Lecture 4:
Annotation and Evaluation

First-Year Project 3:
Natural Language Processing

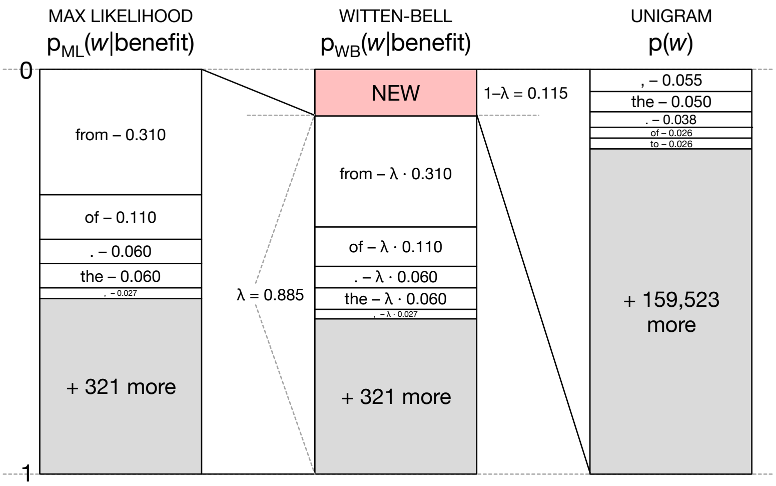
Christian Hardmeier

28 April 2022

N-gram models: Wrapping up

IT UNIVERSITY OF COPENHAGEN

Backing off to lower-order models



IT UNIVERSITY OF COPENHAGEN

Witten-Bell smoothing

- Witten-Bell smoothing treats "seeing a new word" as an event in its own right, so we can model its probability explicitly.
- In training, the "new word" event occurs as many times as we have different words.

$$p(\text{new}|w_1, w_2) = 1 - \lambda = \frac{\#\text{types}(w_1 w_2 \bullet)}{\#\text{tokens}(w_1 w_2 \bullet) + \#\text{types}(w_1 w_2 \bullet)}$$

- In the interpolated model, this probability corresponds to the weight $(1 - \lambda)$:

$$\begin{aligned} p'(w_{k+1}|w_1, \ldots, w_k) = {} & \lambda\, p_{\text{ML}}(w_{k+1}|w_1, \ldots, w_k) \\ & + (1 - \lambda)\, p'(w_{k+1}|w_2, \ldots, w_k) \end{aligned}$$

Absolute discounting

- Subtract a constant discount $(0 < d < 1)$ from the nominator of the counts:

$$p(w_3|w_1, w_2) = \frac{\#\text{tokens}(w_1 w_2 w_3) - d}{\#\text{tokens}(w_1 w_2 \bullet)}$$

- This will have a large effect on small counts, but a small effect on large counts.
- d can be estimated, e.g. from held-out data.

Absolute discounting

$r = f_{\text{MLE}}$	f_{emp}	$f_{\text{add-1}}$
0	0.000027	0.000137
1	0.448	0.000274
2	1.25	0.000411
3	2.24	0.000548
4	3.23	0.000685
5	4.21	0.000822
6	5.23	0.000959
7	6.21	0.00109
8	7.21	0.00123
9	8.26	0.00137

`http://www.cs.cornell.edu/courses/cs6740/2008fa/lectures/smoothing2+backoff.pdf`
Data from Church and Gale, *Computer Speech and Language* 5 (1991) 19–54

Kneser-Ney smoothing

I can't see without my reading ______

- The continuation *glasses* is far more likely than *Kong*.
- But in an English new corpus, *Kong* is more frequent than *glasses*.
- *Kong* only occurs in specific contexts (mostly *Hong Kong*).
  - We only expect to see *Kong* in a bigram we know.
  - We *don't* expect *Kong* to occur in a context we don't know.
  - Contexts we don't know correspond to *backoff situations*.

(Improved) Kneser-Ney smoothing

- Kneser-Ney smoothing uses different distributions for the *higher-order* and the *backoff* distributions.
- For the higher-order distribution, it uses absolute discounting.
  - Discounts estimated separately for counts 1 and 2.
- Backoff distributions are estimated based on the *number of contexts* a word occurs in:

$$p_{\mathsf{cont}}(w) = \frac{\#\text{types}(\bullet w)}{\#\text{types}(\bullet\bullet)}$$

  - If a word occurs in many contexts, it's *flexible* and may also occur in a new one.
  - If it only occurs in *restricted* contexts, we don't expect it suddenly to show up in a new one.

