Statistical Significance Tests in Machine Translation

Tsuyoshi Okita, Andy Way
CNGL/School of Computing, Dublin City University

Overview

In the context of Machine Translation, there are two popular statistical significance tests : a method based on bootstrap method [Koehn,2004; Zhang and Vogel, 2004] and that based on approximate randomization [Riezler and Maxwell III, 2005]. The latter is more conservative since it increases the likelihood of type-I error than the former.

Bootstrap Test for Statistical Significance Testing

Compute actual statistic of score differences |Sx − Sy| on test data
Calculate sample mean Tb = 1/B \Sum_{b=0}^B |Sxb − Syb| over bootstrap samples b=0,,B
For bootstrap samples b=0,...,B
  Sample with replacement from variable tuples for systems X and Y for test sentences
  Compute pseudo−statistic | Sxb − Syb| on bootstrap data
  If |Sxb − Syb| − Tb (+T) \geq |Sx − Sy| c++
p = (c+1) / (B+1)
Reject null hypothesis if p is less than or equal to specified rejection level

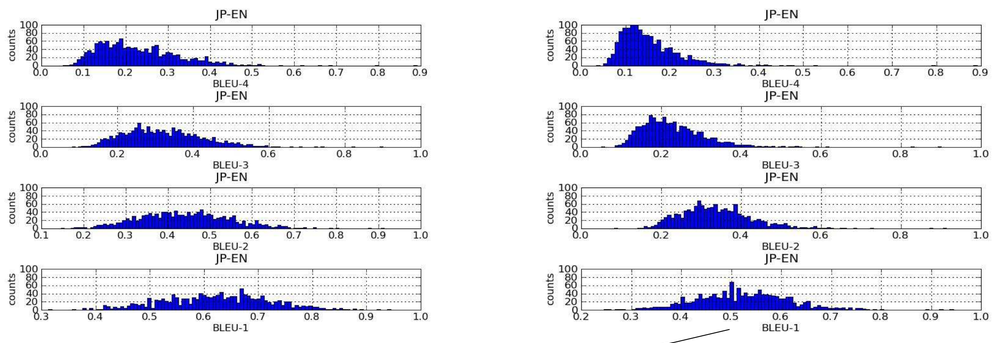
Approximate Randomization Test for Statistical Significance Testing

Compute actual statistic of score differences |Sx − Sy| on test data
For random shuffles r = 0,...,R
  For sentences in test set
    Shuffle variable tuples between system X and Y with probability 0.5
  Compute pseudo−statistic |Sxr−Syr| on shuffled data
  If |Sxr−Syr| \geq |Sx−Sy| c++
p = (c+1) / (R+1)
Reject null hypothesis if p is less than or equal to specified rejection level

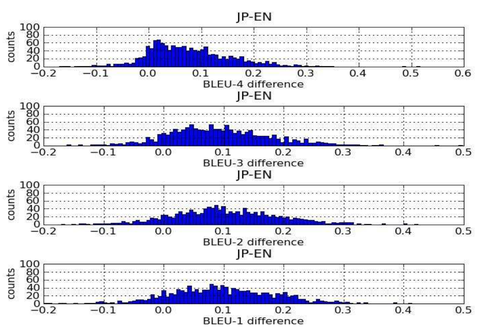
Characteristics in Machine Translation Context

(1) For a given test set, since an MT system does not produce translation outputs in various ways the overall score, which is often measured by BLEU [Papineni et al., 2002], is single and fixed. An idea in the above two methods is to randomly select a paired test set in a sentence level to enable the permutation tests, which seems supported by the stratification of the output [Yeh, 2000; Noreen 1989]. Often test statistic is examined 1000 times. However, is this no problem?

Performance on System A (BLEU−4,3,2,1 per sentence) Performance on System B (BLEU−4,3,2,1 per sentence)



Performance Difference per sentence

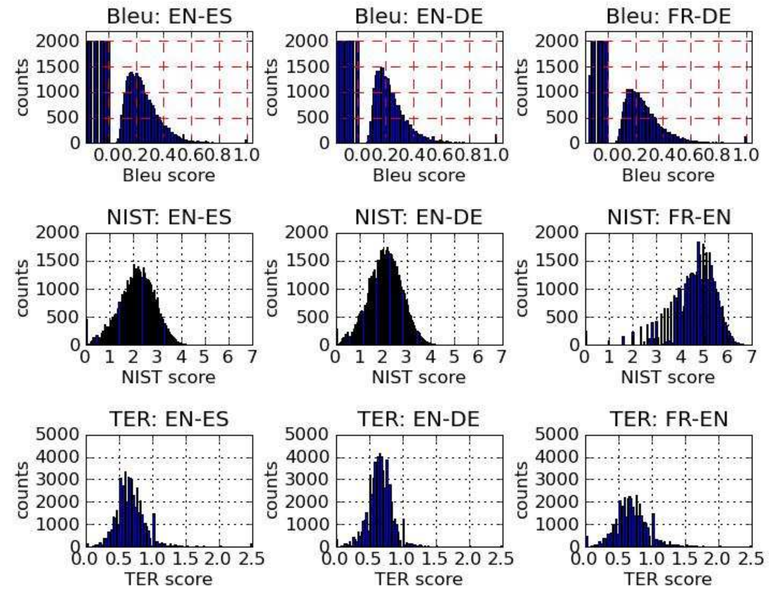


There are a lot of sentences whose BLEU score are both zero.

