Chapter NLP:II

II. Corpus Linguistics

- Empirical Research
- Hypothesis Testing
- Text Corpora
- Data Acquisition
- Data Annotation

Data Annotation

Definition 1 (Annotation)

Annotation is the process of marking or adding information (labels, categories) to data (examples, items, markables) that is required for processing.

- Classes and labels of documents, marked spans, span labels, reference summaries, image descriptions
- An annotation can also specify relations between annotations.
- Annotation is done in annotation tasks, often by human annotators (raters, voters, coders, . . .).

Topic: ”Google revenues“ **Genre:** ”News article“

Data Annotation

Sources of Annotations

- Manual annotation
  Annotations are added by humans. Often called ground truth or gold standard.

- Automatic annotation
  Automatically add annotations from external sources or from different model or algorithm. Sometimes called silver standard.

Data Annotation

Automatic Annotation: Sources

Automatic annotations are cost effective and enable large corpora.

- Self-supervision
  The annotations are part of the original data. e.g. language modeling
- Semi-supervision
  The annotations for new data are derived from already annotated data.
- Weak or distant supervision
  The annotations are derived from relations between the data and external knowledge. Sentiment from user ratings, entity relations from databases
- Simulated annotators like LLMs. [Gilardi, 2023]
  LLMs already outperform crowd workers in some text generation tasks.

Automatic annotations are often noisy and must be filtered or cleaned to improve the quality.

Data Annotation

Manual Annotation: Sources

Manual annotations are time-consuming and expensive but assumed to be correct and of high quality.

- Experts
  Annotations are done by experts trained for the task and in the general area of the annotation (linguistics, psychology, . . .). and expensive.

- Laypeople
  Training and supervising laypeople on a task. This can be a cheaper alternative for easy tasks that need little expertise.

- Crowdsourcing or Click work
  Using a platform to recruit click workers with little training or supervision. Easy to recruit many annotators, but needs good task design and evaluation for good quality results.

Data Annotation

Manual Annotations: Software

- Prodigy [prodi.gy]
  - NLP focussed tool with a deep integration of spacy, LLMs, and active learning support. Allows custom templates via html.
  - Expensive license.
- Label Studio [labelstud.io]
  - General tool with templates for many tasks, some options for task design.
  - Free version with some limitations, difficult to integrate in automated workflows like active learning.
- Doccano [github.com/doccano]
  - Open source, but quite limited in features.

Never implement your own annotation tool without a very good reason.

Data Annotation

Crowdsourcing [Suhr et al., 2021][Callison-Burch et al., 2021]

Crowdsourcing refers to techniques using collective intelligence:

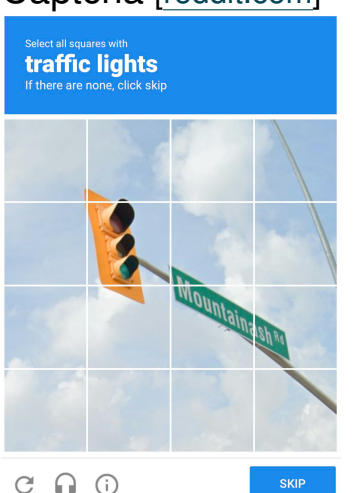
- Distribute annotation work to many independent annotators.
- Use individual expertise on small parts of the whole. Wikipedia, OpenStreetMap, ...
- "Wisdom of the crowds": Even if individual assessments vary, the average is often close to the truth. Citizen Science, Captchas, Francis Galton's Ox, ...
- Diversify the pool of annotators.

Crowdsourcing is good in NLP, AI, and IR when:

- many examples are required,
- the task can be split up and parallelized,
- the individual annotations require little training, and
- the results can be averaged across annotators.

OpenStreetMap [openstreetmap.org]

Captcha [reddit.com]



Data Annotation

Crowdsourcing: Platforms [Suhr et al., 2021][Callison-Burch et al., 2021]

Amazon Mechanical Turk: [mturk.com]

- For microtask (a few seconds up to minutes)
- Supports many (100–10K) but low skilled workers (mostly US/India).
- Allows custom templates via html and javascript.

