Statistical Testing in NLP (I)

CS 690N, Spring 2018

Advanced Natural Language Processing

http://people.cs.umass.edu/~brenocon/anlp2018/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

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.

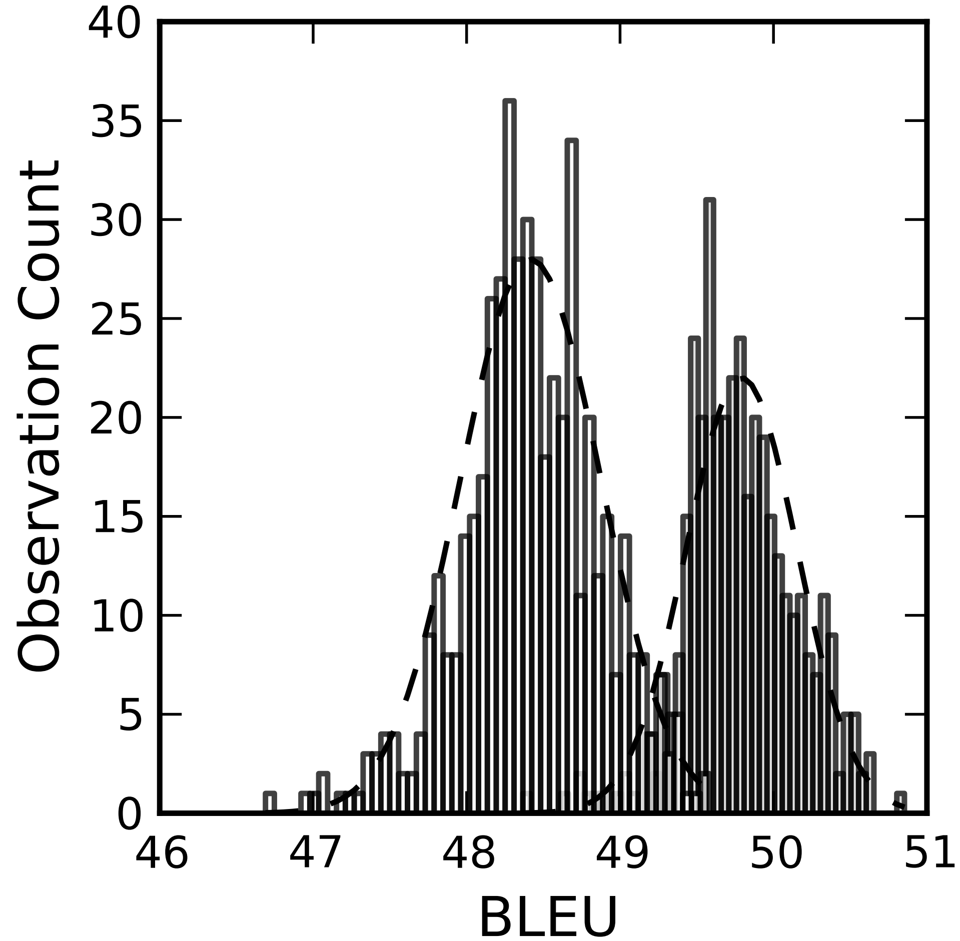

Figure 1: Histogram of test set BLEU scores for the BTEC phrase-based system (left) and BTEC hierarchical system (right). While the difference between the systems is 1.5 BLEU in expectation, there is a non-trivial region of overlap indicating that some random outcomes will result in little to no difference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

[Dyer et al. 2011]

Randomness in learning algo.