Smoothing methods

- **Laplace smoothing**
  - Avoids 0 probabilities, very easy to implement.
  - Performs worse than other methods.
- **(Improved) Kneser-Ney smoothing**
  - One of the best-performing smoothing methods for natural language.
  - Based on absolute discounting, with clever handling of backoff distribution.
  - Makes specific assumptions about the distribution of infrequent tokens that work well for natural language.
  - For sequences with few infrequent tokens, estimation may fail!
- **Witten-Bell smoothing**
  - Good method for sequences that don't meet the Kneser-Ney assumptions.
  - Uses the number of different continuations of an n-gram to estimate how likely yet another new continuation will be.

N-gram modelling tools

- NLTK
  - The n-gram library in `nltk` is a teaching tool.
  - You would *not* use it for real projects.
- KenLM – `https://kheafield.com/code/kenlm/`
  - Very fast and scalable implementation.
  - Only supports one smoothing method (Kneser-Ney).
  - Free software.
- SRILM – `http://www.speech.sri.com/projects/srilm/`
  - Very complete and well-documented package.
  - Supports many different methods and options.
  - Non-free, free of charge for many non-profit use cases.
  - Commercial use costs money.

Data Annotation

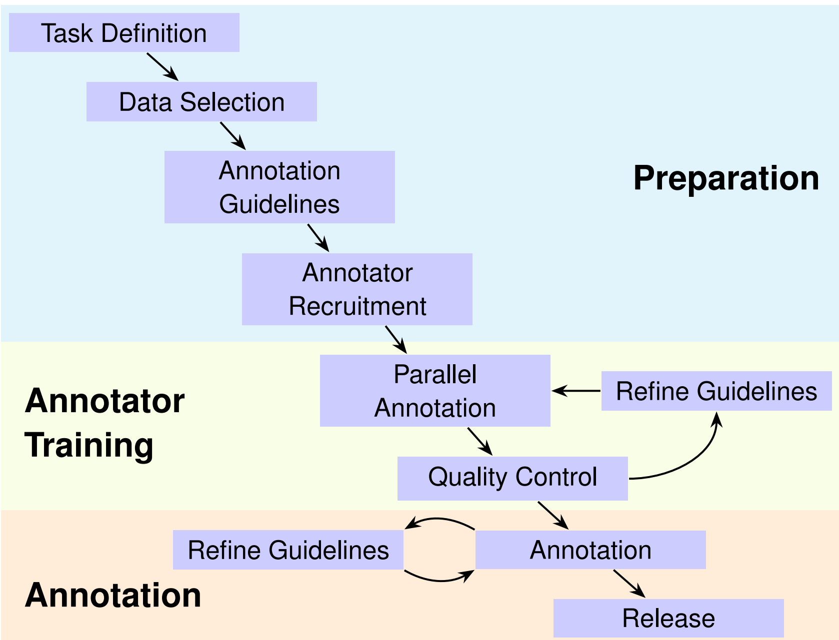
Encoding knowledge for machines

- Explicit rules and recipes

```
if(has long ears) then return "hare"
if(has trunk) then return "elephant"
```

  - Requires deep theoretical understanding.
  - Brittle, sensitive to incorrect assumptions.
  - Expensive, difficult to maintain.
- Unsupervised learning
  - Difficult to get the results you want.
  - Requires correct inductive bias in models.
  - Often used for *pre-training* in modern NLP.
- Data annotation
  - Manually enriching data with additional information.
  - Predominant method for knowledge encoding
    in machine learning scenarios.

Annotation Project Management



Task definition

- What will we use the data for, and how?
- Do our data sources cover the things we look for?
- Can we limit the task to make annotation easier?
  - Trade-off between rich annotation and consistency.
- What data/context do we need to make annotation decisions?

Data selection

Corpus size

- How much data do we need?
  - For training: ca. 1 example/model parameter
  - Less for evaluation
- How much data can we afford to get annotated?

Data source

- Availability of data
- Legality/licensing conditions (Redistribution?)
- Are the phenomena we're interested in frequent enough?

Sampling

- Random sampling
- Complete documents
- Oversampling

Annotation guidelines

What to annotate

- Trade-off between
  - high information content, and
  - categories that are easy to annotate consistently.
- Sources of information:
  - Previous annotation efforts
  - Theoretical literature (linguistics, social science, etc.!)
- Ideally: Simple, clearly answerable questions.