UpWork [upwork.com]

- For recruiting experts and specialists.
- Usually more expensive.

There are other platforms like Toloka or Appen for B2B or AI click work.

Data Annotation

Crowdsourcing: Issues [Suhr et al., 2021][Callison-Burch et al., 2021]

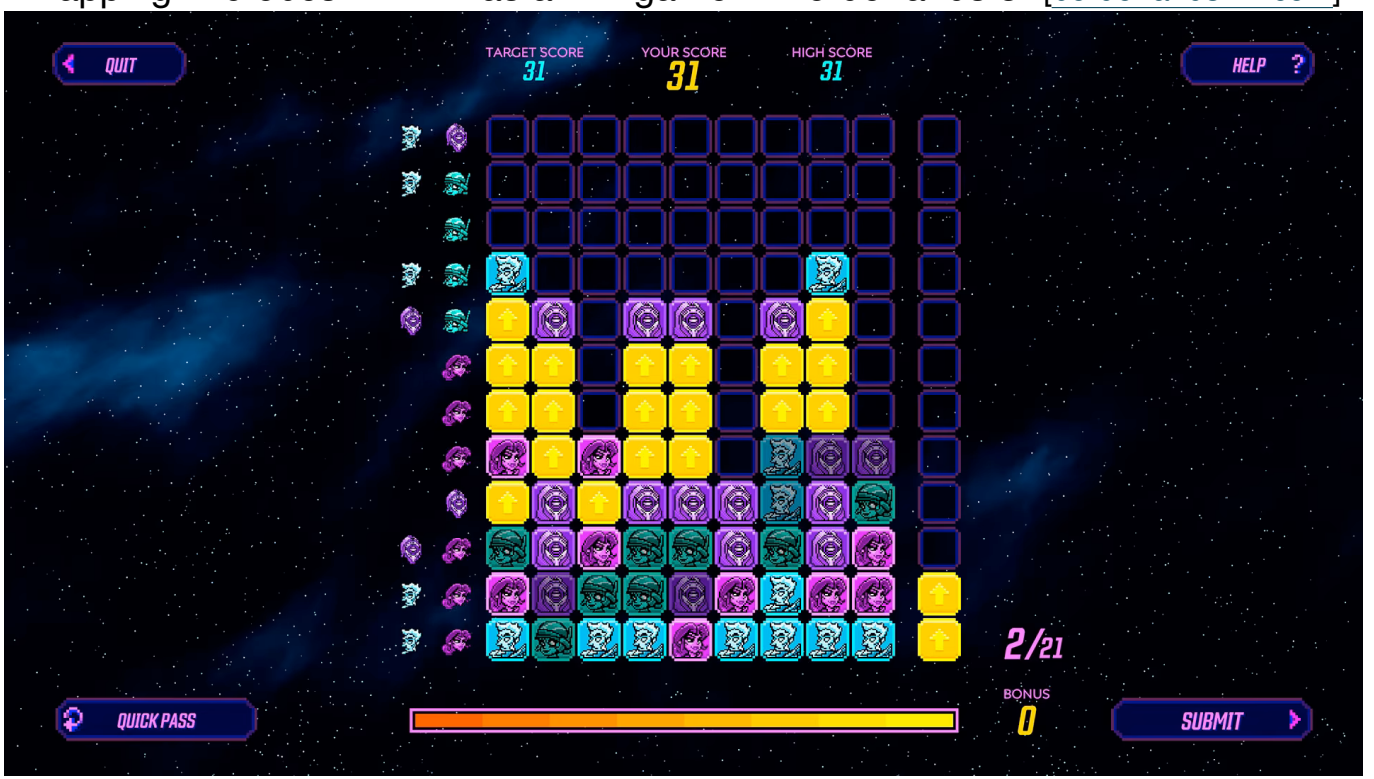
- Recruitment
  Find annotators with a given background
  (Experience in crowd work, location, language)
- Qualifications
  Train annotators and test their abilities to do the task.
- Quality Control
  Test annotations for correctness. Improve correctness via task design.
- Reputation
  Good annotators more often take tasks from reputable organizers. Be fair and pay annotators well and in-time.
- Payment
  Low pay has adverse affects: poor quality annotations, market degradation, research ethics. $\rightarrow$ Time the tasks and pay minimum wage.

