UNIVERSITÄT
LEIPZIG

Research Seminar WS 23/24

Active Learning for Corpus Quality

Maximilian Bley, Masterarbeit

Leipzig, 25.01.24

Gliederung

1. Einleitung
2. Related Work
3. Experiment 1: Annotationsschema und -richtlinien
4. Experiment 2: Active Learning (Ausblick)

1. Einleitung

Relevanz von Trainingsdatenqualität wird stärker diskutiert

⇒ Teilweise extrem viel nicht verwertbarer Text in Webkorpora (Kreutzer et. al (2022), Abadji et. al. (2022))

⇒ Evidenz, dass Qualität sich positiv auf die Performance auswirkt (Longpre et al. (2023))

⇒ Standard in der Preprocessing-Pipeline von LLMs (GPT, BLOOM, LLAMA, …)

1. Einleitung

Gap

- Methoden zur Filterung von großen Datensätzen sind oft **musterbasierte Heuristiken** und/oder **Clfs, die Ähnlichkeit zu Daten vorhersagen** für die eine hohe Qualität angenommen wird (Proxydaten)

⇒ musterbasierte Filterung kommt schnell an Grenzen (z.B. harte Schwellwerte)

⇒ wenig Kontrolle über Proxydaten (z.B. GPT-Reddit-Heuristik)

1. Einleitung

Gap

- Methoden zur Filterung von großen Datensätzen sind oft **musterbasierte Heuristiken** und/oder **Clfs, die Ähnlichkeit zu Daten vorhersagen** für die eine hohe Qualität angenommen wird (Proxydaten)

⇒ musterbasierte Filterung kommt schnell an Grenzen (z.B. harte Schwellwerte)

⇒ wenig Kontrolle über Proxydaten (z.B. GPT-Reddit-Heuristik)

Vorschlag

- Datenfilterung als **multi-label Textklassifikation** zu modellieren
- Training mit **Active-Learning**

1. Einleitung

Beitrag

- Modell, das im besten Fall …

  → … ähnliche Fehlermuster wie eine Heuristik erkennt

  → … durch die ML-Eigenschaft kontrollierbare Qualitätslabels besitzt

  → … zusätzliche Texteigenschaften lernen kann

  → … nach der Evaluation die Machbarkeit einer Datenfilterung mittels ML-Klassifikation und Active-Learning erklärt

Gliederung

1. Einleitung
2. **Related Work**
3. Experiment 1: Annotationsschema und -richtlinien
4. Experiment 2: Active Learning (Ausblick)

2. Related Work:

Kreutzer et. al (2022)

- Analyse von Sprachkorpora mittels manueller Annotation von Qualitätsmerkmalen → z.B. Tag “correct” mit 53.52% in CCAligned-Sample

Abadji et. al. (2022)

- neue OSCAR-Version mit u.a. automatischen Qualitätsannotationen (“header/footer”, “tiny”, “noisy”)
- Analyse von z.B. Deutsch ergibt Label “clean” kommt nur mit 12% vor

2. Related Work:

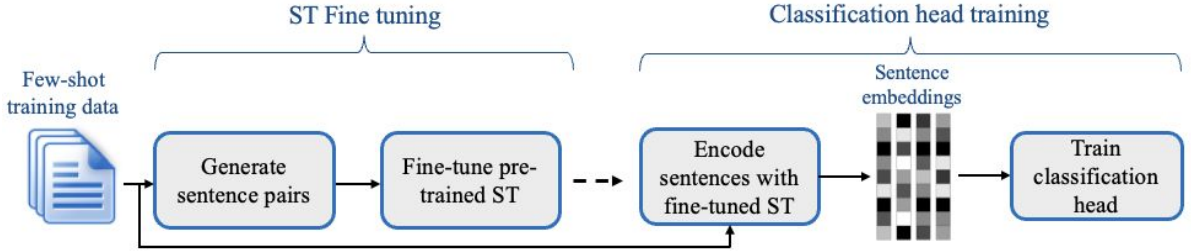
Longpre et al. (2023)

- Clf der Qualitätswert von 0 (hoch) bis 1 (niedrig) vorhersagt → Filterung von Trainingsdaten (C4) nach verschiedenen Schwellwerten
- Pretraining LLMs + Finetuning + Evaluation auf u.a. QA-Tasks und Toxic-ID

  → bis zu 6% Verbesserung bei 10% weniger Pretrainingsdaten

2. Related Work: Text-Klassifikation und Active Learning

SETFIT (Sentence Transformer Finetuning) (Tunstall et. al., 2022)



GBERT (Chan et al., 2020) → 89% Web, 11% “saubere” Daten

Logit-Modell → “one-vs-rest”

Pool-based Active-Learning → “uncertainty x diversity”

Gliederung

3.1 Ziel

Evaluiertes Annotationsschema und -richtlinie

- Definition von Kategorien, die Qualitätsmerkmale und Eigenschaften von Sätzen aus Webcrawls repräsentieren
- Entwicklung für Nicht-Expert:innen
- Evaluation mittels Inter-Annotator-Agreement

Vorarbeit für Experiment 2

- Goldstandard + getestetes AL-Setup

Gliederung

3.2 Methodik: Daten und Datenanalyse

Daten

- Wortschatz-Leipzig-Webcrawl 2018

  → ca. 600 Mio. Sätze

  → 84% “.de”, 11% “.at” und 5% “.ch”-Domains

Datenanalyse

- zufälliges Durchqueren und musterbasierte Suche
- Ergebnis: initialer Entwurf des Annotationsschemas bzw. der -richtlinie

3.2 Methodik: Agile Annotation

Rahmen

- 3 Annotatoren, 3 Runden, Budget von 8h
- iteratives Vorgehen: Evaluation nach jeder Runde + Überarbeitung

Batches

- Active-Learning-Batch (n = 460)
- Random-Batch (n = 600)
- Manual-Batch (n = 275)

3.2 Methodik: Metriken

Inter-Annotator-Agreement

- Cohens-Kappa: zwei Rater, ein Label
- Multi-Kappa: drei Rater, Jaccard-Distanz bei mehr als einem Label

