Foundations of Natural Language Processing
Lecture 10
Methods in Annotation and Evaluation

Henry S. Thompson

(Based on slides from Sharon Goldwater, with slides from Alex Lascarides and Nathan Schneider)

14 February 2019

School of informatics

David Hume

Today we will look at. . .

- Annotation
  - Why "gold" $\neq$ perfect
  - Quality Control
- Evaluation
  - Experimental setup
  - Significance testing
  - Error analysis
  - Evaluating without Gold Standards:
    How do we evaluate when there is more than one right answer?

Factors in Annotation

Suppose you are tasked with building an annotated corpus. (E.g., with part-of-speech tags.) In order to estimate **cost** in time and money, you need to decide on:

- Source data (genre? size? licensing?)
- Annotation scheme (complexity? guidelines?)
- Annotators (expertise? training?)
- Annotation software (graphical interface?)
- Quality control procedures (multiple annotation, adjudication?)

Annotation Scheme

- Assuming a competent annotator, some kinds of annotation are straightforward for most inputs.

Annotation Scheme

- Assuming a competent annotator, some kinds of annotation are straightforward for most inputs.

- Others are not.
  - Text may be ambiguous
  - There may be gray area between categories in the annotation scheme

You play annotator

Noun or adverb?

- **Yesterday** was my birthday .
- **Yesterday** I ate a cake .
- He was fired **yesterday** for leaking the information .
- I read it in **yesterday** 's news .
- I had not heard of it until **yesterday** .

You play annotator

Verb, noun, or adjective?

- We had been **walking** quite briskly
- **Walking** was the remedy, they decided
- In due time Sandburg was a **walking** thesaurus of American folk music.
- we all lived within **walking** distance of the studio
- a woman came along carrying a folded umbrella as a **walking** stick
- The **Walking** Dead premiered in the U.S. on October 31, 2010, on the cable television channel AMC

Annotation: Not as easy as you might think

Pretty much any annotation scheme for language will have some difficult cases where there is gray area, and multiple decisions are plausible.

- Because human language needs to be **flexible**, it cuts corners and is reshaped over time.

- Not just syntax: wait till we get to semantics!

Annotation Guidelines

However, we want a dataset's annotations to be as clean as possible so we can use them reliably in systems.

Documenting conventions in an annotation manual/standard/guidelines document is important to help annotators produce **consistent** data, and to help end users interpret the annotations correctly.

Annotation Guidelines

- Penn Treebank: 36 POS tags (excluding punctuation).
- Tagging guidelines (3rd Revision): 34 pages
  - "The temporal expressions *yesterday*, *today* and *tomorrow* should be tagged as nouns (NN) rather than as adverbs (RB). Note that you can (marginally) pluralize them and that they allow a possessive form, both of which true adverbs do not." (p. 19)
  - An entire page on nouns vs. verbs.
  - 3 pages on adjectives vs. verbs.
- Penn Treebank bracketing (tree) guidelines: $>$300 pages!

Annotation Quality

But even with extensive guidelines, human annotations won't be perfect:

- Simple error (hitting the wrong button)
- Not reading the full context
- Not noticing an erroneous pre-annotation
- Forgetting a detail from the guidelines
- Cases not anticipated by or not fully specified in guidelines (room for interpretation)

"Gold" data will have some tarnish. How can we measure its quality?

Inter-annotator agreement (IAA)

- An important way to estimate the reliability of annotations is to have multiple people independently annotate a common sample, and measure **inter-annotator/coder/rater agreement**.
- **Raw agreement rate**: proportion of labels in agreement
- If the annotation task is perfectly well-defined and the annotators are well-trained and do not make mistakes, then (in theory) they would agree 100%.
- If agreement is well below what is desired (will differ depending on the kind of annotation), examine the sources of disagreement and consider additional training or refining guidelines.
- The agreement rate can be thought of as an upper bound (**human ceiling**) on accuracy of a system evaluated on that dataset.

IAA: Beyond raw agreement rate

- Raw agreement rate counts all annotation decisions equally.

- Some measures take knowledge about the annotation scheme into account (e.g., counting singular vs. plural noun as a minor disagreement compared to noun vs. preposition).