Data Annotation

Crowdsourcing: Gamification

Idea: Recruit motivated annotators by hiding the task in games or designing the annotation task as a game.

Mapping microbes in DNA as a minigame in Borderlands 3. [borderlands.2k.com]



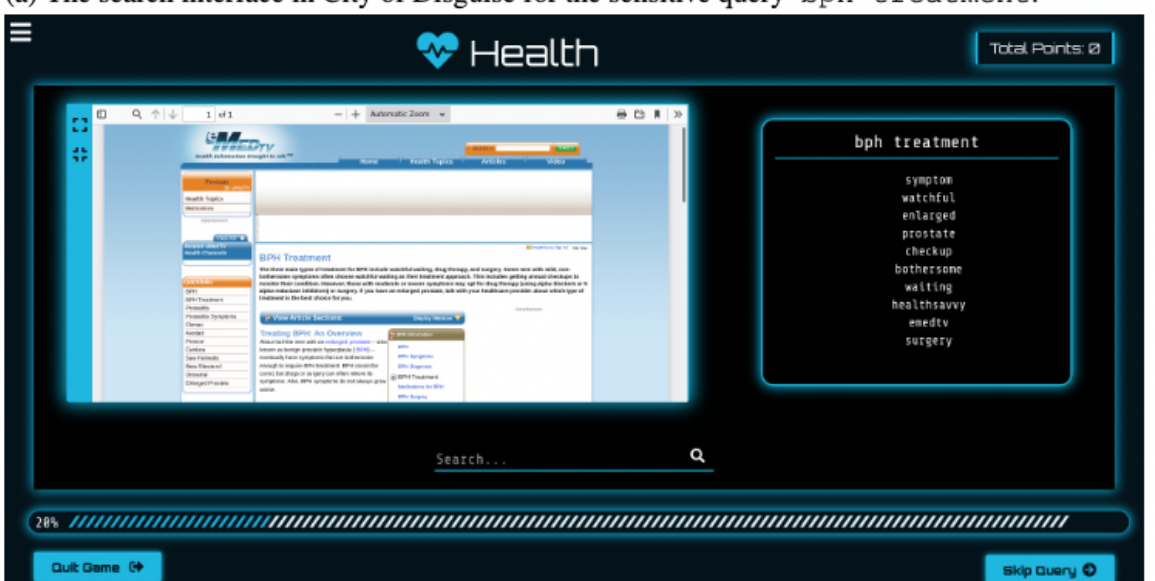
Data Annotation

Crowdsourcing: Gamification

Idea: Recruit motivated annotators by hiding the task in games or designing the annotation task as a game.

Obfuscating search queries to hide sensitive information in City in Disguise. [Fröbe, 2022]

(a) The search interface in City of Disguise for the sensitive query `bph treatment`.



(b) Categories in City of Disguise.

(c) Scoring for a successful obfuscation.

Data Annotation

Annotation Tasks

Annotation Tasks: the process of producing correct and reproducible annotations in sufficient quantity within a given budget.

- Correct
  The annotations can be trusted, e.g. experts have a high agreement.
- Reproducible
  Annotators produce the same annotations when repeating the task.

