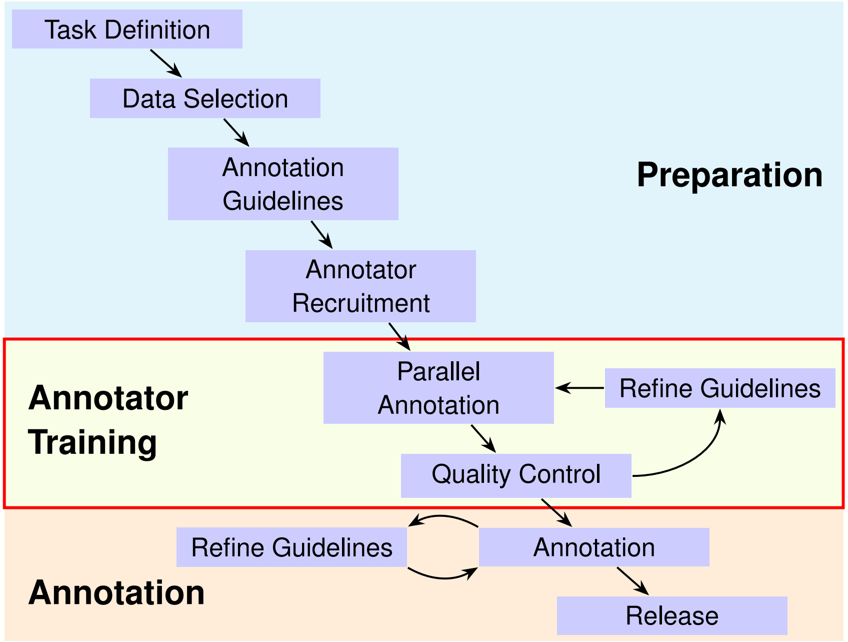
Lecture 5: Annotation and Evaluation

First-Year Project 3:
Natural Language Processing

Christian Hardmeier

3 May 2022

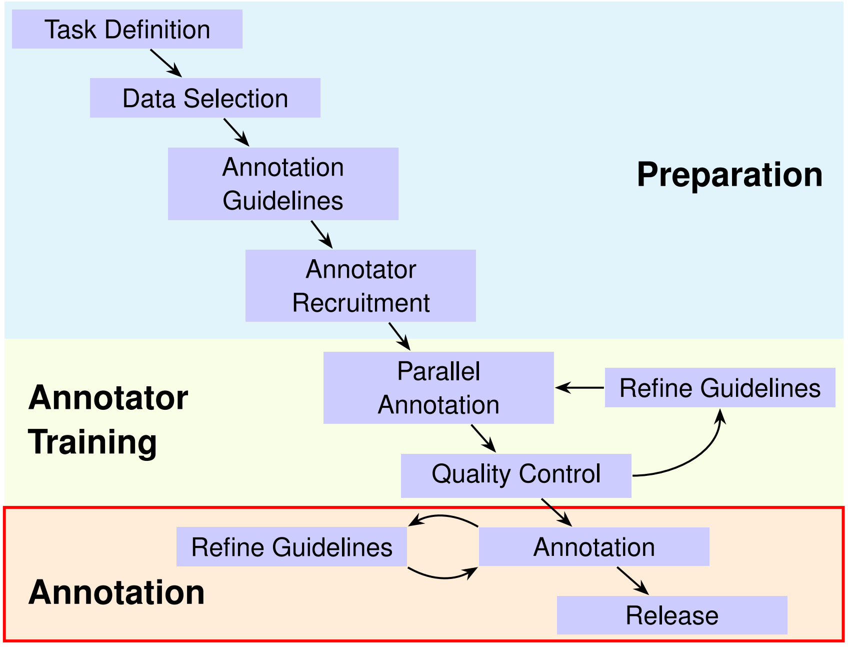
IT UNIVERSITY OF COPENHAGEN



Annotator training

- Let two or more annotators work in parallel.
- Discuss difficult cases frequently in the beginning.
- After completing a small portion of the data, compute *inter-annotator agreement* (IAA) and discuss differences.
- Refine guidelines where necessary. Add examples to guidelines.
- Repeat until no further improvement is seen, and think about whether the IAA is satisfactory.