- What if some decisions (e.g., POS tags) are far more frequent than others?
  - If 2 annotators both tagged *hell* as a noun, what is the chance that they agreed **by accident**? What if they agree that it is an interjection (rare tag)—is that equally likely to be an accident?
  - **Chance-corrected** measures such as Cohen's kappa (κ) adjust the agreement score based on label probabilities.
  - . . . but they make modeling assumptions about how "accidental" agreement would arise; important that these match the reality of the annotation process!
  - More below on hypothesis testing/statistical significance.

Crowdsourcing

- Quality control is even more important when eliciting annotations from "the crowd".

- E.g., **Amazon Mechanical Turk** facilitates paying anonymous web users small amounts of money for small amounts of work ("Human Intelligence Tasks").

- Need to take measures to ensure annotators are qualified and taking the task seriously.
  - Redundancy to combat noise: Elicit 5+ annotations per data point.
  - Embed data points with known answers, reject annotators who get them wrong.

The Nature of Evaluation

- Scientific method rests on making and testing hypotheses.

- Evaluation is just another name for testing.

- Evaluation not just for public review:
  - It's how you manage internal development
  - And even how systems improve themselves (see ML courses).

What Hypotheses?

About existing linguistic objects:

- Is this text by Shakespeare or Marlowe?

About output of a language system:

- How well does this language model predict the data?
- How accurate is this segmenter/tagger/parser?
  - Is this segmenter/tagger/parser better than that one?

About human beings:

- How reliable is this person's annotation?
- To what extent do these two annotators agree? (IAA)

Gold Standard Evaluation

- In many cases we have a record of 'the truth':
  - The best human judgement as to what the correct segmentation/tag/parse/reading is,
    or what the right documents are in response to a query.
- Gold standards used both for training and for evaluation
- But testing must be done on unseen data (held-out test set; train/test split)

Don't ever train on data that you'll use in testing!!

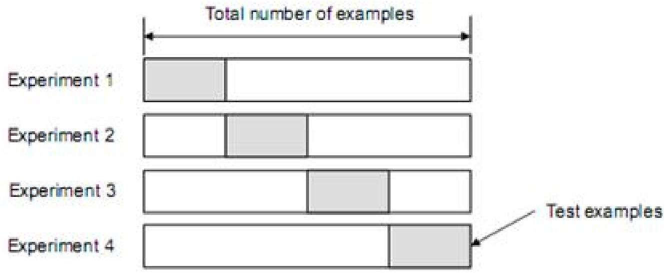
Tuning

- Often, in designing a system, you'll want to **tune** it by trying several configuration options and choosing the one that works best empirically.
  - E.g., Lidstone (add-λ) smoothing; choosing features for text classification.
- If you run several experiments on the test set, you risk **overfitting** it; i.e., the test set is no longer a reliable proxy for new data.
- One solution is to hold out a second set for tuning, called a **development** ("dev") set. Save the test set for the very end.

Cross-validation

What if my dataset is **too small** to have a nice train/test or train/dev/test split?

- ***k*-fold cross-validation**: partition the data into k pieces and treat them as mini held-out sets. Each **fold** is an experiment with a different held-out set, using the rest of the data for training:



- After k folds, every data point will have a held-out prediction!

- If tuning the system via cross-validation, still important to have a separate blind test set.

Measuring a Model's Performance

Accuracy: Proportion model gets right:

$$\frac{|\text{right}|}{|\text{test-set}|} \times 100$$

E.g., POS tagging (state of the art $\approx 96\%$).

Measuring a Model's Performance

Precision, Recall, F-score

- For isolating performance on a particular label in multi-label tasks, or
- For chunking, phrase structure parsing, or anything where word-by-word accuracy isn't appropriate.
- F_1-score: Harmonic mean of
  precision (proportion of model's answers that are right) and
  recall (proportion of test data that model gets right).
- E.g., for the POS tag `NN`:

$$
\begin{aligned}
P &= \frac{|\text{tokens correctly tagged } \texttt{NN}|}{|\text{all tokens automatically tagged } \texttt{NN}|} &&= \frac{\text{TP}}{\text{TP+FP}} \\
R &= \frac{|\text{tokens correctly tagged } \texttt{NN}|}{|\text{all tokens gold-tagged } \texttt{NN}|} &&= \frac{\text{TP}}{\text{TP+FN}} \\
F_1 &= \frac{P \cdot R}{P+R}
\end{aligned}
$$

Upper Bounds, Lower Bounds?

Suppose your POS tagger has 95% accuracy? Is that good? Bad??

Upper Bound: Turing Test:

- When using a human Gold Standard, check the agreement of humans against that standard.

