EFFICIENT COMPUTATION OF CONFIDENCE INTERVALS FOR WORD ERROR RATES

Juan Miguel Vilar

jvilar@lsi.uji.es
Departamento de Lenguajes y Sistemas Informáticos
Universitat Jaume I
12071 Castellón (SPAIN)

ABSTRACT

Word Error Rate is a standard measure of quality for different tasks such as Speech Recognition, OCR or Machine Translation. As such, it is important to compute it together with confidence intervals. Previous works in the literature employ Monte Carlo methods in order to compute those intervals. We show how to compute them without simulations. We also adapt a method that compares two systems over the same test data so that it can be used without simulations.

Index Terms— error analysis, word error rate

1. INTRODUCTION

In many tasks the output of the system is a sentence (or a collection of them), examples are Speech Recognition, OCR or Machine Translation. The evaluation of those systems is traditionally carried out by comparing the output with a reference. The Word Error Rate (WER) is one measure of the difference between the output of the system and the reference. It is measured as the number of edition operations (insertions, deletions and substitutions) needed to transform the reference in the output divided by the length of the reference.

In order to properly compare different systems, it is necessary to provide not only the value of the WER but also a confidence interval for it. In [1] the authors propose a method for computing those intervals using *bootstrapping*, ie. repeatedly sampling the test sentences in order to find relevant statistics (Monte Carlo estimates). In [2] an approximation to the standard error of the WER is provided, but it is not formally derived and its validity is not proven.

Here we formally derive an estimate for the distribution of the WER and show how to use it for computing confidence intervals. We also show how to efficiently (ie. without simulations) compute the probability that a system is an improvement over another (the so called probability of improvement [1]).

Work partially supported by the Spanish Ministerio de Educación y Ciencia (TIN2006-12767 and Ingenio 2010 project MIPRCV, CSD2007-00018), the Generalitat Valenciana (GV06/302), and Bancaixa (P1 1B2006-31).

2. FORMULATION OF THE PROBLEM

Following [1], we will assume that the evaluation has been performed over s sentences. We will use n_i for the length of sentence i and e_i for the number of errors in it. We can represent the whole test as the sequence:

$$x = (n_1, e_1), \ldots, (n_s, e_s). \tag{1}$$

The total average WER is then:

$$w = \frac{\sum_i e_i}{\sum_i n_i}. \tag{2}$$

The bootstrap procedure consists in repeatedly sampling x with replacement in order to produce B different samples:

$$x_b^* = (n_1^{*b}, e_1^{*b}), \ldots, (n_s^{*b}, e_s^{*b}), \tag{3}$$

for $b = 1, \ldots, B$. For these, we obtain the corresponding WERs:

$$w_b^* = \frac{\sum_i e_i^{*b}}{\sum_i n_i^{*b}}. \tag{4}$$

These can be regarded as samples from a random variable[1] W^* from which to obtain the confidence intervals.

Our problem then is to find a confidence interval for the distribution of the random variable W^* defined as

$$W^* = \frac{\sum_i E_i}{\sum_i N_i}, \tag{5}$$

where the E_i and N_i are discrete random variables such that the different (E_i, N_i) are iid[2], but for any i the corresponding E_i and N_i need not be iid.

Note that the difficulty of the problem lies in this lack of independence. If E_i and N_i were independent, so would be their sums and the distribution of W^* could be easily computed. The key idea is to find a related expression that is a sum of independent variables.

[1] Following standard conventions, we will use uppercase letters for random variables and lowercase letters for concrete values of those variables.

[2] This is because they come from a sampling with replacement.

3. DISTRIBUTION OF THE WER

First, consider the distribution function of W^*:

$$P(W^* < x) = P\left(\frac{\sum_i E_i}{\sum_i N_i} < x\right). \quad (6)$$

Since the lengths of the sentences are all positive, we get:

$$P(W^* < x) = P\left(\sum_i (E_i - xN_i) < 0\right). \quad (7)$$

Define

$$Z_i^x = E_i - xN_i. \quad (8)$$

Remember that the pairs (E_i, N_i) are iid., therefore the Z_i^x are also iid. and the central limit theorem can be applied:

$$P(W^* < x) = P\left(\frac{\sum_i Z_i^x - s\,\mathrm{E}(Z^x)}{\sqrt{s}\,\mathrm{sd}(Z^x)} < \frac{-\sqrt{s}\,\mathrm{E}(Z^x)}{\mathrm{sd}(Z^x)}\right) \approx \Phi\left(\frac{-\sqrt{s}\,\mathrm{E}(Z^x)}{\mathrm{sd}(Z^x)}\right), \quad (9)$$

where E represents the expected value, sd represents the standard deviation, and Φ is the distribution function of a variable under a normal distribution with mean 0 and variance 1.