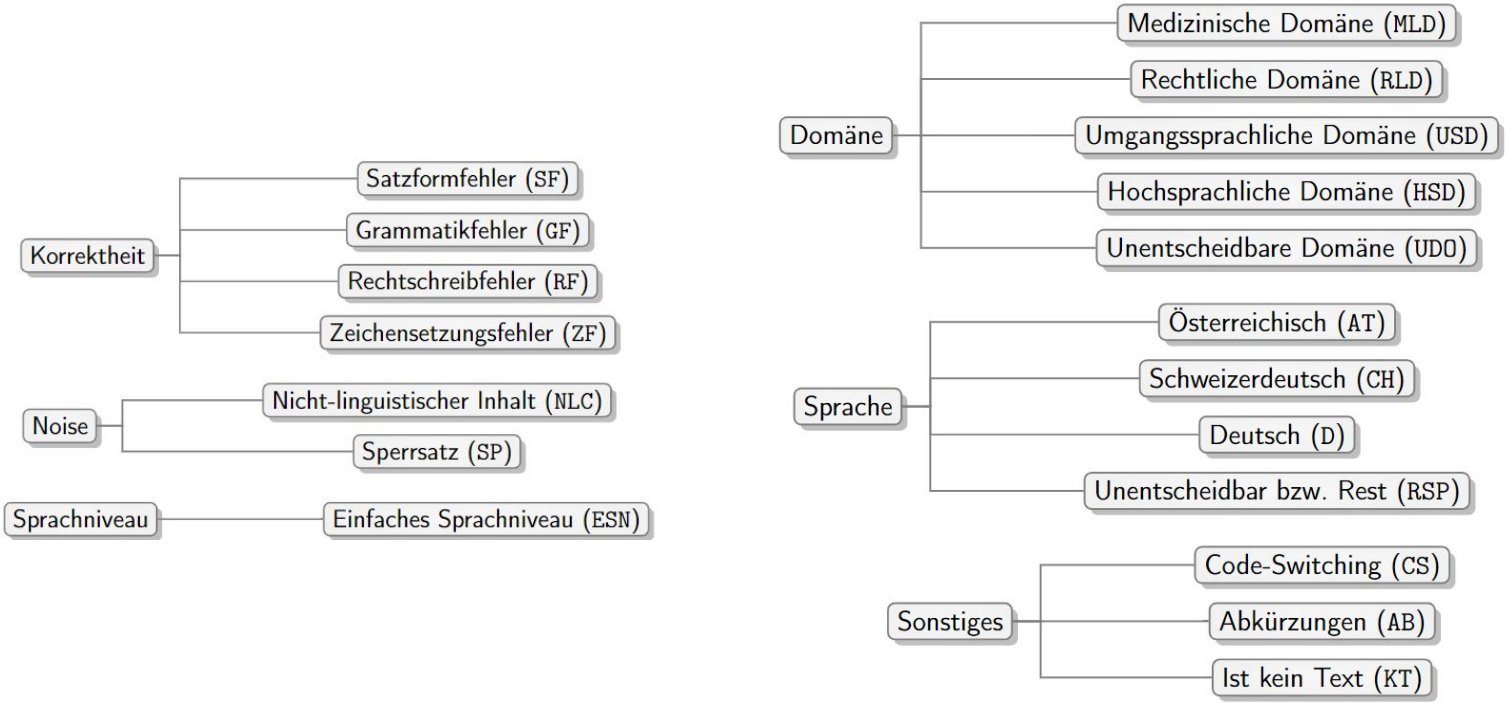
⇒ Formel (Artstein & Poesio, 2008): $$\kappa = \frac{A_o - A_e}{1 - A_e}$$

⇒ A_o = “observed agr. between two coders” (Multi-K.: $_{avg}A_o$)

⇒ A_e = “expected agr. between two coders by chance” (Multi-K.: $_{avg}A_e$)

Gliederung

3.3 Ergebnisse: Finales Annotationsschema



3.3 Ergebnisse: Finale Annotationsrichtlinie

Definitionen/Erklärungen

3.2.1 Kategorie: Korrektheit

Tags: `SF` (Satzformfehler), `GF` (Grammatikfehler), `RF` (Rechtschreibfehler), `ZF` (Zeichensetzungsfehler)

Definition: Ein Annotationsitem wird getaggt, wenn Satzform- oder Grammatik-, Rechtschreib- und Zeichensetzungsfehler[7] vorkommen.

Satzformfehler werden vergeben, wenn die Satzstruktur fehlerhaft oder unvollständig ist. Beispielsweise soll ein Tag vergeben werden, wenn der Satzanfang klein geschrieben, mit einem untypischen (Satz-)Zeichen beginnt oder das Satzendzeichen (Punkt, Ausrufezeichen, Fragezeichen, Auslassungspunkte oder ein Anführungszeichen, das auf ein Satzendzeichen folgt) fehlt. (...)

3.3 Ergebnisse: Finale Annotationsrichtlinie

Positivbeispiele (Kat. Korrektheit)

- Template-Modifikationen durch → `SF` `GF`
  Textfragment ohne Satzendzeichen und grammatikalisch inkorrekt, weil ein Prädikat fehlt.

Negativbeispiele (Kat. Noise)

- Meersalzgrotte – Klangentspannung
  Sprachtypisches Vorkommen des Bindestrichs und immer noch Text (vgl. Abb. 3.2). Insbesondere also Satzformfehler (`SF`) und durch die Wortreihung ein Grammatikfehler (`GF`).

3.3 Ergebnisse: Inter-Annotator-Agreement

IAA pro Batch

Multi-Kappa, Jaccard-Distanz				
	init	1. Rev.	2. Rev.	3. Rev.
AL-Batch	0.61908	0.70036	0.74329	**0.75609**
Random-Batch	0.75742	0.79377	**0.80707**	x
Manuel-Batch	0.61586	**0.71136**	x	x

⇒ Batch-interne Verbesserungen nach Diskussion/Anpassung der Kategorien

3.3 Ergebnisse: Inter-Annotator-Agreement

IAA pro Batch

Multi-Kappa, Jaccard-Distanz				
	init	1. Rev.	2. Rev.	3. Rev.
AL-Batch	0.61908	0.70036	0.74329	**0.75609**
Random-Batch	0.75742	0.79377	**0.80707**	x
Manuel-Batch	0.61586	**0.71136**	x	x

⇒ Batch-interne Verbesserungen nach Diskussion/Anpassung der Kategorien

⇒ Manuel-Batch? Neue, nicht diskutierte Beispiele! (CS, SP, CH, AT)

3.3 Ergebnisse: Inter-Annotator-Agreement

⇒ Positive Auswirkung des Relabelings von klassenspezifischen Beispielen (effizienter als komplettes Relabeling!)

⇒ z.B. Random-Batch:

Verbesserung nach Diskussion aller ESN-Bsp. von 0.18672 → 0.75723!