Lower Bound: Performance of a 'simpler' model (**baseline**)

- Model always picks most frequent class (**majority baseline**).

- Model assigns a class randomly according to:

  1. Even probability distribution; or
  2. Probability distribution that matches the observed one.

Suitable upper and lower bounds depend on the task.

Measurements: What's Significant?

- We'll be measuring things, and comparing measurements.

- What and how we measure depends on the task.

- But all have one issue in common:

  Are the differences we find significant?

- In other words, should we interpret the differences as down to pure chance? Or is something more going on?

- Is our model significantly better than the baseline model?
  Is it significantly worse than the upper bound?

Example: Tossing a Coin

- I tossed a coin 40 times; it came up heads 17 times.
- Expected value of fair coin is 20. So we're comparing 17 and 20.
- If this difference is *significant*, then it's (probably) not a fair coin.
  If not, it (probably) is.

Normal distributions

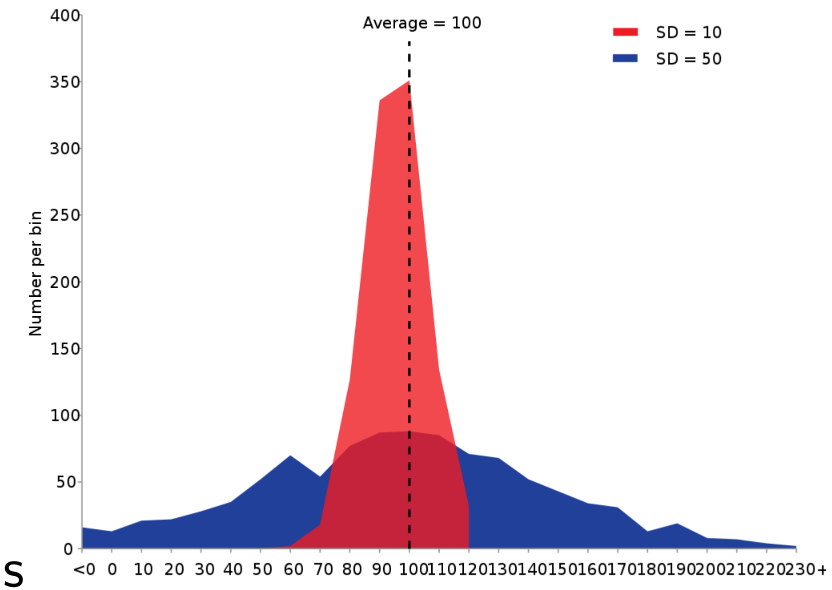
Significance measurement can be complex to understand, but the basic idea is simple (for normal distributions):

- Measure difference in terms of standard deviation

Standard deviation is essentially a measure of how representative the mean is

- The more outliers, and the further they are from the mean
- The less representative the mean is

- The standard deviation quantifies this



Mean and standard deviation

Definitions:

Mean of N measurements

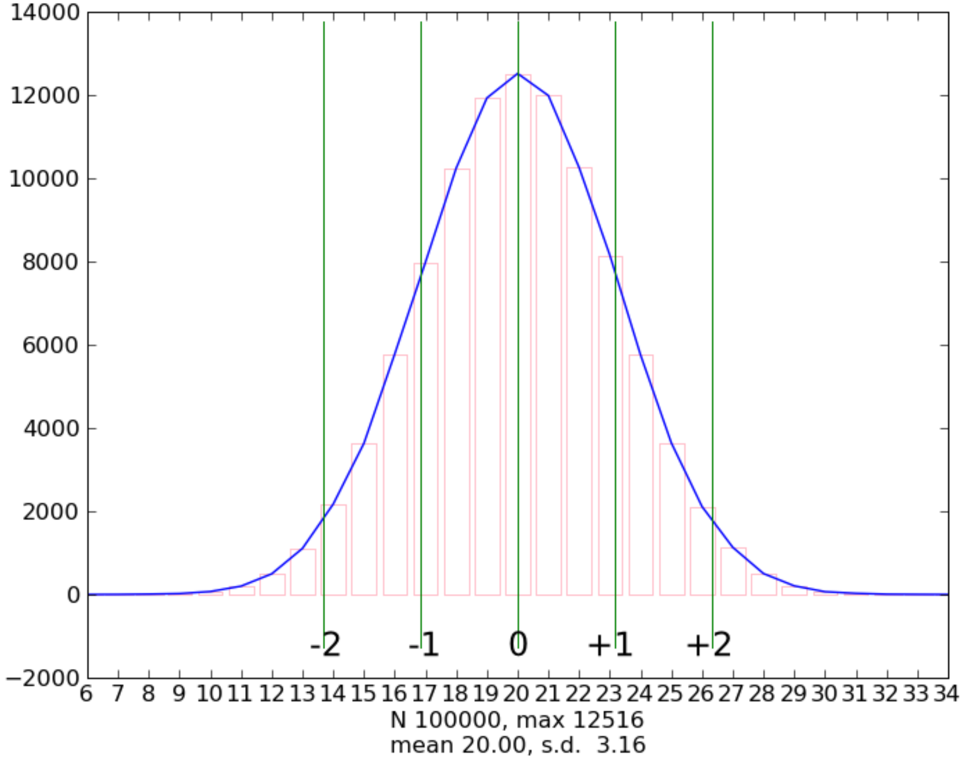
$\frac{n_1+n_2+n_3+\cdots+n_N}{N}$ Call this μ

