Statistical variability in NLP

• How to trust experiment results, given many sources of variability?
  • How was the text data sampled?
  • How were the annotations sampled?
    • How variably do the human annotators behave?
  • How variable are the computational algorithms?
Computational variability

- Randomness in algorithm?
- Arbitrariness in hyperparameters?
- Options to control
  - Maximize settings on development data
  - Average over randomness
Randomness in learning algo.

Figure 1: Histogram of test set BLEU scores for the BTEC phrase-based system (left) and BTEC hierarchical system (right). While the difference between the systems is 1.5 BLEU in expectation, there is a non-trivial region of overlap indicating that some random outcomes will result in little to no difference being observed.

[Dyer et al. 2011]
optimizer samples, we have a statistic that jointly quantifies the impact of test set effects and optimizer instability on a test set. We call this statistic \( s_{\text{sel}} \). Different values of this statistic can suggest methodological improvements. For example, a large \( s_{\text{sel}} \) indicates that more replications will be necessary to draw reliable inferences from experiments on this test set, so a larger test set may be helpful.

To compute \( s_{\text{sel}} \), assume we have \( n \) independent optimization runs which produced weight vectors that were used to translate a test set \( n \) times. The test set has \( R_1, R_2, \ldots, R_j \). Let \( X_1, X_2, \ldots, X_n \) where each \( X_i = X_{i1}, X_{i2}, \ldots, X_{ik} \) is the list of translated segments from the \( i \)th optimization run list of the \( j \) translated segments of the test set. For each hypothesis output \( X_i \), we construct \( k \) bootstrap replicates by drawing \( m_{ij} \) segments uniformly, with replacement, from \( X_i \), together with its corresponding reference. This produces \( k \) virtual test sets for each optimization run \( i \). We designate the score of the \( j \)th virtual test set of the \( i \)th optimization run with \( m_{ij} \).

\[
\begin{align*}
\text{If } m_{ij} = 1 \\
\text{then } s_i &= \frac{1}{n} \sum_{i=1}^{n} \left( m_{ij} - m_i \right)^2
\end{align*}
\]

4.2 Comparing Two Systems

In the previous section, we gave statistics about the distribution of evaluation metrics across a large number of experimental samples (Table 1). Because of the large number of trials we carried out, we can be extremely confident in concluding that for both pairs of systems, the experimental manipulation accounts for the observed metric improvements, and furthermore, that we have a good estimate of the magnitude of that improvement. However, it is not generally feasible to perform as many replications as we did, so here we turn to the question of how to compare two systems, accounting for optimizer noise, but without running 300 replications.

We begin with a visual illustration how optimizer instability affects test set scores when comparing two systems. Figure 1 plots the histogram of the 300 optimizer samples each from the two BTEC Chinese-English systems. The phrase-based system's distribution is centered at the sample mean 48.4, and the hierarchical system is centered at 49.9, a difference of 1.5 BLEU, corresponding to the widely replicated result that hierarchical phrase-based systems outperform conventional phrase-based systems in Chinese-English translation. Crucially, although the distributions are distinct, there is a non-trivial region of overlap, and experimental samples from the overlapping region could suggest the opposite conclusion!

To further underscore the risks posed by this overlap, Figure 2 plots the relative frequencies with which different BLEU score deltas will occur, as a function of the number of optimizer samples used. The expected difference is 0.2 BLEU. While there is a reasonably high chance of observing a non-trivial improvement (or even a decline) for 1 sample, the distribution quickly peaks around the expected value given just a few more samples.

\[\text{[Dyer et al. 2011]}\]
Human variability
Human variability

- Linguistic/content annotations are naturally variable and subjective
  - Is this a positive or negative view of a product?
  - Is this sentence grammatical?
Human variability

- Linguistic/content annotations are naturally variable and subjective
  - Is this a positive or negative view of a product?
  - Is this sentence grammatical?
- Does the task setup usefully get at the phenomenon?
  - “Is this sentence grammatical” vs.
    “Which of these two sentences is more grammatical?”
Human variability

- Linguistic/content annotations are naturally variable and subjective
  - Is this a positive or negative view of a product?
  - Is this sentence grammatical?
- Does the task setup usefully get at the phenomenon?
  - “Is this sentence grammatical” vs.
    “Which of these two sentences is more grammatical?”
- Typical measurement is via agreement rate between two human annotators
  - Training? Tiredness? Attention?
  - Self-agreement? (psychometrics: test-retest correlation)
Human variability

• Linguistic/content annotations are naturally variable and subjective
  • Is this a positive or negative view of a product?
  • Is this sentence grammatical?
• Does the task setup usefully get at the phenomenon?
  • “Is this sentence grammatical” vs. “Which of these two sentences is more grammatical?”
• Typical measurement is via agreement rate between two human annotators
  • Training? Tiredness? Attention?
  • Self-agreement? (psychometrics: test-retest correlation)
• Discrete data
  • Agreement rate (accuracy)
  • Cohen’s kappa: control for agreement due to chance (randomly guessing the base rates)
Human variability

- Linguistic/content annotations are naturally variable and subjective
  - Is this a positive or negative view of a product?
  - Is this sentence grammatical?
- Does the task setup usefully get at the phenomenon?
  - “Is this sentence grammatical” vs.
    “Which of these two sentences is more grammatical?”
- Typical measurement is via agreement rate between two human annotators
  - Training? Tiredness? Attention?
  - Self-agreement? (psychometrics: test-retest correlation)
- Discrete data
  - Agreement rate (accuracy)
  - Cohen’s kappa: control for agreement due to chance (randomly guessing the base rates)
- Real-valued data: correlation, rank correlation, MAE, etc.
Text data variability

• Do results generalize to ....
  • new domains?
  • new authors?
  • new documents?
  • new sentences?

• (Typically things get worse if anything changes)
• Also of interest: even if only care about text similar to our current one, did we “get lucky” in our selection of sentences/documents/etc?
Text data variability

• Mathematically, the easiest case to analyze: What if we resampled the tokens/sentences/documents from a similar population as our current data sample?

• Assume units are sampled i.i.d.; then apply your favorite statistical significance/confidence interval testing technique
  • T-tests, binomial tests
  • Bootstrapping
  • Paired tests
Significance tests and CIs

• Given how small the data sample is, how much information do we really have about the true parameter $\theta$
• (e.g. accuracy if we could access the population)

• Null hypothesis testing / p-values:
  chance of seeing as extreme/interesting result, given an uninteresting null hypothesis

• Confidence intervals with A% confidence
  • 1. Probability the true value is in this set
    • Bayesian interpretation; useful intuition, typically not used for experimental results, but sometimes similar
  • 2. Following this CI inference algorithm, A% of all experiments will have the true value contained within them
  • Frequentist interpretation

• CI view of null hypothesis testing:
  e.g. Does the CI not include zero?
Statistical tests

• Closed-form tests
  • t-tests, exact binomial test, chi-square tests....
• Bootstrapping: very flexible!