Annotation tools

- Ease of use
- Visualise all necessary information
- Data formats

Adequacy: Please rank the three translations according to *how adequately the translation of the last sentence reflects the meaning of the source, given the context.*

Fluency: Please rank the three translations according to *how fluent the last sentence is, in terms of grammaticality, naturalness and consistency, taking into account the context of the previous sentences.*

Table 2: Instructions to human annotators

(Wang, Hardmeier and Sennrich, NODALIDA 2021)

ParCorFull: A Parallel Corpus Annotated with Full Coreference

Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

Lapshinova-Koltunski, Ekaterina; Hardmeier, Christian and Krielke, Pauline, 2018, *ParCorFull: A Parallel Corpus Annotated with Full Coreference*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11372/LRT-2614.

Share:

LRT + Open Submissions

Authors	Lapshinova-Koltunski, Ekaterina ; Hardmeier, Christian ; Krielke, Pauline
Item identifier	http://hdl.handle.net/11372/LRT-2614
Project URL	https://github.com/chardmeier/parcor-full
Referenced by	http://www.lrec-conf.org/proceedings/lrec2018/summaries/941.html
Date issued	2018-05-08
Type	corpus, text
Size	158919 tokens
Language(s)	English , German

Annotator Recruitment

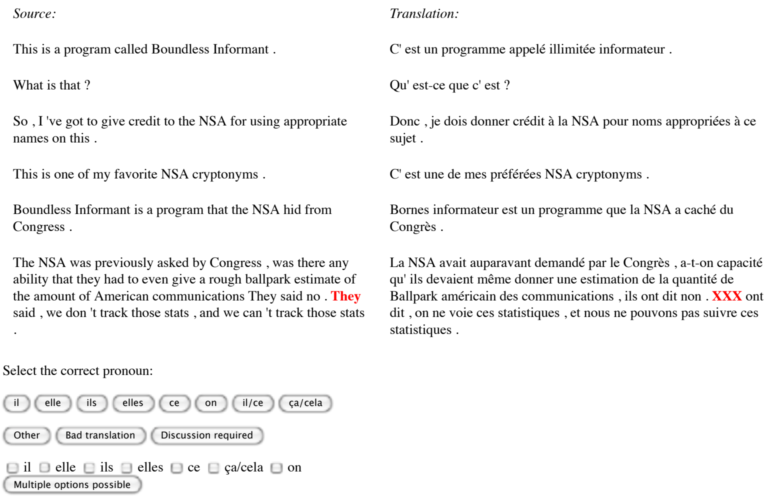
- Level of expertise required
  - Linguistic proficiency
  - Theoretical knowledge
- Where to recruit from?
  - Existing collaborators
  - Hiring people
  - Crowdsourcing
- Cost

IT UNIVERSITY OF COPENHAGEN

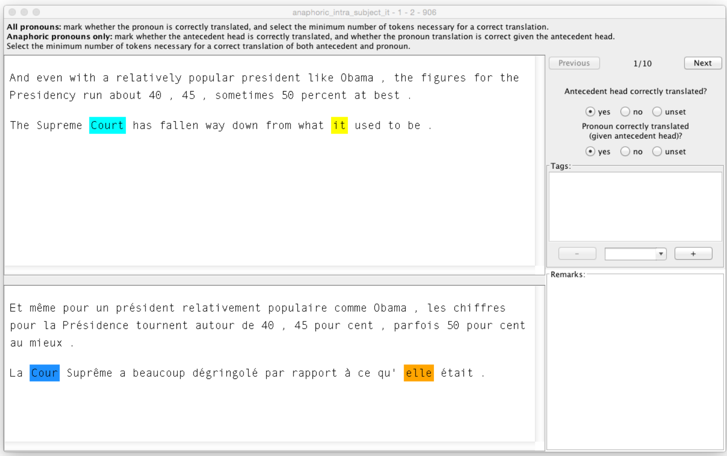
	N (N_m)	Age: range	mean	σ
English	42 (36)	22–70	37.1	11.3
French	42 (22)	18–55	30.9	10.0
German	31 (25)	18–55	31.9	10.9
Italian	43 (31)	18–48	29.7	8.3
Spanish	45 (27)	18–67	33.0	9.7
	203 (141)	18–70	32.4	10.2

Participants in crowdsourcing study
N: Total participants
N_m: Participants satisfying proficiency requirements