> *"I have a very large collection of clean labeled data"* – No One

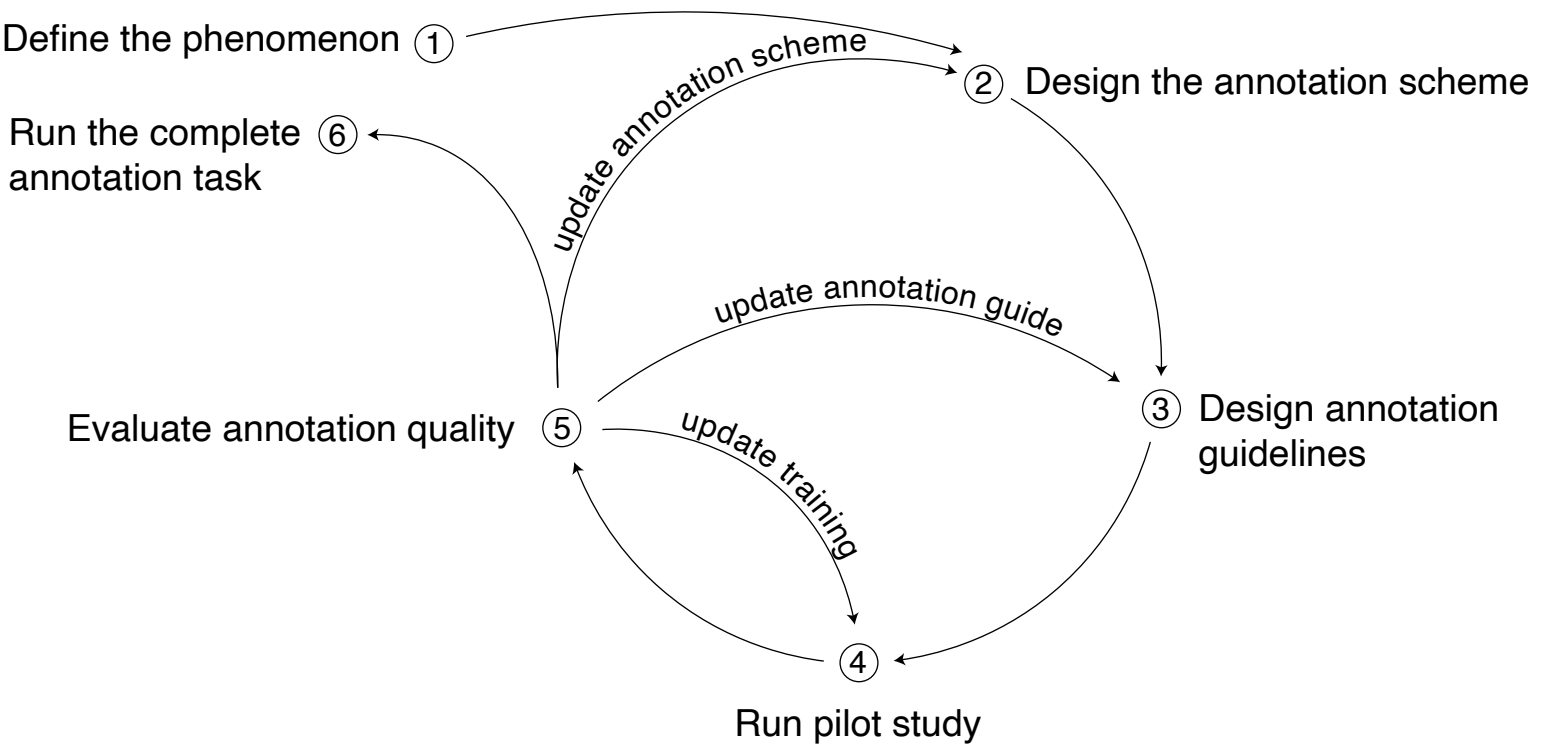
Challenges:

- Disagreement
  In many cases, there are different beliefs of what is a valid annotation.
- Budget, size, and correctness trade-off
  Different annotation strategies trade correctness against size.
  Some noise is acceptable for many projects.

Data Annotation

Annotation Tasks

Designing annotation tasks is iterative (similar to software development or human-centered design).