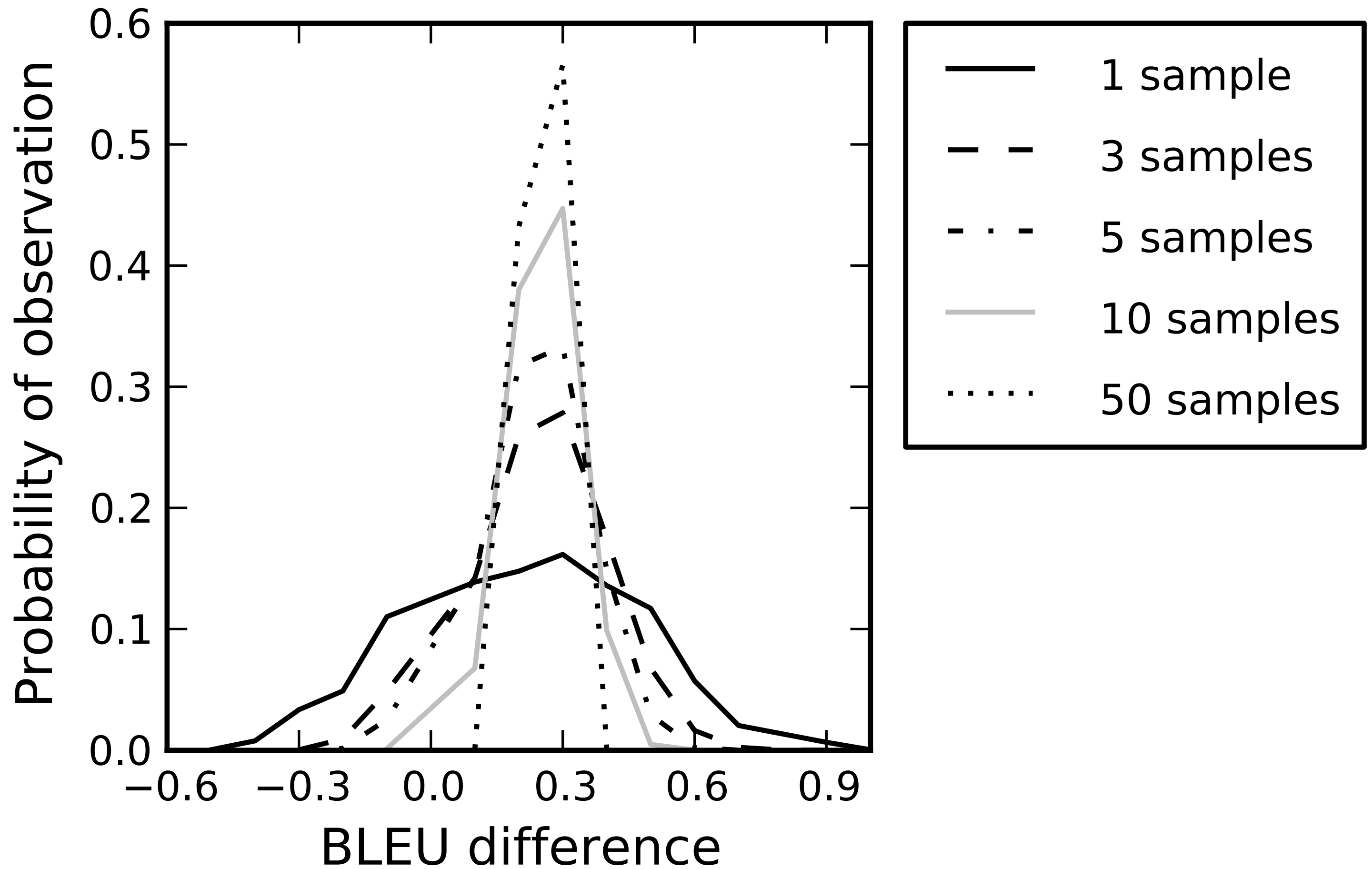

Figure 2: Relative frequencies of obtaining differences in BLEU scores on the WMT system as a function of the number of optimizer samples. The expected difference is 0.2 BLEU. While there is a reasonably high chance of observing a non-trivial improvement (or even a decline) for 1 sample, the distribution quickly peaks around the expected value given just a few more samples.

[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 θ
  - (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!