Finding the confidence interval now consists in fixing the values of Φ and solving for x. For instance, for a 90% confidence interval, we want the values that make Φ equal to 0.05 and 0.95, which are approximately -1.64 and 1.64, respectively. Suppose that we have found that our limit is l (in the previous example, $l = -1.64$ for the lower limit and $l = 1.64$ for the upper limit). Then we have:

$$\frac{-\sqrt{s}\,\mathrm{E}(Z^x)}{\mathrm{sd}(Z^x)} = l. \quad (10)$$

By the definition of Z^x, we have:

$$\mathrm{E}(Z^x) = \mathrm{E}(E) - x\,\mathrm{E}(N), \quad (11)$$

$$\mathrm{sd}(Z^x) = \sqrt{\mathrm{var}(E) + x^2\,\mathrm{var}(N) - 2x\,\mathrm{cov}(E,N)}, \quad (12)$$

where var is the variance and cov the covariance. Therefore, we rewrite (10) as:

$$\frac{-\sqrt{s}(\mathrm{E}(E) - x\,\mathrm{E}(N))}{\sqrt{\mathrm{var}(E) + x^2\,\mathrm{var}(N) - 2x\,\mathrm{cov}(E,N)}} = l. \quad (13)$$

Squaring and rearranging we arrive to:

$$\begin{aligned}&(l^2\,\mathrm{var}(N) - s\,\mathrm{E}^2(N))x^2 \\ &+ (2s\,\mathrm{E}(E)\,\mathrm{E}(N) - 2l^2\,\mathrm{cov}(E,N))x \\ &+ l^2\,\mathrm{var}(E) - s\,\mathrm{E}^2(E) = 0,\end{aligned} \quad (14)$$

which can be easily solved. Note that the squaring introduces a spurious solution, but, if the two values of l that are of interest have the same absolute value and opposite sign (as they usually do), then the two solutions of (14) for any of them are the endpoints of the confidence interval.

Interestingly, if we solve (13) for $l = 0$, we get the median, which is $\mathrm{E}(E)/\,\mathrm{E}(N)$, the empirical value of W. This can be seen as a good reason for using it as estimate of the value of the WER.

4. COMPUTATION OF CONFIDENCE INTERVALS

Using the results from the previous section, the procedure for computing the confidence interval is:

- From the test sentences, compute $\mathrm{E}(E)$, $\mathrm{E}(N)$, $\mathrm{E}(E^2)$, $\mathrm{E}(N^2)$ and $\mathrm{E}(EN)$.
- Compute $\mathrm{var}(E)$ and $\mathrm{var}(N)$ as $\mathrm{E}(E^2) - \mathrm{E}^2(E)$ and $\mathrm{E}(N^2) - \mathrm{E}^2(N)$, respectively.
- Compute $\mathrm{cov}(E,N)$ as $\mathrm{E}(EN) - \mathrm{E}(E)\,\mathrm{E}(N)$.
- Find the value of l for the desired confidence level.
- Solve Equation (14), obtaining x_1, x_2. The confidence interval is then (x_1, x_2).

Note that all these computations can be carried out in a single pass over the test data as the expected values in the first step are just averages.

5. SYSTEM COMPARISON

The confidence intervals found by the above method are appropriate for comparing systems over different test sets. When using a single test set for a direct comparison, other alternatives exist. A measure for direct comparison of two systems is proposed in [1]. This consists in evaluating the probability that the difference of the WER for the systems is less than zero. To this end, they define the difference in WER as

$$\Delta w := w^A - w^B = \frac{\sum_i (e_i^A - e_i^B)}{\sum_i n_i}, \quad (15)$$

where the two systems are A and B. Then, they propose to use bootstrapping to estimate the probability that $\Delta W^* < 0$.

