Using Bootstrapping to Study Human and Automatic Metrics In *MT Summit XI*, To appear. Copenhagen, Denmark.

Hamon, O., A. Popescu-Belis, K. Choukri, M. Dabbadie, A. Hartley, W. Mustafa El Hadi, M. Rajman, and I. Timimi 2006. CESTA: First Conclusions of the Technolangue MT Evaluation Campaign. In *LREC 2006 (5th International Conference on Language Resources and Evaluation)*, 179-184. Genova, Italy.

Koehn, P. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *EMNLP '04 (Conference on Empirical Methods in Natural Language Processing)*, 388-395. Barcelona, Spain.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *10th Machine Translation Summit - (MT SUMMIT)*, 79--86. Phuket, Thailand.

Kulesza, A., and S. Shieber 2004. A learning approach to improving sentence-level MT evaluation. In *10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, 75--84. Baltimore MD.

Kumar, S., and W. Byrne 2004. Minimum Bayes-risk decoding for statistical machine translation. In *HLT-NAACL*, 169-176.

Niessen, S., F. J. Och, G. Leusch, and H. Ney 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *LREC 2000 (2nd International Conference on Language Resources and Evaluation)*, 39-45. Athens, Greece.

Papineni, K. 2002. Machine Translation Evaluation: N-grams to the Rescue. In *LREC 2002 (Third International Conference on Language Resources and Evaluation)*. Las Palmas, Canary Islands, Spain.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. Yorktown Heights, NY: IBM Research Division, T.J.Watson Research Center.

Russo-Lassner, G., J. Lin, and P. Resnik 2005. A Paraphrase-Based Approach to Machine Translation Evaluation. University of Maryland, College Park

Tillmann, C., S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf 1997. Accelerated DP Based Search for Statistical Translation. In *Eurospeech 1997*, 2667--2670. Rhodes, Greece.

Turian, J. P., L. Shen, and I. D. Melamed 2003. Evaluation of Machine Translation and its Evaluation. In *Machine Translation Summit IX*, 386-393. New Orleans, Louisiana, USA.

White, J. S. 2001. Predicting Intelligibility from Fidelity in MT Evaluation. In *Workshop on MT Evaluation "Who did what to whom?" at Mt Summit VIII*. Santiago de Compostela, Spain.

Zhang, Y., S. Vogel, and A. Waibel 2004. Interpreting BLEU/NIST Scores: How Much Improvement do We Need to Have a Better System? In *LREC 2004 (4th International Conference on Language Resources and Evaluation)*, 2051-2054. Lisbon, Portugal.