Multi-Kappa, Binäre Distanz			
	init	1. Rev.	2. Rev.
SF	0.82397	0.82397	0.83996
GF	0.68486	0.68272	0.70502
RF	0.46172	0.46172	0.45610
ZF	0.44606	0.44606	0.45024
NLC	0.74019	0.74019	0.74981
SP	1.00000	1.00000	1.00000
ESN	0.18672	0.75723	0.75723
MZD	0.49646	0.49646	0.49646
RLD	0.54033	0.54033	0.54033
USD	0.40789	0.73523	0.71565
HSD	0.46287	0.61198	0.61175
UDO	0.48587	0.69035	0.69035
AT	0.59893	0.59893	0.59893
CH	-0.00334	-0.00334	-0.00334
D	0.52492	0.62592	0.68622
RSP	0.62789	0.86618	0.86618
CS	1.00000	1.00000	1.00000
AB	0.68517	0.68517	0.68114
KT	0.79726	0.79726	0.79726
-inf	0.55130	0.55130	0.61492

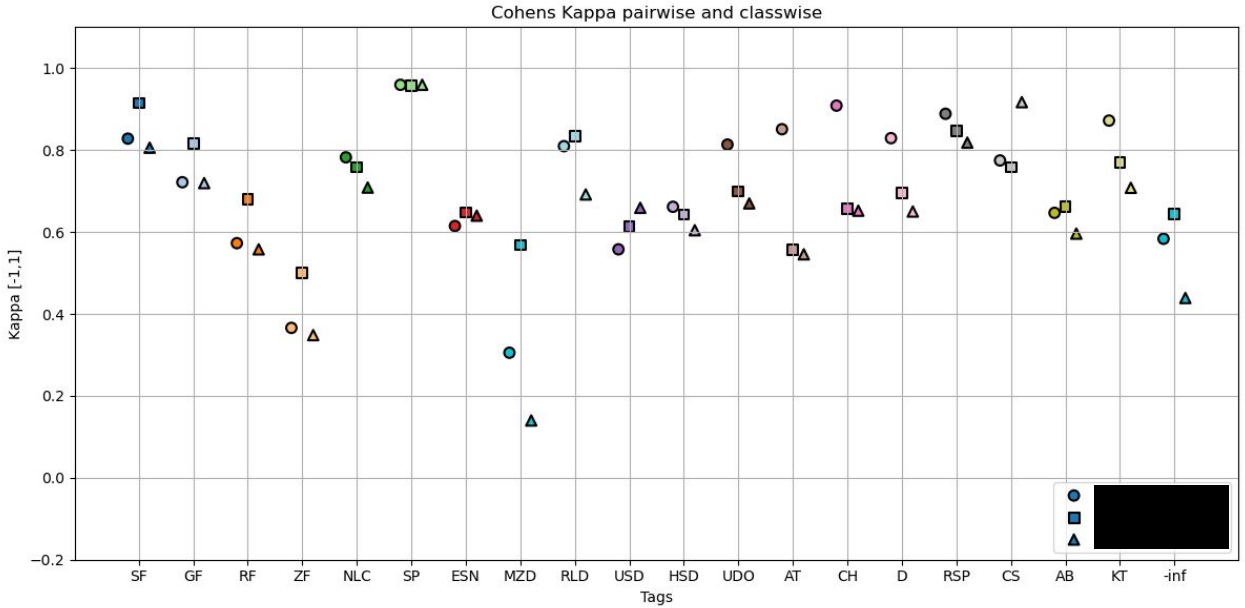
3.3 Ergebnisse: Inter-Annotator-Agreement

Kombi-Batch

⇒ 0.77423 auf allen Beispielen (“substantial”, aber < 0.8!)

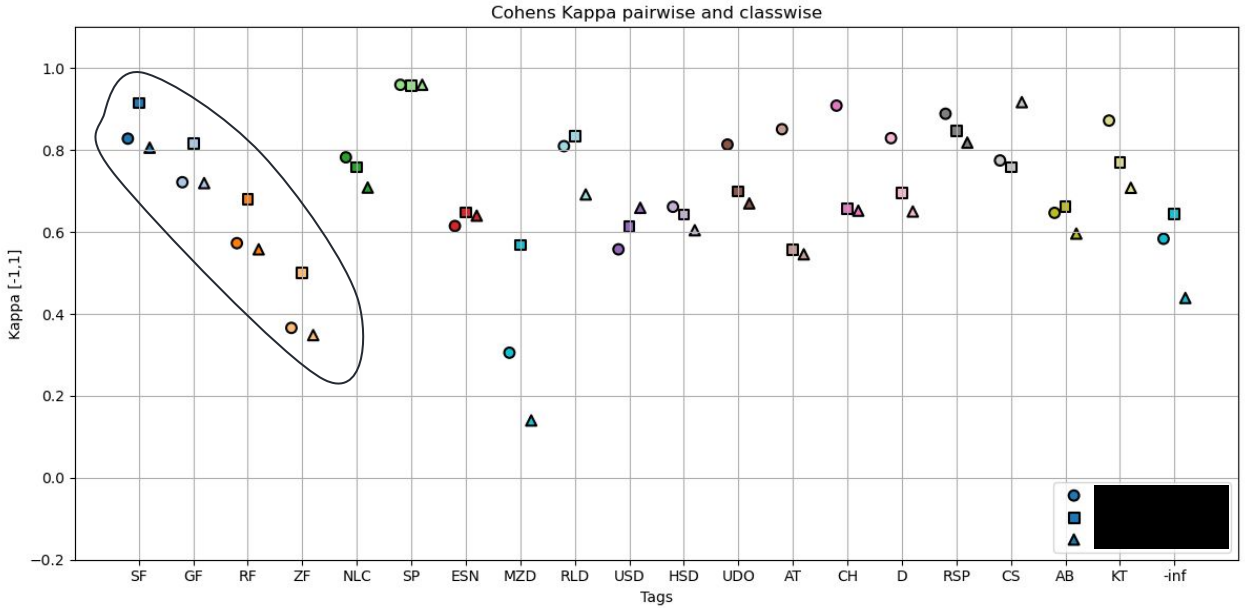
3.3 Ergebnisse: Inter-Annotator-Agreement

Kombi-Batch (detailliert)