IT UNIVERSITY OF COPENHAGEN

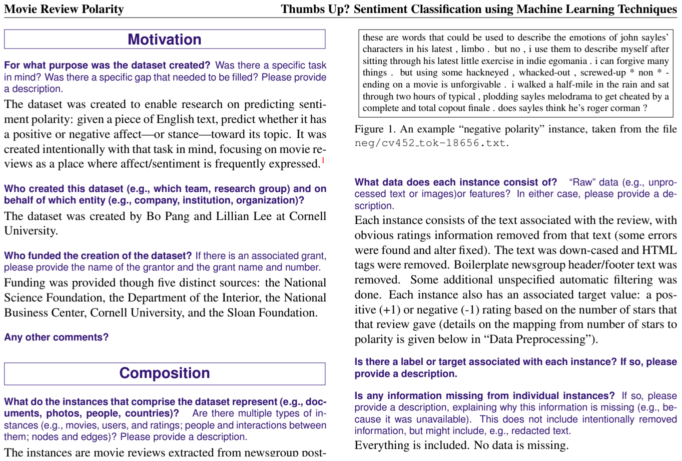


Annotation

- We usually can't afford double annotation for the whole dataset.
- No further IAA calculations are possible.
- Difficult examples will still pop up!
- Discuss/adjudicate and refine guidelines if necessary.
- Update previously annotated parts when guidelines change!

Release

- Do you have the necessary rights?
- Licencing
- Long-term storage (e.g., `https://lindat.mff.cuni.cz/`)
- Ethical aspects: Datasheets for Datasets `https://arxiv.org/abs/1803.09010`
  - Motivation
  - Composition
  - Collection process
  - Recommended uses



Movie Review Polarity **Thumbs Up? Sentiment Classification using Machine Learning Techniques**

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity: given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. It was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided though five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup post-

> these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and alter fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

(Gebru et al, 2018)

Annotation Quality Control

Intra-Annotator Agreement

- Let annotators reannotate a small portion after a while.
  - Spaced out, in different order.
  - Mixed with new examples.
  - Or after a period of time has passed.
- Check the *consistency* of annotations.
- Reasons for low intra-annotator agreement:
  - Ill-defined guidelines
  - Priming effects
  - Insufficient information to make decisions

Inter-Annotator Agreement

- Let all coders annotate a common portion of the dataset.
- Check for discrepancies in the annotations.
- Reasons for low inter-annotator agreement:
  - Incomplete or ambiguous guidelines
  - Diverging interpretations of the guidelines
  - Different annotator background (expertise, language proficiency)
  - Different understanding of the task
  - Or any of the reasons mentioned before
- Inter-Annotator Agreement gives an indication of how well-defined and reproducible the task is.

Observed Agreement

$$A_o = \frac{\text{\# matches}}{\text{\# total items}}$$

- Often used for **intra**-annotator agreement.
- Basis for inter-annotator metrics, but not sufficient on its own.
- Agreement between coders might be due to chance!
- Not comparable across studies.
- Higher chance agreement is likely
  - if there are few categories to choose from, or
  - if the categories are unbalanced.

Chance-corrected Agreement

$$\text{Agreement} = \frac{A_o - A_e}{1 - A_e}$$

- Estimate A_e, the probability of chance agreement based on the task design.
- Subtract this from the observed agreement:
  - $1 - A_e$ is the maximum attainable agreement above chance level.
  - $A_o - A_e$ is the actually observed agreement above chance level.
- Methods differ in how A_e is estimated.

