NLP Evaluation:
Bootstrapping & sig tests

CS 585, Fall 2015
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2015/

Brendan O’Connor
College of Information and Computer Sciences
University of Massachusetts Amherst
Questions

- What metrics to use?
- How to deal with complex outputs like translations?
- Are the human judgments ...
  - ... measuring something real?
  - ... reliable?
- Is the sample of texts sufficiently representative?
- How reliable or certain are the results?
Are my results meaningful?
Are my results meaningful?

- System 1 = 87% accuracy.
  System 2 = 89% accuracy.
Are my results meaningful?

- System 1 = 87% accuracy.
  System 2 = 89% accuracy.
- Does this difference mean anything? Key questions:
Are my results meaningful?

• System 1 = 87% accuracy.
  System 2 = 89% accuracy.

• Does this difference mean anything? Key questions:

• Do you trust the human judgments?
  • analyze agreement rates
Are my results meaningful?

- System1 = 87% accuracy.
  System2 = 89% accuracy.
- Does this difference mean anything? Key questions:
  - Do you trust the human judgments?
    - analyze agreement rates
  - Is the data from the right distribution?
    Correct domain/genre?
    - judgment call...?
Are my results meaningful?

- System1=87% accuracy.
  System2=89% accuracy.
- Does this difference mean anything? Key questions:
  - Do you trust the human judgments?
    - analyze agreement rates
  - Is the data from the right distribution?
    Correct domain/genre?
    - judgment call...?
  - Are there enough examples that we can trust it?
    - Statistical question! [Today]
Statistical “Significance”

• Assume data was drawn from a greater population.
• If we were to take a new sample, how much would data differ?
  • Or: how much would a statistic of that data differ?
  • “Confidence interval”
    (better name: Uncertainty Interval)

• How to test stat sig?
  • 1. Bootstrap simulation: handles anything (**)
  • 2. Off-the-shelf tests: for specific situations
  • 3. Quick rule-of-thumb (**)

Sunday, November 22, 15
Bootstrap test

- [blackboard]

- Inputs
  - Original data size N
  - Test statistic: \texttt{stat(data)}. e.g.
    - accuracy (numeric)
    - system1 better than system2? (boolean)

- Algorithm
  - For each of 10,000 replications:
    - Draw \texttt{samp}: a sample with replacement from the original data, again size N. (Many of the original examples will not be in sample)
    - Calculate \texttt{stat(samp)}
  - Save all 10,000 \texttt{stat(samp)} values. Then analyze
    - Numeric: Histogram. Mean, standard deviation, CI
    - Boolean: Proportion that are true?
Bootstrap test

• Two types (many others...)

• 1. Binary null hypothesis (7.3 JM 3ed)
  • Boolean statistic: is null hypo true?
  • p-value: Proportion of replications where null hypo is true
    (pvalue<.05 means a non-null hypothesis is ...
    “significant” ... worth considering)

• 2. Confidence interval (this lecture)
  • Numeric statistic: e.g. accuracy rate
  • The “normal approx” bootstrap CI:
    95% CI = [mean +/- 2*stdev]
Paired tests

- Single dataset. Compare system 1 vs system 2

- Good approach (“paired”): bootstrap sample items, compare system performances

- Bad approach (“unpaired”):
  - 1. bootstrap sample items. calc system1’s acc CI
  - 2. bootstrap sample items. calc system2’s acc CI
  - 3. do the CIs overlap?
  - Why bad?
Power Analysis

• How much data do we have to collect?
• *Power Analysis*: given how big an effect you want to measure, that implies how big N should be

• How to implement
  • Make fake dataset size N, run the bootstrap. Look at whether differences can be detected
  • [IPYNB DEMO]
  • Off-the-shelf formulas, e.g. R `power.t.test()`, `power.prop.test()`, http://www.statmethods.net/stats/power.html
  • Rules of thumb
Rules of thumb: CIs

- **Binomial CI (Agresti-Coull version)**
  K occurrences in N examples.
  Let $k' = K + 2$, $n' = N + 4$, $p' = k'/n'$
  95% CI = $[p' \pm 2\sqrt{p'(1-p')/n'}]$
  ... or more conservatively ...
  95% CI = $[p' \pm 1/\sqrt{n'}]$

- **Rule of Three**
  K=0 occurrences in N examples.
  Prob of occurrence?
  95% CI = $[0 \ldots 3/N]$
Rules of thumb: power analysis

http://www.nrcse.washington.edu/research/struts/chapter2.pdf
Rules of thumb: power analysis

• Rule of three:
  \[ K=0 \implies 3/N \quad 95\% \text{ upper bound} \]

To be sure prob \( \leq p \), how many examples?

http://www.nrcse.washington.edu/research/struts/chapter2.pdf
Rules of thumb: power analysis

• **Rule of three:**
  \[ K=0 \Rightarrow \frac{3}{N} \text{ 95\% upper bound} \]

  To be sure prob \( \leq p \), how many examples?

• \( \frac{3}{p} \)
Are my results meaningful?
Are my results meaningful?

• Statistical significance is neither sufficient nor necessary for a meaningful result! Remember there are three different factors:
Are my results meaningful?

• Statistical significance is neither sufficient nor necessary for a meaningful result! Remember there are three different factors:

• Do you trust the human judgments?
  • analyze agreement rates
Are my results meaningful?

• Statistical significance is neither sufficient nor necessary for a meaningful result! Remember there are three different factors:

• Do you trust the human judgments?
  • analyze agreement rates

• Is the data from the right distribution?
  Correct domain/genre?
  • judgment call...?
Are my results meaningful?

- Statistical significance is neither sufficient nor necessary for a meaningful result! Remember there are three different factors:

- Do you trust the human judgments?
  - analyze agreement rates

- Is the data from the right distribution?
  Correct domain/genre?
  - judgment call...?

- Are there enough examples that we can trust it?
  - Statistical question! [Today]