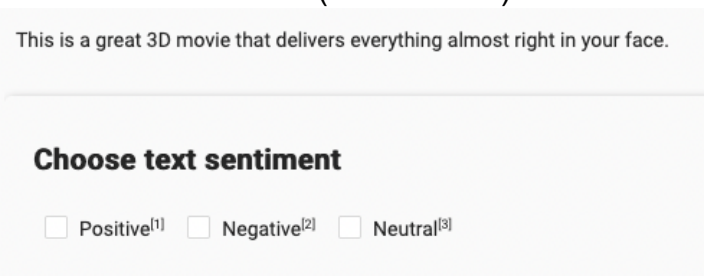


Data Annotation

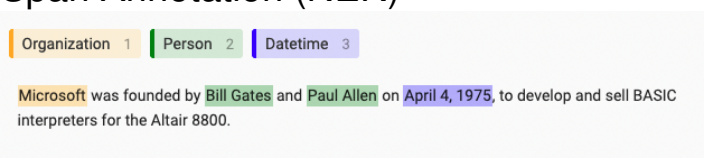
Annotation Schemes

The annotation scheme describes the form (i.e. layout) and scope (i.e options) of the annotation task. Typical schemes for NLP tasks are:

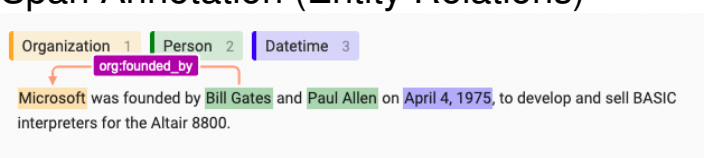
Text Classification (Sentiment)



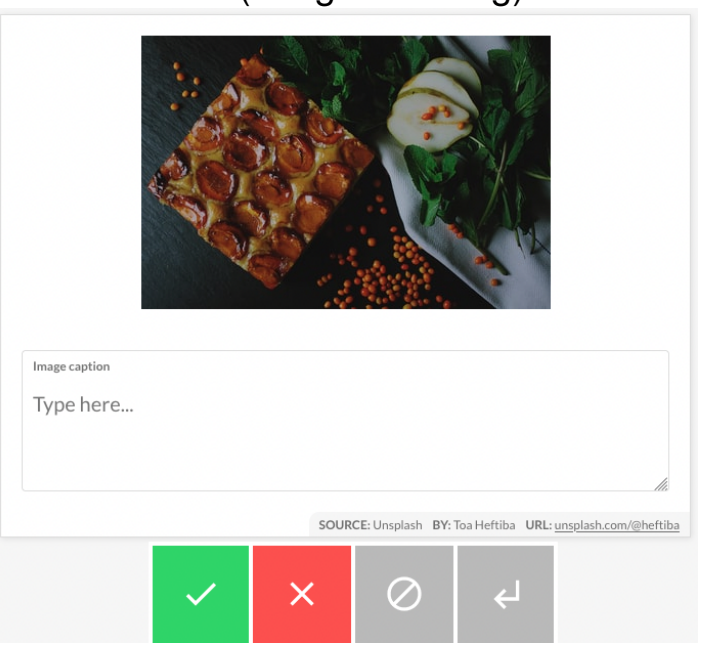
Span Annotation (NER)



Span Annotation (Entity Relations)



Freeform Text (Image Labeling)



Data Annotation

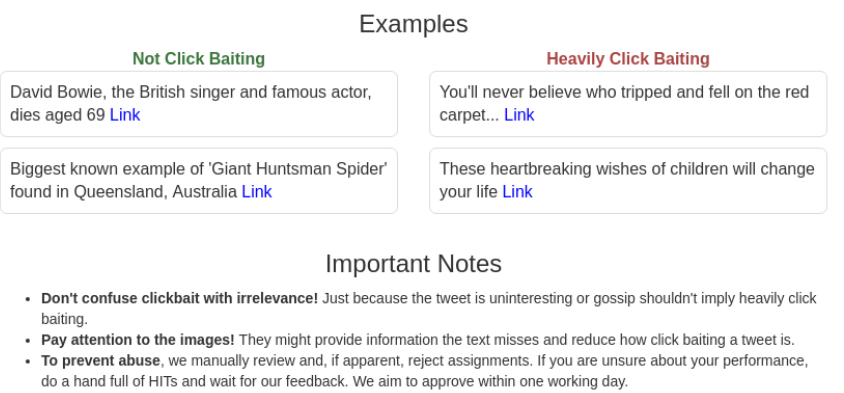
Annotation Schemes: Guidelines

Annotation guidelines are the instructions given to the annotators.
Elements of annotation guidelines:

1. Definitions of task and the phenomena.
2. Definitions of annotation options (classes, . . .).
3. Typical examples.
4. Edge cases: How to annotate atypical examples.