Estimating chance agreement

- Notation
  - $K = \{k_1, k_2, \ldots, k_n\}$: Different categories
  - C_1, C_2: Labels assigned by two coders
  - $\#(C_i, k_j)$: Number of times coder i has assigned label k_j
- Most methods assume *independence of coders*.

  $$p(C_1 = k, C_2 = k) = p(C_1 = k)p(C_2 = k) \text{ for all } k \in K$$

- Expected agreement is the probability of agreeing on *any* label:

  $$A_e = \sum_{k \in K} p(C_1 = k)p(C_2 = k)$$

- Presented for two coders – all scores can be generalised.

  Artstein and Poesio, *Computational Linguistics* 34 (2008) 4, 555–596.

Assumptions: S, π, κ

S If coders were operating by chance alone,
we'd get a *uniform* distribution:

$$p(C_1 = k_i) = p(C_2 = k_l) \text{ for any two categories } k_i, k_j$$

π If coders were operating by chance alone,
we'd get *the same* distribution for each coder.

$$p(C_1 = k) = p(C_2 = k) \text{ for any category } k$$

κ If coders were operating by chance alone,
we'd get *a separate* distribution for each coder.

S coefficient

- Assumption: All categories are equally likely.
- Chance labelling is a draw from a uniform distribution:

$$\begin{aligned} A_e &= \sum_{i=1}^{|K|} p(C_1 = k_i)p(C_2 = k_i) \\ &= \sum_{i=1}^{|K|} \frac{1}{|K|} \cdot \frac{1}{|K|} = |K| \cdot \left[\frac{1}{|K|}\right]^2 \\ &= \frac{1}{|K|} \end{aligned}$$

- Can be artifically increased
  by simply adding more (useless) categories.

Scott's π

- Assumption: All coders have the same preferences.
- Chance labelling is a draw in proportion to the frequency of the labels in the corpus:

$$p(C_i = k) = \frac{\#(\bullet, k)}{\#(\bullet, \bullet)} = \frac{\#(\bullet, k)}{2N}$$

- Expected agreement:

$$\begin{aligned} A_e &= \sum_{k \in K} p(C_1 = k) p(C_2 = k) \\ &= \sum_{k \in K} \left[\frac{\#(\bullet, k)}{2N} \right]^2 \\ &= \frac{1}{4N^2} \sum_{k \in K} \#(\bullet, k)^2 \end{aligned}$$

Cohen's κ

- Each coder has their own preferences (individual annotator bias).
- Individual distributions estimated with relative frequencies:

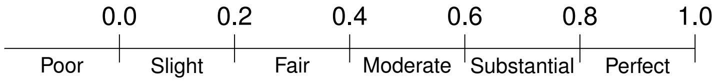
$$p(C_i = k) = \frac{\#(C_i, k)}{\#(C_i, \bullet)} = \frac{\#(C_i, k)}{N}$$

- Expected agreement:

$$\begin{aligned} A_e &= \sum_{k \in K} p(C_1 = k) p(C_2 = k) \\ &= \sum_{k \in K} \frac{\#(C_1, k)}{N} \cdot \frac{\#(C_2, k)}{N} \\ &= \frac{1}{N^2} \sum_{k \in K} \#(C_1, k) \#(C_2, k) \end{aligned}$$

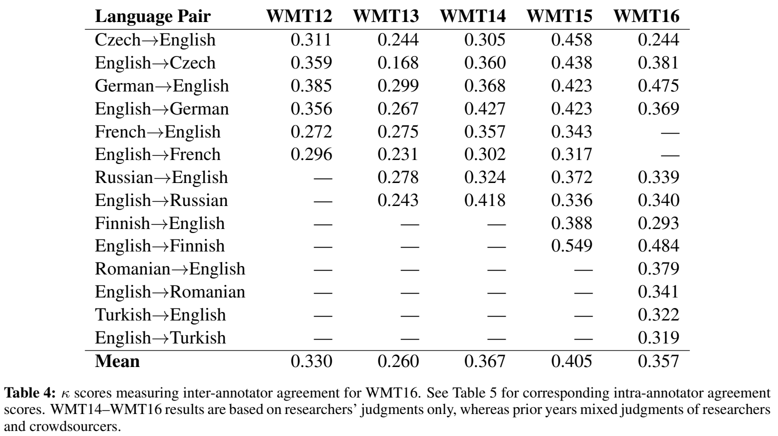
What is good inter-annotator agreement?