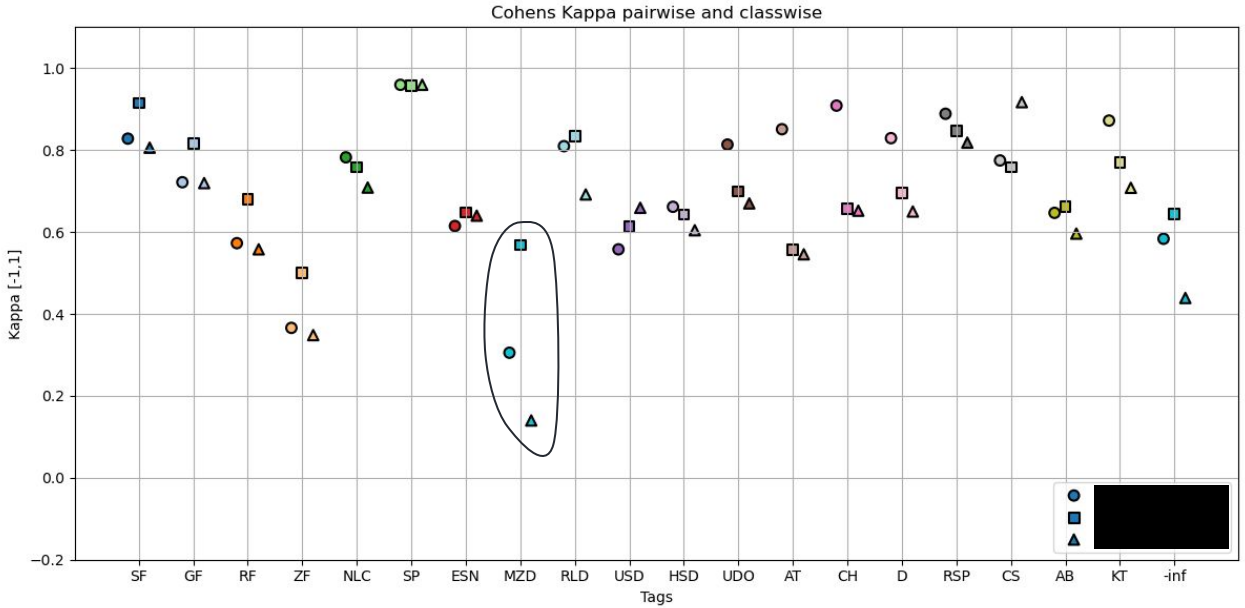
3.3 Ergebnisse: Inter-Annotator-Agreement

Kombi-Batch (detailliert)



3.3 Ergebnisse: Inter-Annotator-Agreement

Kombi-Batch (detailliert)



3.3 Ergebnisse: Inter-Annotator-Agreement

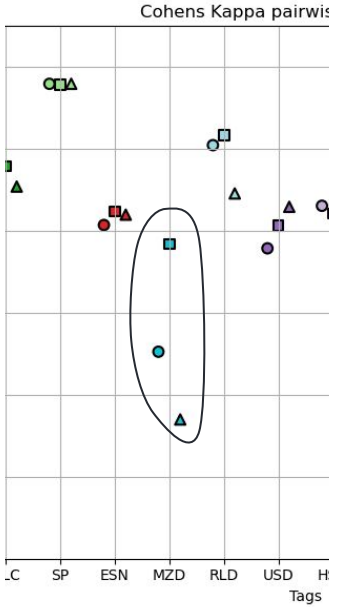
⇒ Medizinische Domäne fällt hier als Ausreißer auf

⇒ Uneindeutigkeit der Beispiele

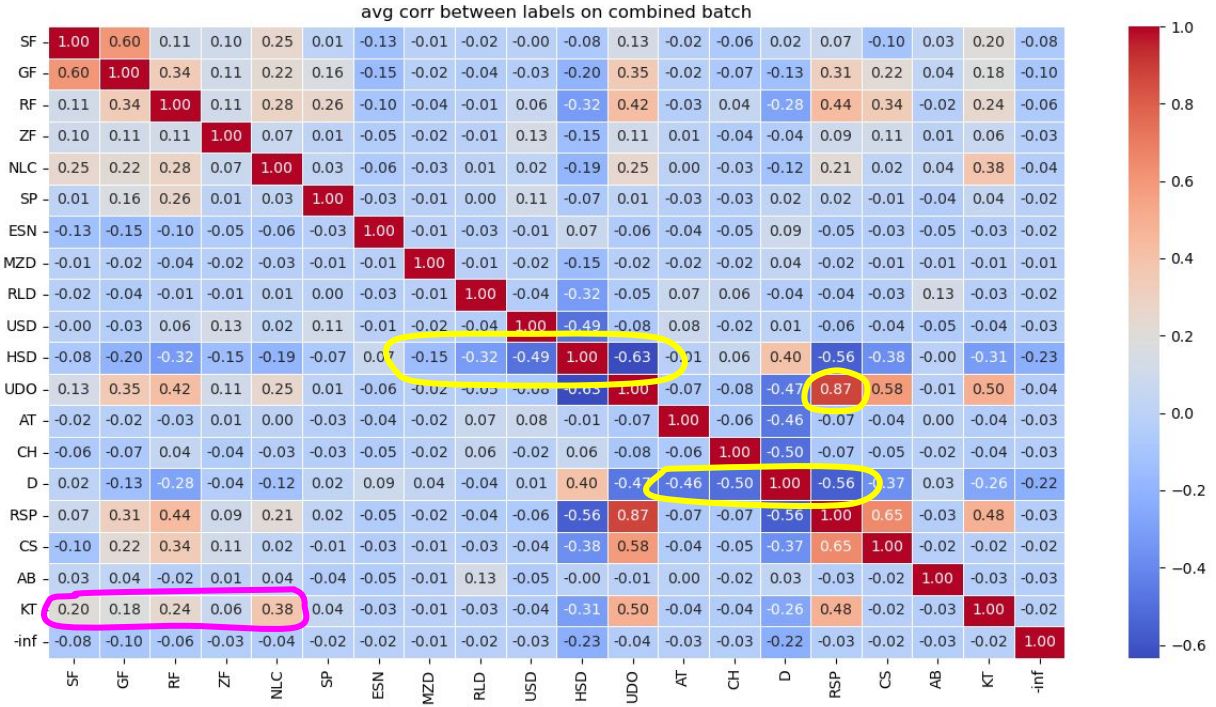
z.B. Sätze wie:

“Meningokokken sind Bakterien, die vor allem für Säuglinge und Kleinkinder eine Gefahr darstellen.”

“Das Protein, das Du über die Nahrung aufnimmst, muss vom Körper verarbeitet werden, wodurch im Abfallprodukte wie Ammoniak und Milchsäure entstehen.”



