Data collection & annotation

CMSC 723 / LING 723 / INST 725

Hal Daumé III [he/him]

19 Sep 2019

Announcements, logistics

- HW2 update is posted
  if you downloaded an old version, please get update

What part of your data is language?

- Input
- Output
- Supplemental annotation
- Other

How do you collect this data

- Place microphones around the Iribe center
- Download stuff from the web
  - Periodicals
  - Blogs / social media (twitter, weibo, facebook, etc.)
  - Discussion fora
  - Political debates / legal docs
- Emails
- Have people come in to a lab and produce language
- Crowdworkers
- Other?

How do you ensure you get the language (and variety) that you want?

- Choice of source
- Language id tools
- Heuristics based on common words

Annotation types

- Document-level labels (classification) or scores (regression)
- Word- or span-level labels (tagging)
- Arbitrary labels (e.g., translation, captioning)

Annotation guidelines

- How you convey to annotators what you want them to annotate
- Can be anywhere from 1 paragraph to >= 104 pages
  - e.g., www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf

Example annotation guidelines

(h/t Naeemul Hassan, UMD Journalism, iSchool; CIKM'15, C+J'15; "ClaimBuster")

- **NFS: Non-factual sentence**
  Subjective sentences (opinions, beliefs, declarations) and many questions fall under this category. These sentences do not contain any factual claim.
  - *But I think it's time to talk about the future*
  - *You remember the last time you said that?*

- **Unimportant Factual Sentence (UFS)**
  They contain factual claims but are not check-worthy. The general public will not be interested in knowing whether these sentences are true or false.
  - *Next Tuesday is Election Day.*
  - *Two days ago we ate lunch at a restaurant*

- **Check-worthy Factual Sentence (CFS)**
  They contain factual claims and the general public will be interested in knowing whether the claims are true.
  - *He voted against the first Gulf War*
  - *Over a million and a quarter Americans are HIV-positive*

To what degree do annotators agree?

- For classification, could just measure % agreement
- Fails to capture “agreement by change”

- Enter Cohen’s Kappa statistic:

Accuracy

Chance

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

Assume annotations are independent

$$p_e = \sum_k \widehat{p_{k12}} = \sum_k \widehat{p_{k1}}\widehat{p_{k2}} = \sum_k \frac{n_{k1}}{N}\frac{n_{k2}}{N} = \frac{1}{N^2}\sum_k n_{k1}n_{k2}$$

k is a category, i is rater, n_{ki} is # times rater i said category k, N=#items

To what degree do annotators agree?

- For classification, could just measure % agreement
- Fails to capture “agreement by change”

- Enter Cohen’s Kappa statistic:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

Accuracy

Chance

< 0.00	Poor agreement	0.41 – 0.6	Moderate agreement
0.00 – 0.2	Slight agreement	0.61 – 0.8	Substantial agreement
0.21 – 0.4	Fair agreement	0.81 – 1.0	Almost perfect agreement

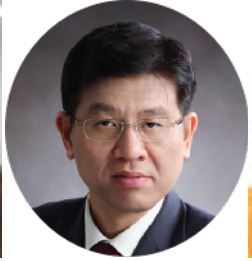
[Landis & Koch, 1977]

Usual annotation procedure

1. Look at the data
2. Develop initial annotation guidelines
3. Have two people go of and annotate a small number of examples
4. Compare results, update annotation guidelines
5. Throw out data from #3
6. Have multiple annotators annotate lots of data
7. Compute agreement
8. If too low, go back to #4
9. Adjudicate differences

Case study: NUCLE

[Dahlmeier, Ng, Wu; 2013]


- English courses at CELC, all NUS undergraduate students
- Designed for students who need language support for academic studies
- Essays written as course assignments on wide range of topics

"Public spending on the aged should be limited so that money can be diverted to other areas of the country's development." Do you agree?
Surveillance technology such as RFID (radio-frequency identification) should not be used to track people (e.g., human implants and RFID tags on people or products). Do you agree? Support your argument with concrete examples.
Choose a concept or prototype currently in research and development and not widely available in the market. Present an argument on how the design can be improved to enhance safety. Remember to consider influential factors such as cost or performance when you summarize and rebut opposing views. You will need to include very recently published sources in your references.