- "[D]eciding what counts as an adequate level of agreement for a specific purpose is still little more than a black art" (Artstein and Poesio, 2008).
- Rules of thumb (Landis and Koch, *Biometrics* 1977):



- Difficult to use: Hard to know what to expect, or what's the minimum to be useful.

Comparing to similar annotations



Language Pair	WMT12	WMT13	WMT14	WMT15	WMT16
Czech→English	0.311	0.244	0.305	0.458	0.244
English→Czech	0.359	0.168	0.360	0.438	0.381
German→English	0.385	0.299	0.368	0.423	0.475
English→German	0.356	0.267	0.427	0.423	0.369
French→English	0.272	0.275	0.357	0.343	—
English→French	0.296	0.231	0.302	0.317	—
Russian→English	—	0.278	0.324	0.372	0.339
English→Russian	—	0.243	0.418	0.336	0.340
Finnish→English	—	—	—	0.388	0.293
English→Finnish	—	—	—	0.549	0.484
Romanian→English	—	—	—	—	0.379
English→Romanian	—	—	—	—	0.341
Turkish→English	—	—	—	—	0.322
English→Turkish	—	—	—	—	0.319
Mean	0.330	0.260	0.367	0.405	0.357

Table 4: κ scores measuring inter-annotator agreement for WMT16. See Table 5 for corresponding intra-annotator agreement scores. WMT14–WMT16 results are based on researchers' judgments only, whereas prior years mixed judgments of researchers and crowdsourcers.

(Bojar et al., WMT 2016)

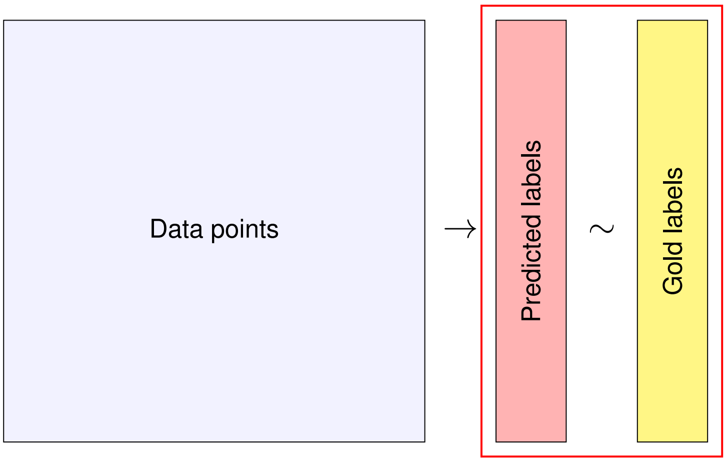
```
scipy.stats.percentileofscore(score_list, score)
```

Practicalities and Further Reading

- Inter-annotator agreement metrics are implemented in `nltk.metrics.agreement.AnnotationTask`
- You will use the multi-coder generalisations of the metrics.
- For more details, consult this paper:
  Ron Artstein and Massimo Poesio: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34 (2008) 4, 555–596.
  `https://www.aclweb.org/anthology/J08-4004.pdf`