3.3 Ergebnisse: Labelkorrelationen



3.3 Ergebnisse: Zusammenfassung

- Annotationsschema, -richtlinie mit “substantial” Agreement
- Teilweise große Verbesserungen mit Agiler Annotation (Mehraufwand!)
- Vorbereitung für Experiment 2 ist (fast) fertig

Gliederung

1. Einleitung
2. Related Work
3. Experiment 1: Annotationsschema und -richtlinien
4. **Experiment 2: Active Learning (Ausblick)**

4. Experiment 2: Active Learning (Ausblick)

Rahmen → 8h Budget, trainierter Annotator, neue Daten

Methodik → AL mit “uncertainty” x “diversity”, SETFIT mit Logit-Modell

Evaluation

- Goldstandard aus Experiment 1
- Vgl. der Vorhersagen (bzw. “sauberen” Menge) mit “sentence-cleaner” von Wortschatz-Leipzig

Danke für die Aufmerksamkeit!

Literatur

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, et al.. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Longpre, S., Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David M. Mimno and Daphne Ippolito. 2023. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. ArXiv:2305.13169

Tunstall, Lewis & Reimers, Nils & Jo, Unso & Bates, Luke & Korat, Daniel & Wasserblat, Moshe & Pereg, Oren. 2022. Efficient Few-Shot Learning Without Prompts. ArXiv:.2209.11055.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28h International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Literatur

Ron Artstein and Massimo Poesio. 2008. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Landis, J Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33: 159-74 .

3.2 Methodik: Agile Annotation - Prozess

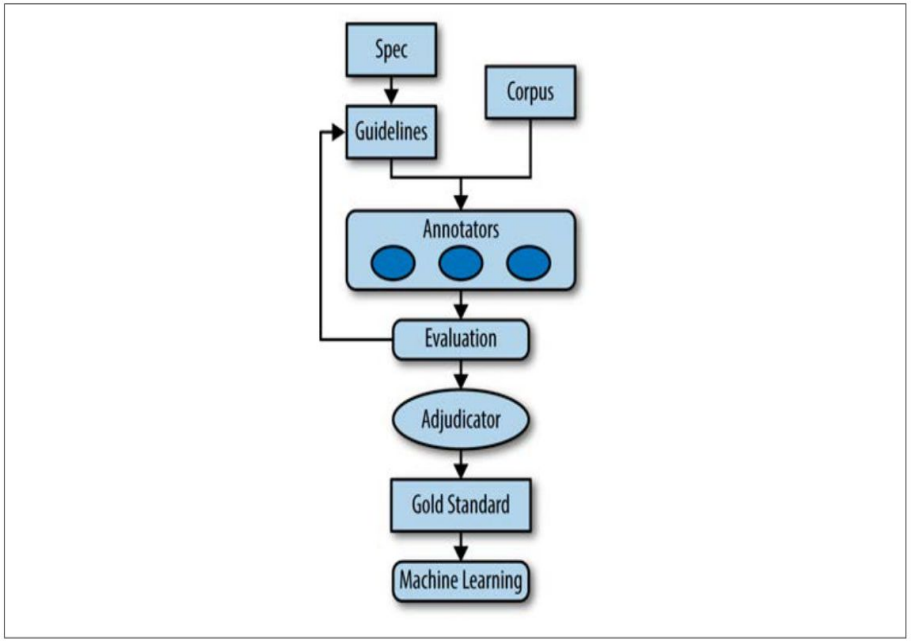


Figure 6-1. The annotation process

Grafik: Pustejovsky & Stubbs, 2012

Appendix: Distanzen

- Jaccard-Distanz vs. MASI-Distanz:

