CSCI 544, Lecture 10: Experiment design; Annotation

Ron Artstein

2022-09-22

Administrative notes: deadlines

Written Assignment Peer Grading due tonight

- So far, about 72% of the students have completed the grading

Coding Assignment 2 due September 27

Project:

Due Date	Task
September 20	Form project teams (52 teams)
September 20–29	**Initial discussion with TA**
October 4	**Project proposal**
November 3	Project status report
Nov 29/Dec 1	Poster presentations (in class)
December 1	Final report
December 3	Self-evaluation and peer grading

Pick your TA!

Evaluation

Typical evaluation of NLP: compare to “gold standard” reference

- Accuracy
- Precision, recall, F-measure

Evaluation of generation: lexical similarity to reference

- Word Error Rate (speech recognition)
- BLEU, METEOR (translation); ROUGE (summarization)
- Multiple references capture lexical variation

Are lexical similarity measures good for dialogue?

- Much more lexical variation in appropriate utterances

Current dialogue evaluation practices

Mix of methods (Finch and Choi 2020)

- Automated: similarity to reference, similarity to context
- Human-rated: appropriateness, coherence, consistency
- ☛ Judge quality of a contribution

Automated measures that try to predict human ratings

Experiment design

Independent variables Manipulated by the experimenter; control

Dependent variables Measured to see if affected by the independent variables

How can we tell if the dependent variable really is affected?

> **If you need statistics to prove the results of your experiment, then you ought to have designed a better experiment.**
>
> *Attributed to Lord Ernest Rutherford*

Hypothesis testing

Is the observed outcome due to random sampling?

Null hypothesis No effect; results due to random sampling

Alternative hypothesis Results **not** due to random sampling

Statistical model: probability of various outcomes if only random sampling is at play

- If probability of observed results is low, reject null hypothesis

Some common statistical tests

- t-test: means of two samples
- ANOVA: means of multiple samples
  - main effects; interactions; simple effects
- chi-squared test: compare frequencies of categories
- correlation, regression: two (or more) continuous variables

Confounds and controls

Outcome may not be the result of random sampling, but not our variable either

Confound Another variable that may affect the result

Control variable Not interested in it, but may affect results

Random variable A variable we can't control

Participants and language

Human participants may be affected by order, etc.

Between subjects Assign different participants to each condition

Within subjects (repeated measures) Same participants in all conditions.

In language studies, the choice of items may also have an effect: not all verbs/nouns are the same.

> Herbert H. Clark, The language-as-fixed-effect fallacy, Journal of Verbal Learning and Verbal Behavior 12(4):335–359, 1973

The empirical revolution (1990s)

Large-scale, broad coverage systems

Emphasis on **formal evaluation** of performance

- Track improvement over time, compare systems
- Typically single component (e.g. parser, tagger)
- Competitions: single task, multiple teams

Quantitative evaluation

- Reference target (“gold standard”)
- Precision, recall, F-measure
- Objective measures (task success, time on task)

Automatic evaluation allows machine learning

Annotated corpora

Annotated corpora are needed for:

- Supervised learning – training and evaluation
- Unsupervised learning – evaluation
- Hand-crafted systems – evaluation
- Analysis of text

Annotations need to be **correct.**

Why measure annotator agreement

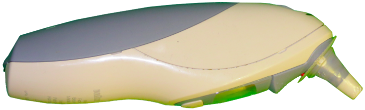
Agreement	is measured between annotations of a single text.
Reliability	measures consistency of an instrument.
Validity	is the correctness relative to a desired standard.

Reliability is a property of a process

Repeated measures with two thermometers

Mercury $\pm 0.1\,^{\circ}\mathrm{C}$

Infrared $\pm 0.4\,^{\circ}\mathrm{C}$



The mercury thermometer is more reliable.

- But what if it's not calibrated properly?

Reliability is a **minimum requirement** for an annotation process.

- Qualitative evaluation also necessary.

Reliability and agreement

Reliability = **consistency** of annotation

- Needs to be measured on the same text.
- Different annotators.
- Work **independently**

If independent annotators mark a text the same way, then:

- They have internalized the same scheme (instructions).
- They will apply it consistently to new data.
- Annotations may be correct.

Results **do not generalize** from one domain to another.

Observed (pairwise) agreement

Observed agreement: proportion of items on which 2 coders agree.