Example 1: Penn Treebank guideline for grammar annotation (318 pages). [Bies 1995]

Example 2: Clickbait in microblogs



Data Annotation

Annotation Schemes: Disagreement [Sandri, 2023]

Disagreement: Annotators make different decisions.

Causes of disagreement:

- Carelessness because of low pay, no consequences, high volume, unclear tasks
- Ambiguity (Users misunderstand the content, because of metaphors, irony, rhetorical moves, word plays, citations)

  `Who knew a side effect of COVID would be gross incompetence.`
- Missing context

  `Dude this guy is serious? And trump retweeted this?????? Please anonymous take them out`
- Subjectivity disagreement due to the annotators' identity, beliefs and background

  `#DemocratsAreDestroyingAmerica #Black- LivesMatter is a terrorist organization`

Data Annotation

Annotation Schemes: Disagreement

Dealing with disagreement:

- Vote aggregation (3, 5, . . . votes).
  - Collect multiple annotations for each example and aggregate them (wisdom of the crowds). Typical are three or five annotators.
  - Works well for classification, difficult for span or freeform text.

Means of vote aggregation:

	Data	Use Case
Majority/Mode	nominal*	Select the class with the most votes.
Mean	interval	Select the class closest to the average.
Median	ordinal	Select the class in the middle after ordering.
Minority/Threshold	binary	n positive votes = positive example.

*What happens when there are as many classes as annotators?

Data Annotation

Annotation Schemes: Disagreement

Dealing with disagreement:

- Vote aggregation (3, 5, . . . votes).
- Review (2 votes).
  - A layperson annotates, an expert reviews and corrects.
  - When annotations are labor intensive (span annotation).
  - When there are many difficult edge cases.
- LLM augmentation.
  - LLM cast the tie when two annotators disagree.
  - LLM decides when an expert needs to review.
- Learning with disagreement.
- Develop a more prescriptive schema. [Röttger, 2022]

Data Annotation

Annotation Schemes: Disagreement

Dealing with careless annotators:

- Evaluate and filter.
  - Check instances.
    Add some clear and easy examples with known annotations.
  - Attention checks.
    `Raise your hand if you still pay attention`
  - Dwell time.
  - Agreement with other annotators.

	Status	Annotations	Checks	Total Time	
	Approved	5 4 5 5 3 4 6 109 116 3 3 3	116 109	1.13 min	expand approve reject
	Approved	5 5 6 5 482 5 6 4 8 71 4 8	482 71	1.23 min	expand approve reject
	Approved	5 4 6 5 7 7 8 5 98 82 6 5	98 82	39 s	expand approve reject

Answer	Time	Text	Media
5	4.08 s	If you can't take the heat... Link	
4	2.88 s	ICE agent shoots and wounds man during arrest attempt: Link	