Standard deviation of N measurements

$\sqrt{\frac{(n_1-\mu)^2+(n_2-\mu)^2+(n_3-\mu)^2+\cdots+(n_N-\mu)^2}{N}}$

The outcomes if the coin is fair

The distribution over 100,000 trials (i.e., in each trial toss coin 40 times) is similar to a normal distribution curve.



- The peak is at 50% heads, but lots of other plausible outcomes.
- Even a result more than two standard deviations out will come up a bit less than 1 in 20 trials (i.e., 2,000/100,000)

Which Significance Test?

- Parametric when the underlying distribution is normal.
  - t-test, z-test,. . .
  - You don't need to know the mathematical formulae; available in statistical libraries!

- Non-Parametric otherwise.
  - Usually do need non-parametric tests: remember Zipf's Law!
  - Can use McNemar's test or variants of it.

See Smith (2011, Appendix B) for a detailed discussion of significance testing methods for NLP.

Error Analysis

- Summary scores are important, but don't always tell the full picture!
- Once you've built your system, it's always a good idea to dig into its output to identify patterns.
  - Quantitative *and* qualitative (look at some examples!)
  - You may find bugs (e.g., predictions are always wrong for words with accented characters)
  - Or think of ways to improve your system
- A **confusion matrix** can help in spotting problem areas

Confusion Matrices

		Estimated Emotion							
		Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral	*Emotion Recog. Rate*
True Emotion	Anger	**19**	0	2	0	**3**	0	0	79.2%
	Boredom	1	**8**	1	1	0	1	**7**	42.1%
	Disgust	0	1	**6**	0	1	0	3	54.5%
	Fear	1	3	2	**7**	2	0	1	43.8%
	Happiness	**3**	0	3	2	**5**	0	2	33.3%
	Sadness	0	0	0	0	0	**14**	0	100.0%
	Neutral	0	**5**	1	0	0	0	**13**	68.4%
	HMM Recog. Rate	79.2%	47.1%	40.0%	70.0%	45.5%	93.3%	50.0%	

Figures on the main diagonal are for the correct answer

- Others are mistakes
- Mistakes in bold are perhaps large enough to worry about...

Tasks where there is > 1 right answer

Example: A Paraphrasing Task

- Estimate that *John enjoyed the book* means *John enjoyed reading the book*.

- Lots of closely related words to *read* are good too:
  skim through, go through, peruse, etc.

Tasks where there is > 1 right answer

Example: A Paraphrasing Task

- Estimate that *John enjoyed the book* means *John enjoyed reading the book*.
- Lots of closely related words to *read* are good too:
  skim through, go through, peruse, etc.

Evaluation: 'Turing Test'

- Classify candidate paraphrases as high, medium or low probability.
- Measure correlation between human vs. machine's judgements.
- Result was 0.64. Is that good?
- Upper bound: average correlation between two human judges! That's 0.74.
- Can use above tests to measure if these are significantly different.

Summary

- Lots of things we might be evaluating.
- Generally, NLP systems evaluated against gold standard data, which is often quite expensive to collect.
- All that is "gold" does not glitter. Important to remember where the data came from and measure reliability.
- You compare performance of your model against: upper bound, baseline model, someone else's model, and use an appropriate significance test to see if differences are 'real' or within margin of error (i.e., likely due to chance).