Table 4: Example question prompts from the NUCLE corpus.

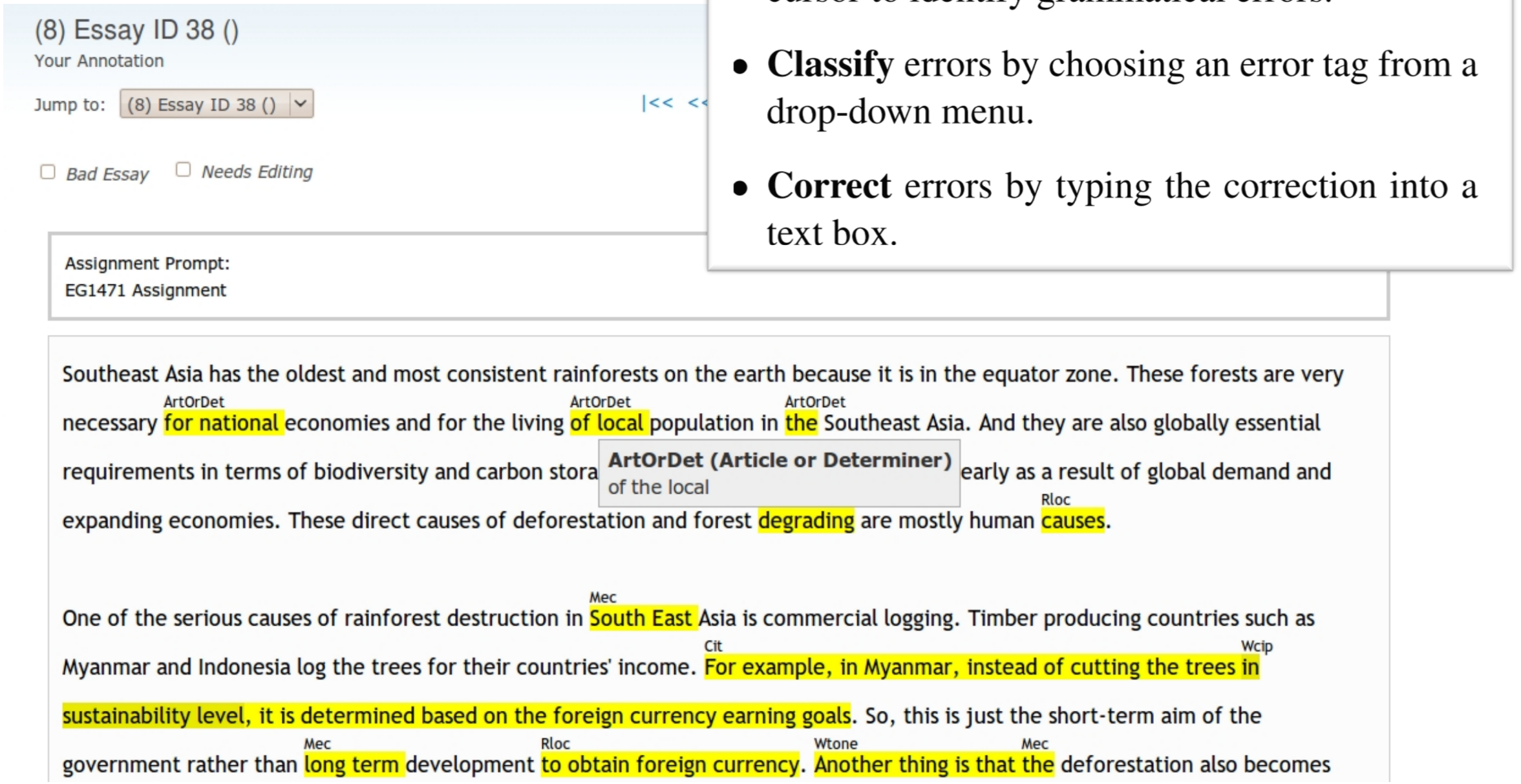
Annotators

- 10 CELC English instructors
- 1414 essays, over 1.2m words
- Only doubly-annotated a subset of the corpus

Case study: NUCLE

[Dahlmeier, Ng, Wu; 2013]

- **Select** arbitrary, contiguous text spans using the cursor to identify grammatical errors.
- **Classify** errors by choosing an error tag from a drop-down menu.
- **Correct** errors by typing the correction into a text box.