Detailed Listing

Item	Coder 1	Coder 2
a	Boxcar	Tanker
b	Tanker	Boxcar
c	Boxcar	Boxcar
d	Boxcar	Tanker
e	Tanker	Tanker
f	Tanker	Tanker
	⋮	⋮

Contingency Table

	Boxcar	Tanker	Total
Boxcar	41	3	44
Tanker	9	47	56
Total	50	50	100

Agreement: $\frac{41 + 47}{100} = 0.88$

High agreement, low reliability

Two psychiatrists evaluating 1000 patients.

	Normal	Paranoid	Total
Normal	990	5	995
Paranoid	5	0	5
Total	995	5	1000

- Observed agreement = 990/1000 = 0.99
- Most of these patients probably aren't paranoid
- No evidence that the psychiatrists identify the paranoid ones
- High agreement **does not indicate** high reliability

Chance agreement

Some agreement is expected by chance alone.

- Randomly assign two labels $\rightarrow$ agree half of the time?
  - Depends on the distributon!
- The amount expected by chance varies depending on the annotation scheme and on the annotated data.

Meaningful agreement is the agreement **above chance.**

Correction for chance

How much of the observed agreement is above chance?

	A	B	Total
A	44	6	50
B	6	44	50
Total	50	50	100

$$\underset{\textit{Total}}{\boxed{\begin{matrix}\mathbf{44} & 6 \\ 6 & \mathbf{44}\end{matrix}}}_{\mathbf{88}} = \underset{\textit{Chance}}{\boxed{\begin{matrix}\mathbf{6} & 6 \\ 6 & \mathbf{6}\end{matrix}}}_{\mathbf{12}} + \underset{\textit{Above}}{\boxed{\begin{matrix}\mathbf{38} & 0 \\ 0 & \mathbf{38}\end{matrix}}}_{\mathbf{76}}$$

Agreement: 88/100
Due to chance: 12/100
Above chance: 76/100

Expected agreement

Observed agreement (A_o)*: proportion of actual agreement*
Expected agreement (A_e)*: expected value of* A_o

Amount of agreement above chance: $A_o - A_e$
Maximum possible agreement above chance: $1 - A_e$

Proportion of agreement above chance attained: $\dfrac{\mathbf{A_o - A_e}}{\mathbf{1 - A_e}}$

Scott's Pi, Fleiss's Kappa, Siegel and Castellan's K

Total number of judgments: $N = \sum_q \mathbf{n}_q$

Probability of one coder picking category q: $\frac{\mathbf{n}_q}{N}$

Prob. of two coders picking category q: $\left(\frac{\mathbf{n}_q}{N}\right)^2$ *[biased estimator]*

Prob. of two coders picking same category: $A_e = \sum_q \left(\frac{\mathbf{n}_q}{N}\right)^2$

	Normal	*Paran*	*Total*
Normal	*990*	*5*	*995*
Paranoid	*5*	*0*	*5*
Total	*995*	*5*	*1000*

$A_o = 0.99$

$A_e = .995^2 + .005^2 = 0.99005$

$K = \frac{0.99 - 0.99005}{1 - 0.99005} \approx -0.005$

Interpreting agreement

Agreement measures
are not
hypothesis tests

- Evaluating magnitude, not existence/lack of effect
- Not comparing two hypotheses
- No clear probabilistic interpretation

Textbook usage paradigm

Conduct a reliability study with:

- Written annotation guidelines
- Generally available coders
- Representative sample of annotation materials

In order to validate annotation **scheme** and **procedure**

With a good procedure, will annotations will be correct?

Annotation model

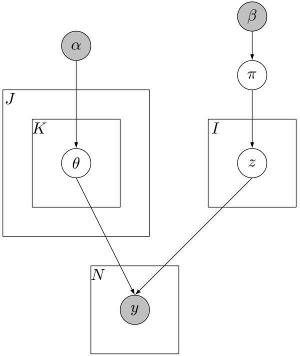
Annotation errors are not random

Affected by:

- Item category
- Annotator
- (Item difficulty)

Use this information to infer true label

- Graphical model
- Multiple annotations
- EM algorithm
- Confidence level



Passonneau and Carpenter (2014), Figure 1

Differences in the annotated material

Kang et al. 2012 , AAMAS: identify smiles in videos

- Smiles are easier to detect on some people than others

Not all coders are equal

Scott, Barone and Koeling, LREC 2012

- Annotate hedges in medical text as likelihood

> **Possible** early pneumonia. . .
> . . . **could** represent pneumonia. . .

- Two annotator populations differ in **medical training**
- Systematic differences between annotators: medically trained interpret hedges as expressing greater likelihood

Each population of coders (instrument) has a certain reliability, but one is probably more correct.