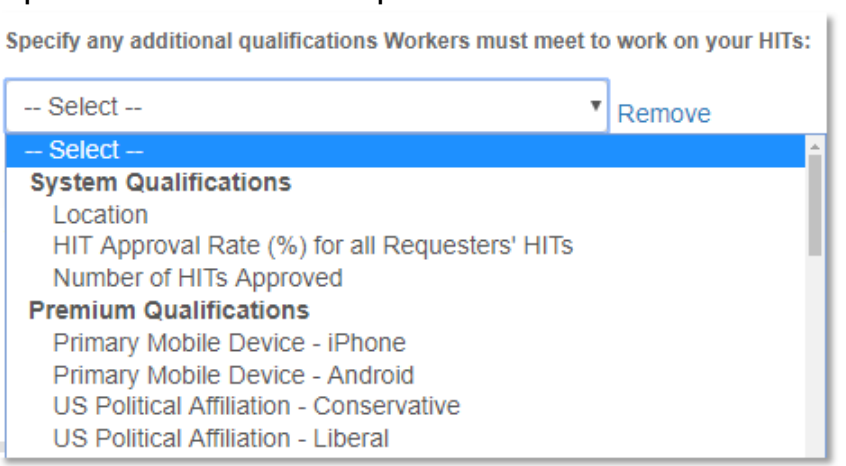
Data Annotation

Annotation Schemes: Disagreement

Dealing with careless annotators:

- Evaluate and filter.
- Make recruitment more restrictive.
  - Require more experience, more qualifications, . . .
  - For subjective tasks: restrictive criteria (area, language skills) might reduce diversity and add biases.

Options for annotator qualifications on AMT



Data Annotation

Annotator Agreement: Observed Agreement

Annotation quality is evaluated via annotator agreement:
A low agreement indicates that annotations differ by annotator.

Idea: Measure the ratio of examples where the annotators agree.

The observed agreement A_{obs} is the percentage of examples i where all annotators independently agree.

$$\text{agr}_i = \begin{cases} 1 \text{ if same category assigned} \\ 0 \text{ else} \end{cases}$$

$$A_o = \frac{1}{\mathbf{i}} \sum \text{agr}_i$$

Annotations for $k \in \{0, 1\}$

	i	**Annotator** c		agr_i
		c_1	c_2	
Category	1	1	1	1
	2	0	1	0
	3	1	1	1
	4	0	1	0
	5	0	0	1
$A_o = 0.6$				

Data Annotation

Annotator Agreement: Observed Agreement

Annotation quality is evaluated via annotator agreement:
A low agreement indicates that annotations differ by annotator.

Idea: Measure the ratio of examples where the annotators agree.

Problem: Observed agreement is not corrected for chance.

What happens if annotators chose randomly?

Case 1:
Annotators chose 0 in 50% of cases and 1 in 50%.
The overlap agreement will be 0.5.

Case 2:
Annotators chose 0 in 10% of cases and 1 in 90%.
The overlap agreement will be 0.82.

Annotations for $k \in \{0, 1\}$

	i	**Annotator** c		agr$_i$
		c_1	c_2	
Category	1	1	1	1
	2	0	1	0
	3	1	1	1
	4	0	1	0
	5	0	0	1
	$A_o = 0.6$			

Data Annotation

Annotator Agreement: Observed Agreement

Annotation quality is evaluated via annotator agreement:
A low agreement indicates that annotations differ by annotator.

Idea: Measure the ratio of examples where the annotators agree.

Problem: Observed agreement
is not corrected for chance.

- Reference value (random annotation) is different for each schema and task.
- Schemas with fewer classes will have a higher agreement.

$\rightarrow$ Changes to the task are hard to evaluate

Annotations for $k \in \{0, 1\}$

	i	**Annotator** c		agr_i
		c_1	c_2	
Category	1	1	1	1
	2	0	1	0
	3	1	1	1
	4	0	1	0
	5	0	0	1
$A_o = 0.6$				

Data Annotation

Annotator Agreement: Cohen's κ [Artstein, 2008]

Idea: Measure by how much the observed agreement A_o agreement is above the agreement A_e expected by chance.

$$\kappa = \frac{\text{Observed above chance}}{\text{Possible above chance}} = \frac{A_o - A_e}{1 - A_e}$$

Estimating A_e:

- Cohen's κ assumes that each annotator has his own prior distribution (bias).

$$A_e = \sum_k P(k|c_1) \cdot P(k|c_2)$$

- The prior distributions are estimated from the observations: The percentage of examples i annotated with category k by annotator c_j

$$P(k|c_j) = \frac{\mathbf{n}_{c_j k}}{\mathbf{i}}$$

Data Annotation

Annotator Agreement: Cohen's κ [Artstein, 2008]