We can proceed like in Section 3. First, we examine the distribution function of ΔW^*:

$$P(\Delta W^* < 0) = P\left(\frac{\sum_i (E_i^A - E_i^B)}{\sum_i N_i} < 0\right). \quad (16)$$

Now, define ΔE_i as $E_i^A - E_i^B$. Since the N_i are all positive, we can leave out the denominator. Therefore

$$P(\Delta W^* < 0) = P\left(\sum_i \Delta E_i < 0\right) = P\left(\frac{\sum_i \Delta E_i - s\,\mathrm{E}(\Delta E)}{\sqrt{s}\,\mathrm{sd}(\Delta E)} < \frac{-\sqrt{s}\,\mathrm{E}(\Delta E)}{\mathrm{sd}(\Delta E)}\right) \approx \Phi\left(\frac{-\sqrt{s}\,\mathrm{E}(\Delta E)}{\mathrm{sd}(\Delta E)}\right). \quad (17)$$

This immediately suggests an algorithm:

- For each sentence compute the difference in errors between system A and B. Use them to find $\mathrm{E}(\Delta E)$ and $\mathrm{sd}(\Delta E) = \sqrt{\mathrm{E}(\Delta E^2) - \mathrm{E}^2(\Delta E)}$.
- Use $\Phi(-\sqrt{s}\ \mathrm{E}(\Delta E)/\,\mathrm{sd}(\Delta E))$ as the estimate for the probability of improvement.

6. EXPERIMENTS

We tested our proposal using the submissions for the shared task of the Second Workshop on Statistical Machine Translation [3]. Note that although the WER is not the best measure for translation systems, it is often used for evaluating them. The corresponding files are available on-line in the page `http://www.statmt.org/wmt07/`.

We computed the WER and the confidence intervals for two subtasks: Spanish to English and German to English over the Europarl test, which consists in 2000 sentences (55 383 words). That makes the confidence intervals quite narrow.

The results can be seen in tables 1 and 2. The first column identifies the system, the second is the WER, the third is the interval obtained by 1000 repetions of bootstrapping, the fourth column corresponds to our method, finally, the fifth column has been obtained by the formula in [2]. In all cases, the intervals correspond to a 95% confidence level.

Taking the bootstrap columns as references, it is clear that the proposed method gets very adjusted intervals. So does the method from [2], although the intervals are slightly wider. This is probably due to some assumption that is not completely fulfilled. This effect is magnified in artificial situations. For instance, if the samples are 500 sentences of length one with one error and other 500 sentences of length ten without errors, the WER is 0.09, the bootstrap interval is (0.06, 0.13), identical to the result of our method and the interval from [2] is (0.03, 0.15).

Looking at the tables, we see a large degree of overlap, so we can compute the probability of improvement. That is reflected in tables 3 and 4 for the German to English task. In each entry, there are two figures, the first is computed by bootstrapping and the second using the method from section 5. Again, the numbers are very close. We see that in the Spanish to English task there are reasonable doubts about the relative order of systems cmu-syntax and uedin on one side and upv and nrc on the other. In the case of German to English, the same happens with the relative position of the systems liu, saar, cmu-uka and nrc.

7. OTHER MEASURES

Note that the derivation of the intervals depends only on the fact that the measure can be expressed as the quotient of two summations, each one over the same set of indexes. As such, the derivation can be used for many different measures like PER (position independent word error rate) [4] or TER (translation edit rate) [5]. It is only a matter of changing the meaning of the E_i and N_i in the formulae above.

It would be interesting to find a similar algorithm for other measures like BLEU [6], but it seems more difficult. The main problem lies in the form of the score: a weighted geometric average that does not lend itself to manipulations similar to those we have done here.

8. CONCLUSIONS

A new method for computing confidence intervals for WER has been presented. The only assumption for this method is that the boostrapping procedure is valid, therefore it can be applied in a wide range of situations. Its main advantage is the ease of computation, a single pass over the data suffices, without resorting to Monte Carlo simulations. Using the same ideas, a method for computing the probability of improvement is also simplified.

9. REFERENCES

[1] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proceedings of the ICASSP'04*, Montreal (Canada), 2004, vol. 1, pp. 409–412.

[2] David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs, "Human evaluation of machine translation through binary system comparisons," in *Second Workshop on Statistical Machine Translation*, Prague (Czech Republic), June 2007, pp. 96–103.

[3] Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, Eds., *Proceedings of the Second Workshop on Statistical Machine Translation*, ACL, Prague (Czech Republic), June 2007.