  ⇒ MASI gewichtet noch die Anzahl der unterschiedlichen Labels

  ⇒ z.B.: x = {1,2,3}, y = {1}, z = {1,2}

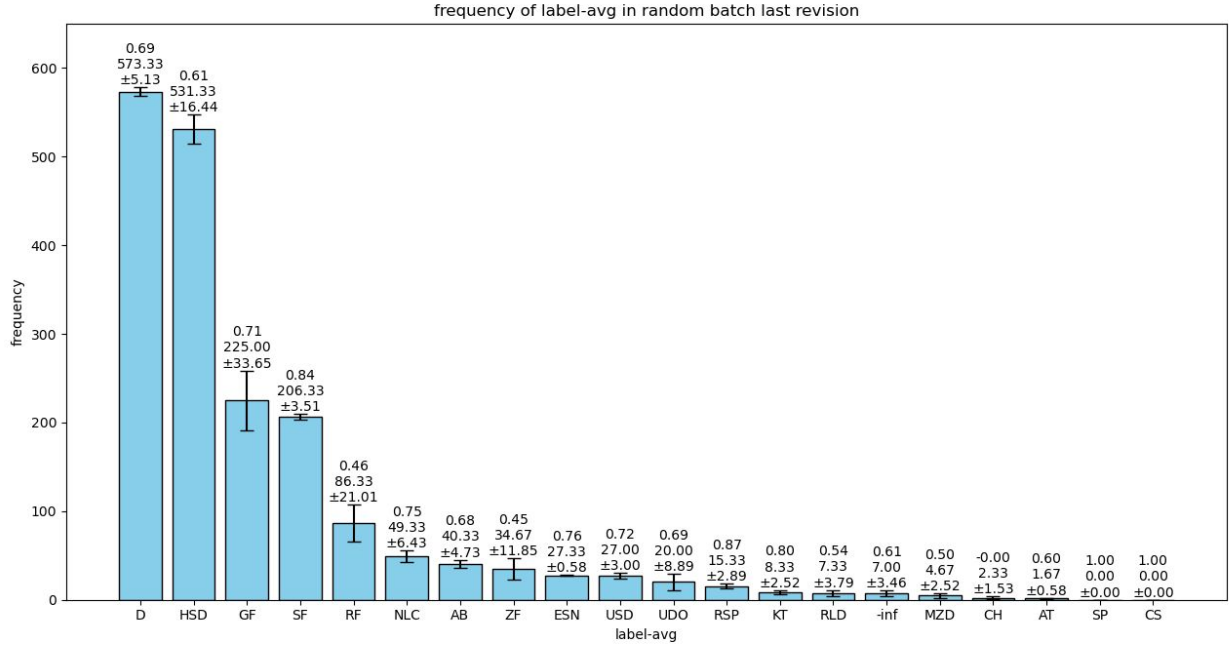
  ⇒ MASI(x,y) = 0.7766

  ⇒ MASI(x,z) = 0.5533

  ⇒ Jaccard(x,y) = 0.6666

  ⇒ Jaccard(x,z) = 0.3333

Appendix: Labelhäufigkeiten im Random-Batch



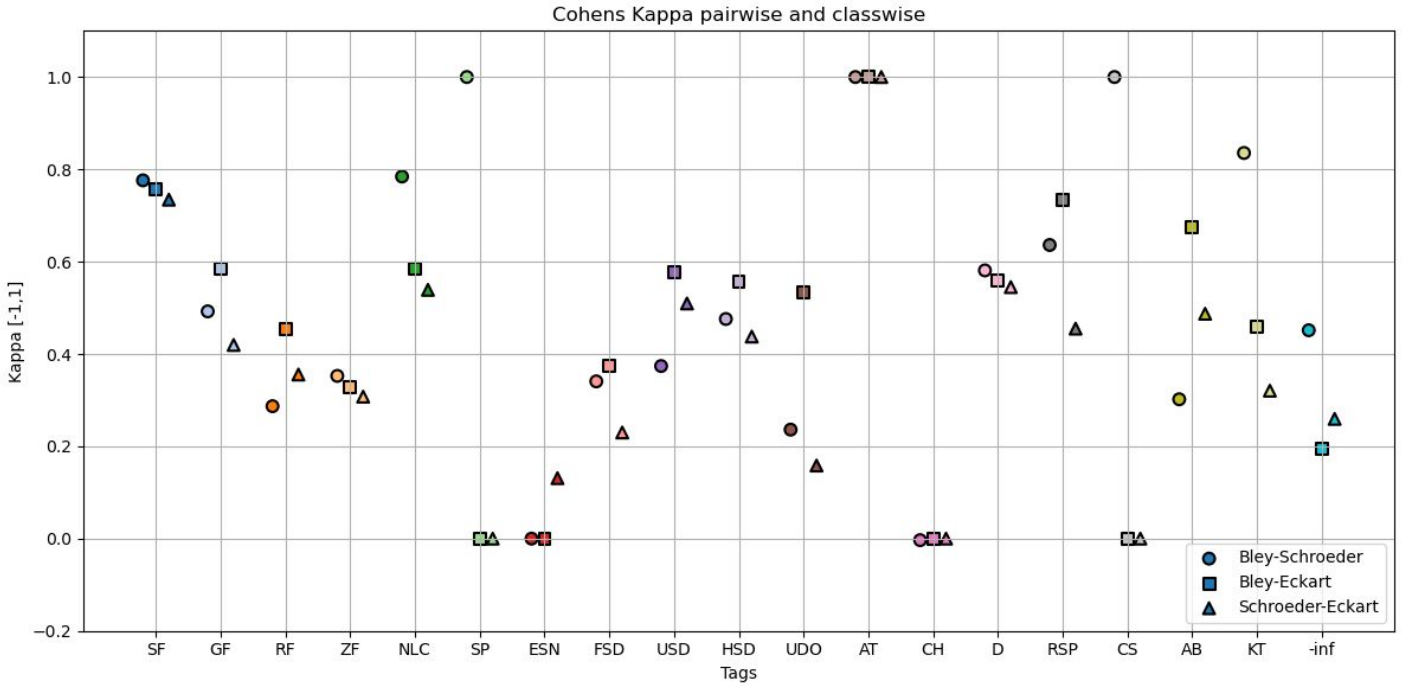
Appendix: Typische Labelkombinationen

	label_set		LC	LDiv
{'D', 'HSD'}	392		3,34606741	171
{'SF', 'D', 'HSD', 'GF'}	154			
{'HSD', 'D', 'GF'}	56			
{'CH', 'HSD'}	38			
{'D', 'HSD', 'ESN'}	37			
{'D', 'GF', 'SF', 'RF', 'HSD'}	33			
{'HSD', 'AT'}	33			
{'HSD', 'RF', 'D', 'GF'}	30			
{'SF', 'D', 'HSD'}	28			
{'AB', 'D', 'HSD'}	24			
{'-inf'}	22			
{'RF', 'D', 'HSD'}	20			
{'RSP', 'UDO', 'GF', 'CS', 'RF'}	19			
{'D', 'GF', 'SF', 'HSD', 'NLC'}	18			
{'HSD', 'RF', 'CH'}	16			
{'RF', 'D', 'HSD', 'NLC'}	13			
{'D', 'GF', 'SF', 'RF', 'HSD', 'NLC'}	13			
{'RF', 'RSP', 'GF', 'UDO'}	12			
{'USD', 'D'}	12			

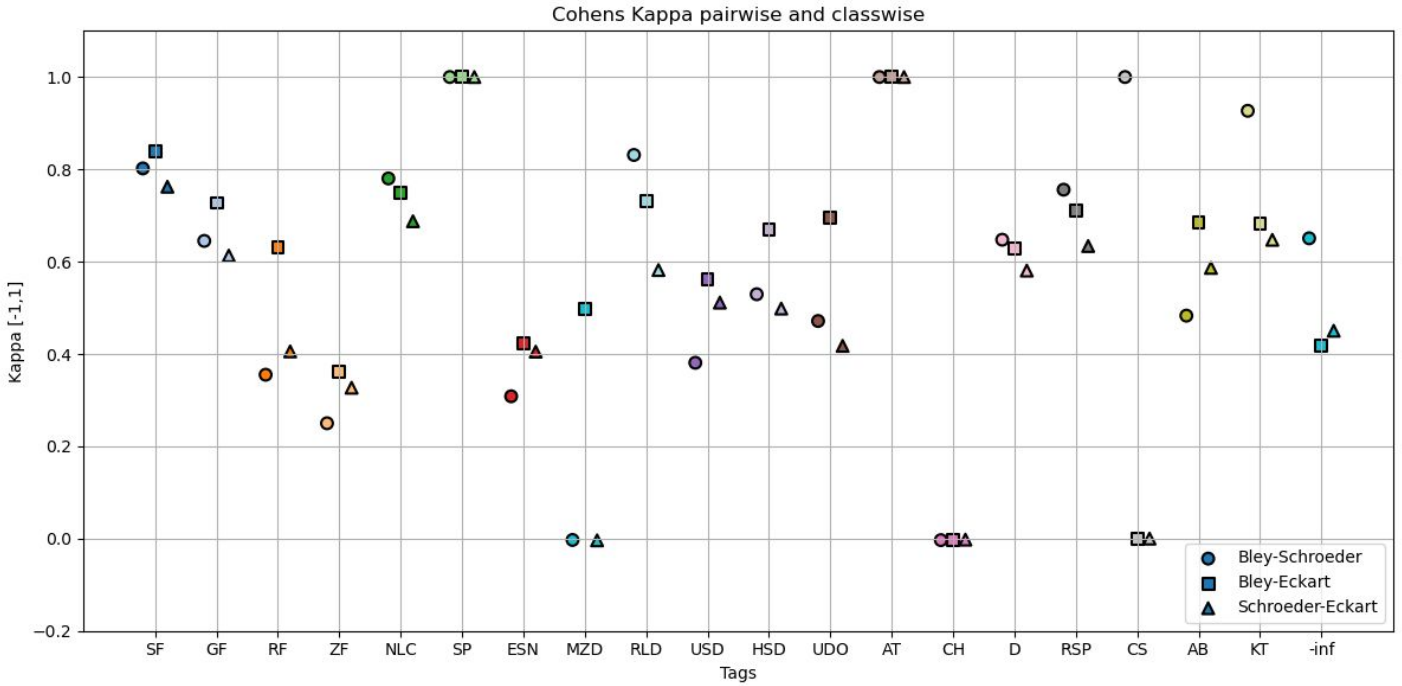
Appendix: Typische Sätze und Labelkombinationen