Evaluating Classifiers

Supervised Classification



Example: Part of speech tagging

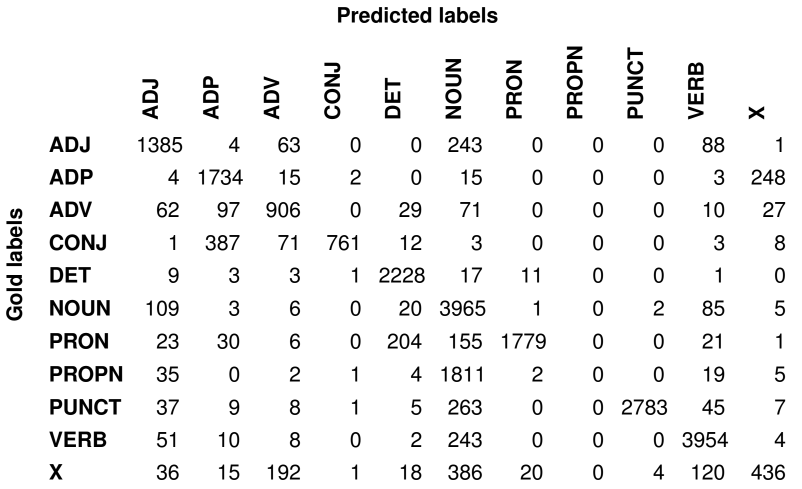
- Parts of speech: Word classes determining syntactic properties.
- 11 classes (reduced from original annotation).
- Adjectives, prepositions, adverbs, conjunctions, determiners, nouns, pronouns, proper names, punctuation, verbs and others.
- Tagged with NLTK and evaluated against manual annotations.

Part-of-speech tagging

The old man the boat.

	Predicted	**Gold**
The	DET	DET
old	ADJ	NOUN
man	NOUN	VERB
the	DET	DET
boat	NOUN	NOUN
.	PUNCT	PUNCT

Confusion matrix

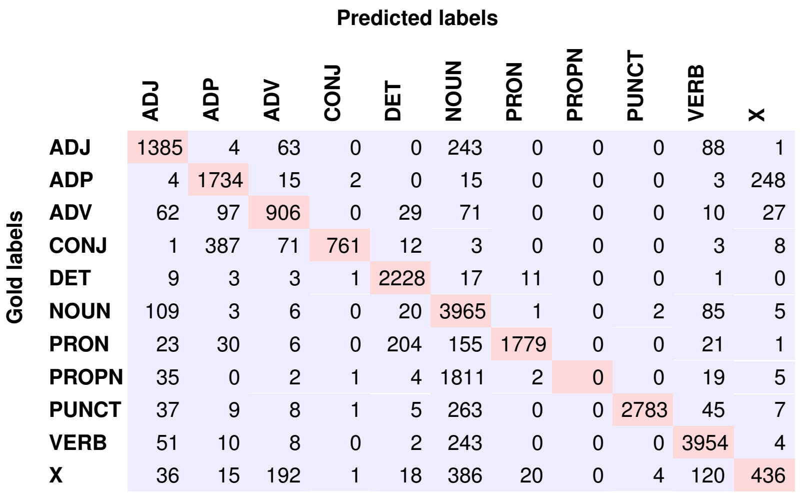


		Predicted labels										
		ADJ	ADP	ADV	CONJ	DET	NOUN	PRON	PROPN	PUNCT	VERB	X
Gold labels	ADJ	1385	4	63	0	0	243	0	0	0	88	1
	ADP	4	1734	15	2	0	15	0	0	0	3	248
	ADV	62	97	906	0	29	71	0	0	0	10	27
	CONJ	1	387	71	761	12	3	0	0	0	3	8
	DET	9	3	3	1	2228	17	11	0	0	1	0
	NOUN	109	3	6	0	20	3965	1	0	2	85	5
	PRON	23	30	6	0	204	155	1779	0	0	21	1
	PROPN	35	0	2	1	4	1811	2	0	0	19	5
	PUNCT	37	9	8	1	5	263	0	0	2783	45	7
	VERB	51	10	8	0	2	243	0	0	0	3954	4
	X	36	15	192	1	18	386	20	0	4	120	436

You've already seen this!