Case study: NUCLE

[Dahlmeier, Ng, Wu; 2013]

Error Tag	Error Category	Description / Example
Word Order		
WOinc	Incorrect sentence form	Why can **[not we — we not]** choose more intelligent and beautiful babies?
WOadv	Adverb/adjective position	It is similar to the murder of many valuable lives **[only based — based only]** on the couple's own wish.
Transitions		
Trans	Link words/phrases	In the process of selecting the gender of the child, ethical problems arise **[where — because]** many innocent lives of unborn fetuses are taken away.
Mechanics		
Mec	Punctuation, capitalization, spelling, typos	The **[affect — effect]** of that policy has yet to be felt.
Redundancy		
Rloc	Local redundancy	Currently, abortion is available to end a life only **[because of — because]** the fetus or embryo has the wrong sex.
Citation		
Cit	Citation	Poor citation practice.
Others		
Others	Other errors	Any error that does not fit into any other category, but can still be corrected.
Um	Unclear meaning	The quality of the passage is so poor that it cannot be corrected.

Error Tag	Error Category	Description / Example
Verbs		
Vt	Verb Tense	A university **[had conducted — conducted]** the survey last year.
Vm	Verb modal	No one **[will — would]** bother to consider a natural balance.
V0	Missing verb	This **[may — may be]** due to a traditional notion that boys would be the main labor force in a farm family.
Vform	Verb form	Will the child blame the parents after he **[growing — grows]** up?
Subject-verb agreement		
SVA	Subject-verb-agreement	The boy **[play — plays]** soccer.
Articles/determiners		
ArtOrDet	Article or Determiner	From the ethical aspect, sex selection technology should not be used in **[non-medical — a non-medical]** situation.
Nouns		
Nn	Noun Number	Sex selection should therefore be used for medical **[reason — reasons]** and nothing else.
Npos	Noun possessive	The education of **[mother's — mothers]** is a significant factor in reducing son preference.
Pronouns		
Pform	Pronoun form	90% of couples seek treatment for family balancing reasons and 80% of **[those — them]** want girls.
Pref	Pronoun reference	Moreover, children may find it hard to communicate with **[his/her — their]** parents.
Word choice		
Wcip	Wrong collocation/idiom/preposition	Singapore, for example, has invested heavily **[on — in]** the establishment of Biopolis
Wa	Acronyms	Using acronyms without explaining what they stand for.
Wform	Word form	Sex-selection may also result in **[addition — additional]** stress for the family.
Wtone	Tone	**[Isn't it — Is it not]** what you always dreamed for?
Sentence Structure		
Srun	Runons, comma splice	**[Do spare some thought and time, we can make a difference! — Do spare some thought and time. We can make a difference!]** (Should be split into two sentences)
Smod	Dangling modifier	**[Faced — When we are faced]** with the unprecedented energy crisis, finding an alternative energy resource has naturally become the top priority issue.
Spar	Parallelism	The use of sex selection would prevent rather than **[contributing — contribute]** to a distorted sex ratio.
Sfrag	Fragment	Although he is a student from the Arts faculty.
Ssub	Subordinate clause	It is the wrong mindset of people that boys are more superior than girls **[should — that should]** be corrected.

Case study: NUCLE

[Dahlmeier, Ng, Wu; 2013]

Source	:	This phenomenon opposes the real .
Annotator A	:	This phenomenon opposes (the → ϵ (ArtOrDet)) (real → reality (Wform)) .
Annotator B	:	This phenomenon opposes the (real → reality (Wform)) .