[4] C. Tillmann, S. Vogel, H. Ney, H. Sawaf, and A. Zubiaga, "Accelerated DP-based search for statistical translation," in *Proceedings of the EuroSpeech'97*, Rhodes (Greece), Sept. 1997, ESCA, vol. 5, pp. 2667–2670.

[5] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambrigde, Massachusets, (USA), Aug. 2006, AMTA, pp. 223–231.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, Pennsylvania (USA), 2002, pp. 311–318, ACL.

System	WER	Bootstrap	Proposed	From [2]
cmu-uka	0.61	(0.60, 0.63)	(0.60, 0.63)	(0.59, 0.63)
uedin	0.62	(0.61, 0.63)	(0.61, 0.63)	(0.60, 0.64)
cmu-syntax	0.62	(0.61, 0.63)	(0.61, 0.63)	(0.60, 0.64)
upc	0.63	(0.61, 0.64)	(0.62, 0.64)	(0.61, 0.65)
nrc	0.63	(0.62, 0.64)	(0.62, 0.64)	(0.61, 0.65)
upv	0.63	(0.62, 0.65)	(0.62, 0.65)	(0.61, 0.65)
systran	0.67	(0.66, 0.68)	(0.66, 0.68)	(0.65, 0.69)
saar	0.71	(0.70, 0.72)	(0.70, 0.72)	(0.69, 0.73)

Table 1. Estimated confidence intervals for the Spanish to English task. First column is the empirical value of WER, the following columns correspond to confidence intervals computed by bootstrapping, by our proposal and by a formula in [2]. Values in gray are equal to the bootstrap reference.

System	WER	Bootstrap	Proposed	From [2]
uedin	0.67	(0.66, 0.68)	(0.66, 0.68)	(0.66, 0.69)
upc	0.70	(0.69, 0.71)	(0.69, 0.71)	(0.68, 0.71)
systran	0.70	(0.69, 0.71)	(0.69, 0.71)	(0.69, 0.72)
liu	0.72	(0.71, 0.73)	(0.71, 0.73)	(0.70, 0.73)
saar	0.72	(0.71, 0.72)	(0.71, 0.72)	(0.71, 0.73)
cmu-uka	0.72	(0.71, 0.73)	(0.71, 0.73)	(0.70, 0.73)
nrc	0.72	(0.71, 0.73)	(0.71, 0.73)	(0.71, 0.73)

Table 2. Estimated confidence intervals for the German to English task. First column is the empirical value of WER, the following columns correspond to confidence intervals computed by bootstrapping, by our proposal and by a formula in [2]. Values in gray are equal to the bootstrap reference.

	cmu-uka	uedin	cmu-syntax	upc	nrc	upv	systran	saar
cmu-uka	—	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
uedin	0.00 0.00	—	0.71 0.72	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
cmu-syntax	0.00 0.00	0.28 0.28	—	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
upc	0.00 0.00	0.00 0.00	0.00 0.00	—	0.99 0.99	1.00 1.00	1.00 1.00	1.00 1.00
nrc	0.00 0.00	0.00 0.00	0.00 0.00	0.01 0.01	—	0.71 0.71	1.00 1.00	1.00 1.00
upv	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.28 0.29	—	1.00 1.00	1.00 1.00
systran	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	—	1.00 1.00
saar	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	—

Table 3. Probability of improvement for the Spanish to English task. The values of entry (i, j) are the probabilities that system i outperforms system j computed by bootstrapping and by our proposal. Gray values are either 0.00 or 1.00.

	uedin	upc	systran	liu	saar	cmu-uka	nrc
uedin	—	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
upc	0.00 0.00	—	0.97 0.97	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
systran	0.00 0.00	0.03 0.03	—	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
liu	0.00 0.00	0.00 0.00	0.00 0.00	—	0.51 0.50	0.84 0.84	0.93 0.93
saar	0.00 0.00	0.00 0.00	0.00 0.00	0.50 0.50	—	0.83 0.84	0.90 0.92
cmu-uka	0.00 0.00	0.00 0.00	0.00 0.00	0.17 0.16	0.17 0.16	—	0.67 0.67
nrc	0.00 0.00	0.00 0.00	0.00 0.00	0.06 0.07	0.08 0.08	0.33 0.33	—

Table 4. Probability of improvement for the German to English task. The values of entry (i, j) are the probabilities that system i outperforms system j computed by bootstrapping and by our proposal. Gray values are either 0.00 or 1.00.