Predicted → True ↓	Can cer	NC
Cancer	TP	FN
Non Cancer	FP	TN

Confusion matrix



		Predicted labels										
		ADJ	ADP	ADV	CONJ	DET	NOUN	PRON	PROPN	PUNCT	VERB	X
Gold labels	ADJ	1385	4	63	0	0	243	0	0	0	88	1
	ADP	4	1734	15	2	0	15	0	0	0	3	248
	ADV	62	97	906	0	29	71	0	0	0	10	27
	CONJ	1	387	71	761	12	3	0	0	0	3	8
	DET	9	3	3	1	2228	17	11	0	0	1	0
	NOUN	109	3	6	0	20	3965	1	0	2	85	5
	PRON	23	30	6	0	204	155	1779	0	0	21	1
	PROPN	35	0	2	1	4	1811	2	0	0	19	5
	PUNCT	37	9	8	1	5	263	0	0	2783	45	7
	VERB	51	10	8	0	2	243	0	0	0	3954	4
	X	36	15	192	1	18	386	20	0	4	120	436

Confusion matrix

		Predicted labels										
		ADJ	ADP	ADV	CONJ	DET	NOUN	PRON	PROPN	PUNCT	VERB	X
Gold labels	**ADJ**	1385	4	63	0	0	243	0	0	0	88	1
	ADP	4	1734	15	2	0	15	0	0	0	3	248
	ADV	62	97	906	0	29	71	0	0	0	10	27
	CONJ	1	387	71	761	12	3	0	0	0	3	8
	DET	9	3	3	1	2228	17	11	0	0	1	0
	NOUN	109	3	6	0	20	3965	1	0	2	85	5
	PRON	23	30	6	0	204	155	1779	0	0	21	1
	PROPN	35	0	2	1	4	1811	2	0	0	19	5
	PUNCT	37	9	8	1	5	263	0	0	2783	45	7
	VERB	51	10	8	0	2	243	0	0	0	3954	4
	X	36	15	192	1	18	386	20	0	4	120	436

Confusion matrix

		Predicted labels										
		ADJ	ADP	ADV	CONJ	DET	NOUN	PRON	PROPN	PUNCT	VERB	X
Gold labels	**ADJ**	1385	4	63	0	0	243	0	0	0	88	1
	ADP	4	1734	15	2	0	15	0	0	0	3	248
	ADV	62	97	906	0	29	71	0	0	0	10	27
	CONJ	1	387	71	761	12	3	0	0	0	3	8
	DET	9	3	3	1	2228	17	11	0	0	1	0
	NOUN	109	3	6	0	20	3965	1	0	2	85	5
	PRON	23	30	6	0	204	155	1779	0	0	21	1
	PROPN	35	0	2	1	4	1811	2	0	0	19	5
	PUNCT	37	9	8	1	5	263	0	0	2783	45	7
	VERB	51	10	8	0	2	243	0	0	0	3954	4
	X	36	15	192	1	18	386	20	0	4	120	436

Accuracy

- Accuracy is the proportion of elements classified correctly.

$$\text{Accuracy} = \frac{\text{sum of diagonal}}{\text{total sum}}$$

- Accuracy is the simplest metric for classification.
- It can be misleading in unbalanced datasets!
- It provides no information about individual classes.

Precision and Recall

- Class-specific metrics:
  Based on *single rows and columns* of confusion matrix.
- Common metrics in NLP come from *Information Retrieval*.
- In a database search, we want to find
  - all relevant results, and
  - no distracting irrelevant results.

Precision and Recall

- **Precision:**
  Out of the examples we **predicted to be** in a certain class, how many of them are correct?
  (How many irrelevant results did we find?)

$$\text{Precision} = \frac{\text{single diagonal element}}{\text{sum of a single column}}$$

- **Recall:**
  Out of the examples **that actually belong** to a certain class, how many of them did we find?
  (Did we actually find what we were looking for?)

$$\text{Recall} = \frac{\text{single diagonal element}}{\text{sum of a single row}}$$