- **Identification** Agreement of tagged tokens regardless of error category or correction.
- **Classification** Agreement of error category, given identification.
- **Exact** Agreement of error category and correction, given identification.

Case study: NUCLE

[Dahlmeier, Ng, Wu; 2013]

- **Identification** Agreement of tagged tokens regardless of error category or correction.
- **Classification** Agreement of error category, given identification.
- **Exact** Agreement of error category and correction, given identification.

Annotators	Kappa-iden	Kappa-class	Kappa-exact
A – B	0.4775	0.6206	0.5313
A – C	0.3627	0.5352	0.4956
B – C	0.3230	0.4894	0.4246
Average	0.3877	0.5484	0.4838

Re-annotation protocols

- Annotate a small amount of data
- Train a simple system
- Use that system to annotate more data
- Have humans correct the annotations

- Beware the Brill Tagger Effect!
  (aka confirmation bias or automation bias)
  (also note that these annotations are no longer independent)

- Can also do active learning, eg aclweb.org/anthology/W07-1516

What about non-tagging problems?

- Generally no agreed-upon measure of agreement
- Standard practice:
  - Treat one annotator as “gold standard”
  - Compute (insert_your_metric_here) of another annotator against that
  - Repeat for all annotators
- Does *not* take into account “chance agreement”
- But provides useful *skyline* for systems
  - *Baseline-skyline* gap tells you how much to work on this task

When others collected data….

[Mieskes, 2017]

- Do you make your data available? HOW?

Venue	# papers	# data published	Ratio
NAACL	182	57	31.3%
ACL	231	63	27.3%
EMLNP	264	81	30.7%
Coling	337	89	26.4%
LREC	744	414	55.6%
total	1758	704	40.0%

Category	Percentage
Link available	65.2%
Link does not work	15.7%
No Link	31.4%
On Request	1.8%
Proprietary data	< 1%

- How have you protected participants of data collection process?
  - 32.8% (of 704) collected data from social media or other sensitive data
  - 96.5% did not specify anything done to anonymize
  - 7% of LREC data included recordings of elderly, children, FLLs; 90% non-anon

- Is there enough information to *re-collect* the data?

When you collect data….

[Mieskes, 2017; Gebru, Morgenstern, et al., 2018; Bender & Friedman, 2018]

- Has data been collected?
- How was this data collected and processed?
- Was previously available data used/extended – which one?
- Is a link or a contact given?
- Where does it point (private page, research institute, official repository)?

Important to think about these issues *before* you begin collecting and annotating data!

- Data statements for NLP
  - Curation rationale
  - Language variety
  - Speaker demographic
  - Annotator demographic
  - Speech situation
  - Text characteristics
  - Recording quality
  - Other
  - Provenance Appendix
- Datasheets for Datasets
  - Motivation/curators
  - Composition
  - Collection process
  - Preprocessing/cleaning
  - Uses
  - Distribution
  - Maintenance

Other things possibly worth knowing….

- There are various extensions to Cohen's kappa to deal with
  - Lots of annotators
  - Missing data (not every annotator annotated everything)
  - Scalar responses (vs categorical)
  - Significance tests on kappa
  - Decent catchall: Krippendorff's alpha

> Krippendorff's *alpha* brings several known statistics under a common umbrella, each of them has its own limitations but no additional virtues.

[see Wikipedia page]

Just because you *can* collect a dataset…

- Many examples over past few years of “easily broken datasets”
  (I won’t point fingers in the slides…)

- The Riloff Rule (as remembered by Hal):

  If you collect a dataset and distribute it,
  if it’s not high quality and worthwhile,
  you will waste tons of other people’s time.

- Hal’s heuristic:

  If tomorrow someone got 98% accuracy on this dataset,
  would I be *happy* or *sad.* If sad, don’t do it.

Today

- Where do you get your data from?
- How do you annotate it?
- How do you measure annotator agreement?
- How can you make sure you've produced high quality data