- 392x {'D', 'HSD'}: *Dadurch verteilen Bonellfedern den Körperdruck flächenelastisch.*
- 154x {'SF', 'D', 'HSD', 'GF'}: *Maßgeblicher Anteil am Klassenerhalt Damen 1*
- 56x {'HSD', 'D', 'GF'}: *3 kleine Zwiebeln schälen und in schneiden.*
- …
- 10x {'RSP', 'UDO', 'ZF', 'GF', 'SF', 'RF', 'KT', 'NLC'}: *user,grpjquota=quota.*

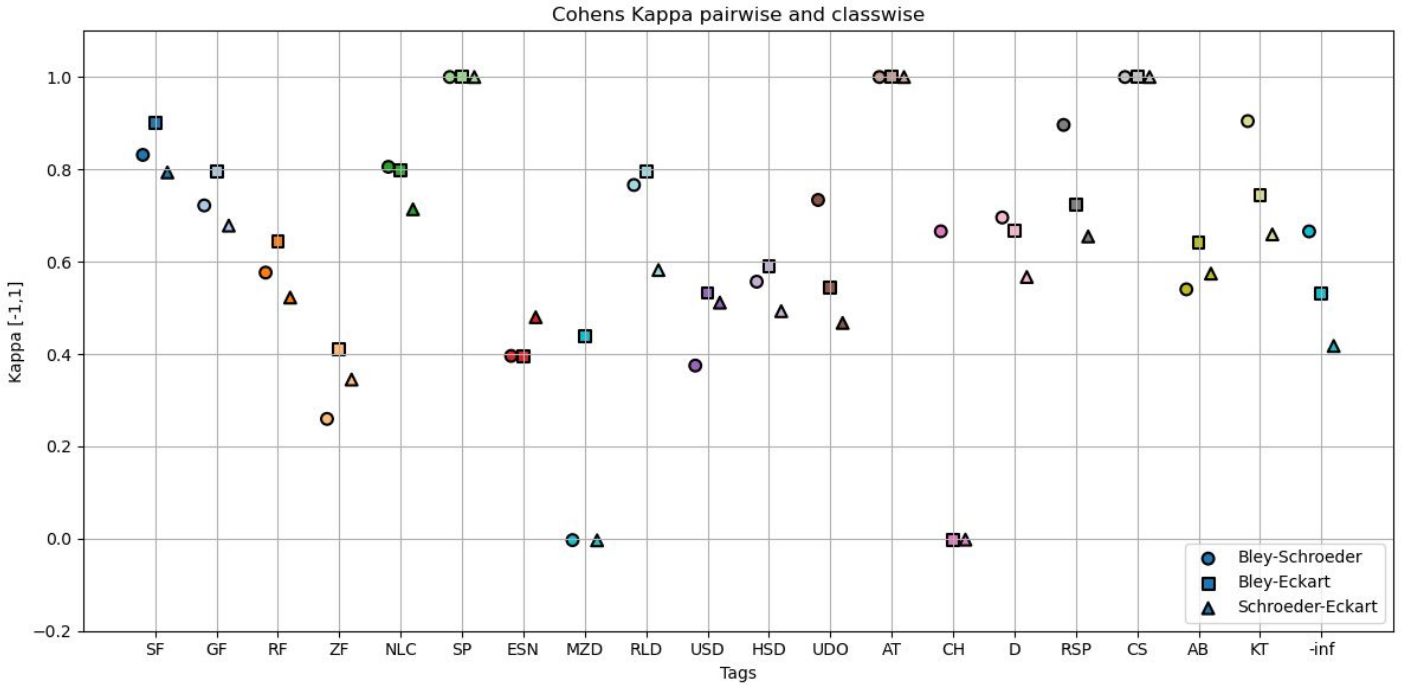
Appendix: IAA Batch AL init → 3rd Version (init)



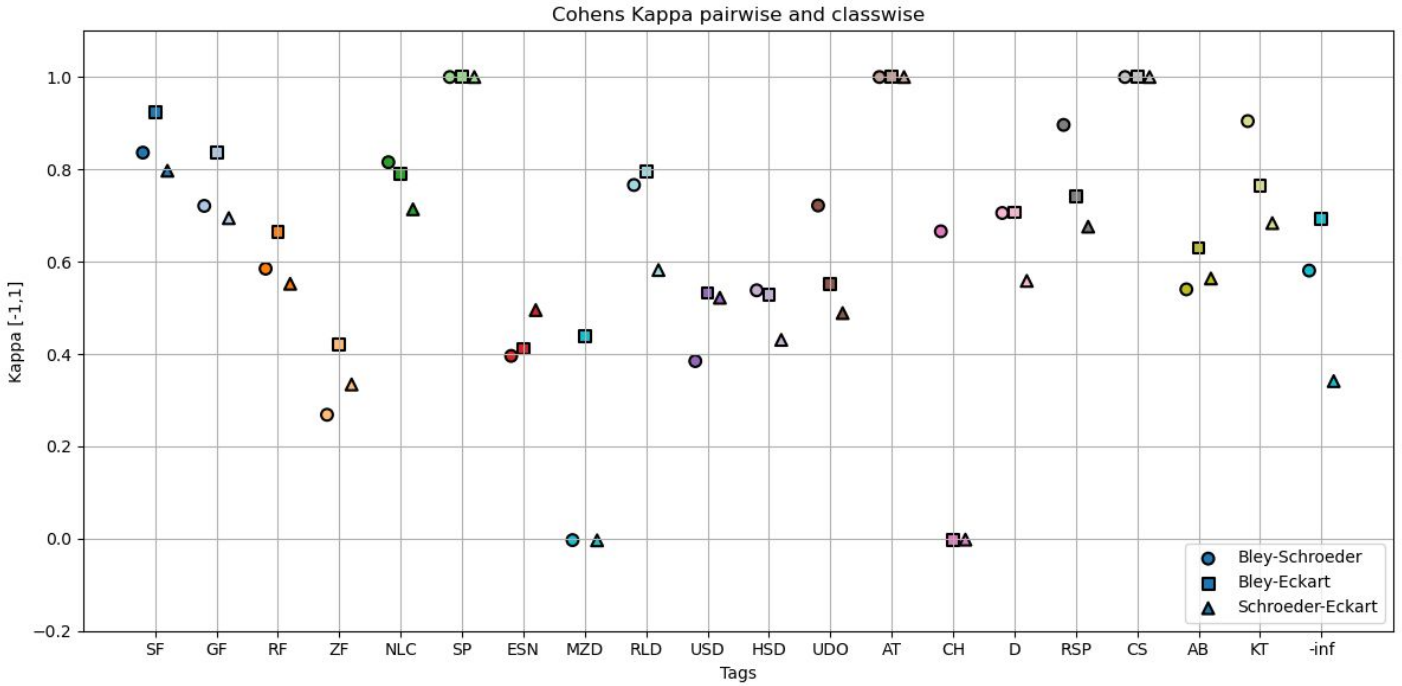
Appendix: IAA Batch AL init → 3rd Version (1st)