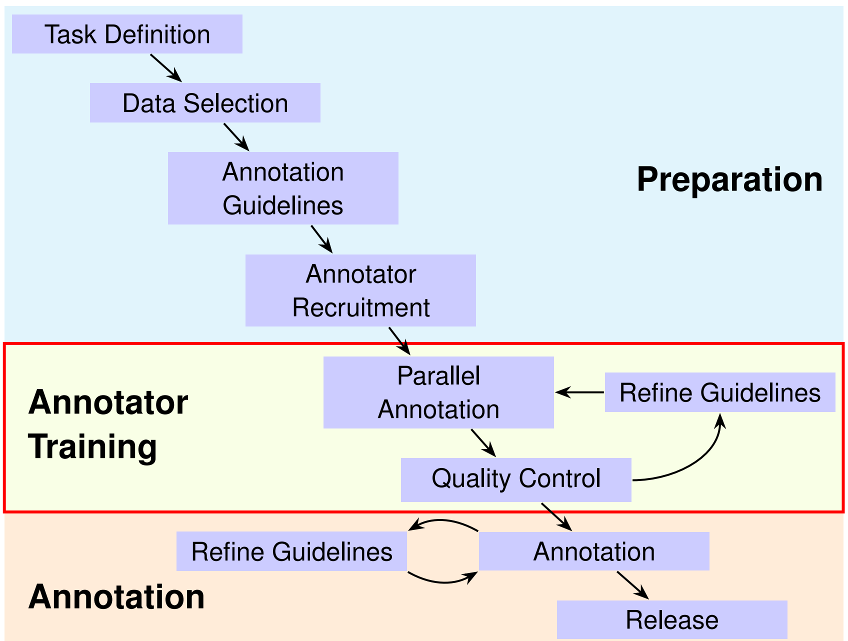
(Bevacqua, Loáiciga, Rohde and Hardmeier, 2021)



(Hardmeier, Nakov, Stymne, Tiedemann, Versley and Cettolo, DiscoMT 2015)

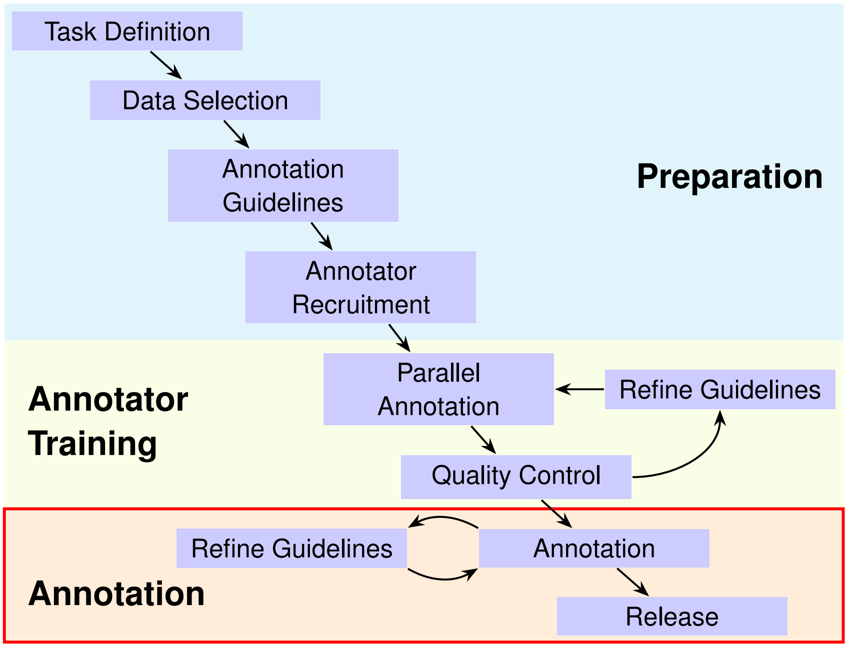


(Guillou and Hardmeier, EMNLP 2018)



Annotator training

- Let annotators work in parallel.
- Discuss difficult cases frequently in the beginning.
- After completing a small portion of the data,
  compute inter-annotator agreement 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.



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

IT UNIVERSITY OF COPENHAGEN

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*.

  $p(C_1 = k, C_2 = k) = p(C_1 = k)p(C_2 = k)$ 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:

$p(C_1 = k_i) = p(C_2 = k_l)$ 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.

$p(C_1 = k) = p(C_2 = k)$ 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:

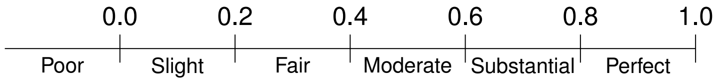
$$p(C_i = k) = \frac{\#(C_i, k)}{\#(C_i, \bullet)} = \frac{\#(C_i, k)}{N}$$

- 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`

Exercise

- Select one of your two TweetEval tasks (but not *emoji*).
- Read up on how the dataset was originally created.
- Use randomly selected IAA subsets.
- Let each member of your group annotate these instances, but
  - do *not* look at the labels in the dataset,
  - do *not* discuss while you annotate, and
  - follow your understanding of the original guidelines as closely as possible.
- Compute inter-annotator agreement between your annotations, with and without including the original labels.
- Discuss the cases you disagreed on and whether there are any patterns.