Idea: Measure by how much the observed agreement A_o agreement is above the agreement A_e expected by chance.

$$\kappa = \frac{\text{Observed above chance}}{\text{Possible above chance}} = \frac{A_o - A_e}{1 - A_e}$$

Estimating A_e from observations:

$$\begin{aligned} A_e &= \sum_k P(k|c_1) \cdot P(k|c_2) \\ &= P(0|c_1) \cdot P(0|c_2) + P(1|c_1) \cdot P(1|c_2) \\ &= \tfrac{3}{5} \cdot \tfrac{1}{5} + \tfrac{2}{5} \cdot \tfrac{4}{5} \\ &= 0.6 \cdot 0.2 + 0.4 \cdot 0.8 = 0.44 \end{aligned}$$

Estimating κ with chance correction:

$$\kappa = \frac{A_o - A_e}{1 - A_e} = \frac{0.6 - 0.44}{1 - 0.44} = 0.29$$

Annotations for $k \in \{0, 1\}$

	i	**Annotator** c		agr$_i$
		c_1	c_2	
Category	1	1	1	1
	2	0	1	0
	3	1	1	1
	4	0	1	0
	5	0	0	1
		$A_o = 0.6$		κ = 0.29

Data Annotation

Annotator Agreement: Fleiss's κ [Artstein, 2008]

Problem: Cohen's κ scales poorly to multiple (3+) annotators.

1. The A_o calculation ignores partial agreement.
2. The A_e calculation expects all annotators to annotate all examples.

Fleiss's κ generalizes the κ to multiple (3+) annotators.

- agr_i is the ratio of pairs of annotators that agree.
  $c(c-1)$ is the number of possible 2-combinations of c. c is the number of annotations per example, not annotators. $\mathbf{n}_{ik}$ is the number of times category k is assigned to example i.

$$\text{agr}_i = \frac{1}{\mathbf{c}(\mathbf{c}-1)} \sum_k \mathbf{n}_{ik}(\mathbf{n}_{ik}-1)$$

- The chance agreement A_e is generalized using the ratio of actual vs. possible assignments of a category k.

$$A_e = \sum_k P(k|c_1) \cdot P(k|c_2) = \sum_k P(k)^2 \qquad P(k) = \frac{1}{\mathbf{c} \cdot \mathbf{i}} \sum_i \mathbf{n}_{ik}$$

Annotations for $k \in \{0, 1\}$

	i	**Annotator** c				agr_i
		c_1	c_2	c_3	c_4	
Category	1	1	1	–	1	1
	2	0	1	1	–	2/6
	3	1	1	–	0	2/6
	4	0	1	0	–	2/6
	5	0	0	0	–	1

$$P_0 = 1/15 \cdot 7 = 0.46$$
$$P_1 = 1/15 \cdot 8 = 0.53$$
$$A_o = 0.6 \qquad A_e = 0.49$$
$$\kappa = 0.22$$

Remarks:

- The κ measures assume that the categories are independent.
- Be mindful when interpreting κ values: Increasing classes and annotator count lowers the agreement. Subjective topics often score lower agreement.
- There are other agreement measures, like Scott's π or S that estimate the chance agreement differently.
- For ordinal or interval data, correlation coefficients (Pearson ρ, Spearman ρ, Kendall's τ) can be better suited.
- Arstein et al. note:

  However, it is important to keep in mind that achieving good agreement cannot ensure validity: Two observers of the same event may well share the same prejudice while still being objectively wrong.

Data Annotation

Non-technical Aspects

There are ethical and legal considerations when working with human-created data.
If in doubt: consult the ethics board before starting an annotation project.

- Personal data
  Annotations can count as or contain personal data.
  Anonymize and request permission for use.

- Legal
  Be mindful of data collection and distribution laws and licenses.

- Harmful text
  Make annotators aware of potential harm beforehand.

- Working Conditions
  Provide compensation. Collect and implement feedback.