Appendix: IAA Batch AL init → 3rd Version (2nd)

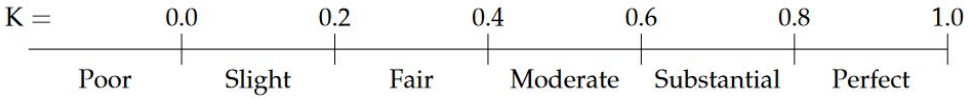


Appendix: IAA Batch AL init → 3rd Version (3rd)



Appendix: Interpretation Kappa

Interpretation (Landis & Koch,1977)



⇒ Diskussion über Interpretation der Scores

⇒ Mind. k > 0.67, noch besser k > 0.8

2. Related Work:

- **Trinh und Le (2018) (WSC), Radford et al. (2019) (GPT2):** Gute Ergebnisse durch saubere Daten (Finetuning vs. LLM)
- **Martin et al. (2020) (CamemBERT):** Relevanz von Datendiversität (implizit Datenqualität wg. OSCAR/CCNet)
- **Caswell et. al (2020) + Kreutzer et. al (2022):** Probleme bei LangID von sog. “under-resourced” Sprachen, was eine genauere Analyse motiviert hat → Annotieren mit Fehler/Qualitätslabels auf Satzebene mit Labels wie “wrong-language”, “non-linguistic”: z.B. CCAligned (65/119 Sprachen im Sample) 29.25% korrekt, WikiMatrix (20/78) 23.74% korrekt (ungewichtet nach Häufigkeit im Korpus) ⇒ “correct/in-language but unnatural” fehlt
- **Abadji et. al. (2022):** Neue OSCAR-Version mit automatischen Annotation wie “header”/”footer”, “short” oder “noisy” (Verhältnis von Buchstaben zu Nicht-Buchstaben > 0.5) ⇒ z.B. Französisch (~500GB) 11% Sätze korrekt

2. Related Work: Text-Klassifikation und Active Learning

Pool-based Active Learning (Settles, 2009)

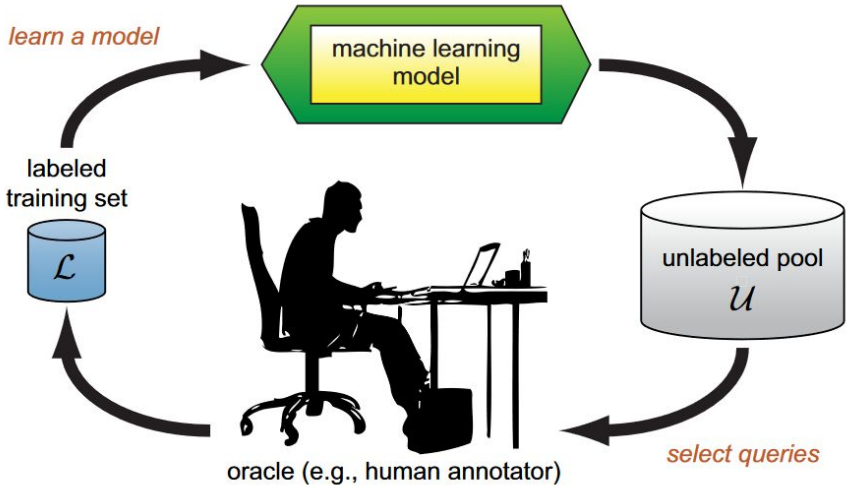


Figure 1: The pool-based active learning cycle.

2. Related Work: Text-Klassifikation und Active Learning

- Li and Guo, 2013: u = “max-margin prediction uncertainty” und c = “label cardinality inconsistency” (jeweils gewichtet mit β bzw. 1-β ∈ [0,1])
- **1.** u: groß/klein → hohe/wenig Unsicherheit in Vorhersage

  ⇒ $f_k(x_i)$ = proba for class k, $\hat{y}_i$ = pos./neg. class predictions (z.B. T < 0.5)

$$sep\_margin(\mathbf{x}_i) = \min_{k \in \hat{\mathbf{y}}_i^+} f_k(\mathbf{x}_i) - \max_{s \in \hat{\mathbf{y}}_i^-} f_s(\mathbf{x}_i)$$

  ⇒ Unwahrscheinlichste pos. Klasse - wahrscheinlichste neg. Klasse

  ⇒ u = 1 / sep_margin(x_i), d.h. je kleiner Nenner, desto größer u

2. Related Work: Text-Klassifikation und Active Learning

- **2.** c groß/klein → hohe/wenig Inkonsistenz zu durchschnittlicher Labelkardinalität (avg_lc) (d.h. bei avg_lc von 2 Labels pro Instanz ist Abweichung eines Bsps. mit 3 Labels gleich 1)

$$c(\mathbf{x}_i) = \left\| \sum_{k=1}^{K} I_{[\hat{\mathbf{y}}_{ik}>0]} - \frac{1}{N_\ell} \sum_{j=1}^{N_\ell} \sum_{k=1}^{K} I_{[\mathbf{y}_{jk}>0]} \right\|_2$$

  ⇒ Euklidische Distanz von erwarteten Labels zu durchschnittlich, observierten Labels

- Insgesamt: Möglichst unsichere und diverse Beispiele