Confusion matrix

Predicted labels

Gold labels

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**NOUN**	**PRON**	**PROPN**	**PUNCT**	**VERB**	**X**
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	243	0	0	0	88	1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	15	0	0	0	3	248
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	9	71	0	0	0	10	27
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	3	0	0	0	3	8
DET	9	3	3	1	2228	17	11	0	0	1	0
NOUN	109	3	6	0	20	3965	1	0	2	85	5
PRON	23	30	6	0	204	155	1779	0	0	21	1
PROPN	35	0	2	1	4	1811	2	0	0	19	5
PUNCT	37	9	8	1	5	263	0	0	2783	45	7
VERB	51	10	8	0	2	243	0	0	0	3954	4
X	36	15	192	1	18	386	20	0	4	120	436

$$\text{Precision} = \frac{\text{diagonal element}}{\text{column sum}}$$

$$= \frac{3965}{7172} = 0.553$$

Confusion matrix

$$\text{Recall} = \frac{\text{diagonal element}}{\text{row sum}}$$

$$= \frac{3965}{4196} = 0.945$$

Predicted labels

Gold labels

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	NOUN	PRON	PROPN	PUNCT	VERB	X
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	243	0	0	0	88	1
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	15	0	0	0	3	248
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	71	0	0	0	10	27
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3	0	0	0	3	8
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2228	17	11	0	0	1	0
NOUN	109	3	6	0	20	3965	1	0	2	85	5
PRON	23	30	6	0	204	155	1779	0	0	21	1
PROPN	35	0	2	1	4	1811	2	0	0	19	5
PUNCT	37	9	8	1	5	263	0	0	2783	45	7
VERB	51	10	8	0	2	243	0	0	0	3954	4
X	36	15	192	1	18	386	20	0	4	120	436

Precision/Recall vs. Sensitivity/Specificity

- Precision and Recall focus on the true positives in the context of *what was found* and *what should have been found*.
- Sensitivity and Specificity focus on *correct identification of positives and negatives*.
- Sensitivity is just another name for Recall, but *Specificity and Precision are different.*
- Se and Sp are like “positive and negative Recall”.

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{R} = \text{Se} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Precision and Recall can be gamed!

- **100% Precision** (good chance): Return only the one example you’re most certain of!
- **100% Recall** (guaranteed): Return the entire dataset.
- But *you can’t game both of them at the same time!*

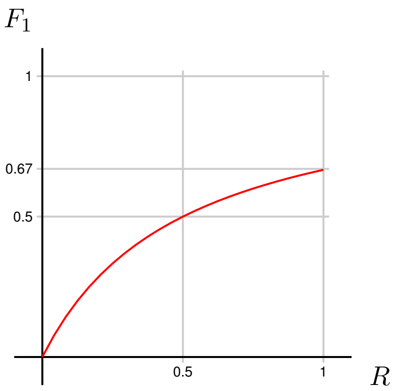
F-score (or F-measure)

- We often want to summarise Precision and Recall in a single number.
- **F-score** (or F_1) is the *harmonic mean* of Precision and Recall.

$$F = 2 \cdot \frac{P \cdot R}{P + R}$$

- If P and R are equal, then F is the same.
- If they are different, then F is closer to the *lower* of them.
- By maximising F-score, we emphasise balanced P and R.
- F-score can be generalised with a parameter to control the balance.

F-score behaviour



Precision fixed at 0.5

Micro-averaging

- Precision and Recall are per class, but sometimes we'd like to have *one single number* to characterise our performance.
- Accuracy is a single number, but is problematic with unbalanced data.
- *Micro-averaged Precision, Recall and F-score*:
  Add the counts of all classes,
  then compute Precision, Recall and F-score.
- If each example has only one label,
  this is *the same as accuracy*.

Macro-averaged scores

- *Macro-averaged Precision, Recall and F-score*:
  Compute P, R and F for each class,
  then take the arithmetic mean.

$$P_{\text{macro}} = \frac{1}{|K|} \sum_{k \in K} P_k \qquad R_{\text{macro}} = \frac{1}{|K|} \sum_{k \in K} R_k$$

- Macro-averaging is sensitive to outlier classes!
  (can enforce balance, but also cause problems)
- **Macro-averaged Recall** is least sensitive to imbalance.

Exercise

Continue working on data annotation and evaluation of inter-annotator agreement.