but this does not matter .	peu importe !
we may find ourselves there once again .	va−t−il en etre de meme cette fois−ci ?
all for the good .	et c' est tant mieux !
but if the ceo is not accountable , who is ?	mais s' il n' est pas responsable , qui alors ?

Evaluation Measure

(2) The test statistic consists of the difference between the evaluated translation outputs in Machine Translation, whose evaluation measure is given from the beginning: either BLEU [Papineni et al., 2002], NIST [Doddington, 2002], METEOR [Bernerjee and Lavie, 200], TER [Snover et al., 2006], or others. Riezler and Maxwell III [2005] say that NIST is more appropriate than BLEU with approximate randomization. Does statistical significance test depend on the evaluation measure?



Hypothesis Test for Dependent Data (block wise error)

(3) When algorithm A and B are compared, it is often the case where these two systems share most of the underlying systems, i.e. there are a lot of dependencies. Church and Mercer [1993] give examples of dependence between test set instances in natural language. Although expected value of the instance results will stay the same, but the chances of getting an unusual result may change. Hence, the chances of getting an unusual result under some null hypothesis requires to incorporate these dependencies. Then, do we need to quantify these dependencies?

Stationary Block Bootstrap [Politis and Romano, 1994]

This method uses blocks of random lengths (not blocks of a fixed length).
In the procedure below, we choose the distribution Fb() to be a geometric distribution with mean equal to the real number b.

1. Start by wrapping the data {X1,...,Xn} around a circle, i.e., define the new series Yt := Xt_mod(N), for t \in N, where mod(N) denotes "module N".
2. Let i0, i1,... be drawn i.i.d. with uniform distribution on the set {1,2,...,N}; these are the starting points of the new blocks.
3. Let b0,b1,..., be drawn i.i.d. from some distribution Fb() that depends on a parameter b (that may depend on N); these are the block sizes.
4. Construct a bootstrap pseudo−seris Y1*, Y2*,..., as follows. For m=0,1,..., let Y_{mb_m + j}^{*} := Y_{i_m+j−1} for j=1,2,...,b_m. This procedure defines a probability measure (conditional on the data X_1,...,X_N) that will be denoted P^{*}; expectation and variance with respect to P^{*} are denoted E^{*} and Var^{*} respectively.
5. Finally, we focus on the first N points of the bootstrap series and construct the bootstrap sample mean \bar{Y_N^{*}} = N^{−1} \sum_{i=1}^N Y_i^{*}. This corresponding estimate of the asymptotic variance of the sample mean is then given by Var^{*} (\sqrt(N \bar{Y}_N^{*}}).

Multiple Hypothesis Tests (family wise error)

(4)It is often the case even though algorithm A is proven to be statistical significant with algorithm B for one set of corpus, it does not often work for another set of corpus whose language pairs are different or whose size are different.

[Type I error rates] At some designated level α and a predefined value $\mathbf{c}_\alpha$, we reject a single hypothesis $\mathbf{H_1}$ when $|\mathbf{T_1}| \geq \mathbf{c}_\alpha$ where $\mathbf{c}_\alpha$ is either the per-comparison error rate (PCER, E(V)/m), the per-family error rate (PFER, E(V)), the family-wise error rate (FWER, $\mathbf{Pr(V \geq 1)}$), the false discovery rate (FDR, E(Q) where Q=V/R if R¿0 and 0 if R=0) [Shaffer, 1995].

Appendix

[BLEU definition] Given the precision $\mathbf{p_n}$ of n-grams of size up to N, the length of the test set in words (c) and the length of the reference translation in words (r),

$$\text{BLEU} = \text{BP} \cdot \mathbf{exp}(\sum_{\mathbf{n=1}}^{\mathbf{4}} \mathbf{log}\, \mathbf{p_n}), \qquad \text{BP} = \mathbf{min}(1, \mathbf{e}^{1-\mathbf{r}/\mathbf{c}})$$

BLEU example system output: Israeli officials responsibility of airport safety
reference : Israeli officials are responsible for airport security
1−gram precision 3/6, 2−gram precision 1/5, 3−gram precision 0/4, 4−gram precision 0/3
Israeli officials airport Israeli officials none none
brevity penalty = 6/7
BLEU−1=3/6 * 6/7=0.42, BLEU−2=3/6*1/5*6/7=0.085

This research is funded by Science Foundation Ireland (SFI) as part of the National Development Plan (NDP) 2007-2013