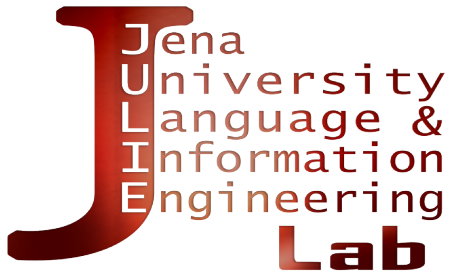
Annotation von Textabschnitten (Sections)

Statistische Auswertung von Annotationsergebnissen mit der Kappa/Alpha-Familie

Jena Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany

http://www.julielab.de



Ergebnisse der ersten Vorannotation

- 1. Runde abgeschlossen
- 30 fertige und verwertbare Annotate aus 20 Dokumenten
- Auswertung von benötigter Zeit und Qualität der Annotate

Ann.	0	1	2	3	4	5	6	7	8	9	10
A	pre	anam	diag	diag	ther	diag	ther	fut	fut	app	app
B	pre	anam	anam	diag	ther	ther	ther	ther	fut	app	app
C	pre	anam	diag	diag	pre	ther	ther	ther	fut	diag	app
D	pre	anam	diag	diag	ther	ther	ther	ther	fut	ther	app
E	pre	anam	diag	diag	pre	diag	ther	fut	fut	diag	app
F	pre	anam	anam	diag	ther	ther	ther	ther	fut	diag	app
G	pre	anam	anam	diag	ther	ther	ther	fut	fut	diag	app

Tabelle: Beispiel – Entscheidungen einer Annotationsaufgabe

Inter-Annotator Agreement

- Annotationsqualität zur Prüfung, wie gut Annotationsguide ist und für Gewährleistung der Reproduzierbarkeit
- Annotationsqualität nicht messbar durch Vergleich mit bspw. Goldstandard (cf. Centroid-Algorithmus etc.)
- deshalb: Inter-Annotator-Agreement (IAA)
- oft beobachtete Übereinstimmung als Maß verwendet
  - aber: keine Berücksichtigung zufälliger Übereinstimmungen
- Familie von Koeffizienten, die beobachtete und zufällige Übereinstimmung gegenüberstellt: Kappa/Alpha
  - Scott's Pi
  - Cohen's Kappa
  - Fleiss' Kappa
  - AC1
  - Krippendorf's Alpha

Fleiss' Kappa

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

$\bar{P} \in [0; 1]$: beobachtete Übereinstimmung, $\bar{P_e} \in [0; 1]$: erwartete (zufällige) Übereinstimmung

- Generalisierung von Scott's π für mehr als zwei Annotatoren
- $\pi, \kappa \in [-1, 1]$
- negative Werte: beobachtete Übereinstimmung $\bar{P}$ geringer als zufällige Übereinstimmung P_e
- bei Durchschnittsbildung über mehrere Tasks problematisch (jeder negative Werte verzerrt stark)
- sehr geringe Variation der Daten (fast alle Token gleiches Label) sehr empfindlich gegenüber Nichtübereinstimmung (geringe Varianz führt zu hohem P_e)
  $\rightarrow$ ungeeignet für einfache und kleine Mengen von Aufgaben

Krippendorff's Alpha

- misst im Gegensatz zu vorherigen Koeffizienten beobachtete und zufällige „Nichtübereinstimmung“
- etwas komplexer Algorithmus, seltener angewendet

$$\alpha = 1 - \frac{D_0}{D_e} = \frac{A_0 - A_e}{1 - A_0} = \frac{(n-1)\sum_c o_{cc} - \sum_c n_c(n_c-1)}{n(n-1)\sum_c n_c(n_c-1)} \quad (2)$$

D_0: beobachtete Nichtübereinstimmung, D_e: erwartete (zufällige) Nichtübereinstimmung

Krippendorff's Alpha

Units u:	1	2	3	4	5	6	7	8	9	10	11	12	
A	1	2	3	3	2	1	4	1	2	.	.	.	
B	1	2	3	3	2	2	4	1	2	5	.	3	
C	.	3	3	3	2	3	4	2	2	5	1	.	
D	1	2	3	3	2	4	4	1	2	5	1	.	
m_u in u	3	4	4	4	4	4	4	4	4	3	2	1	41

Tabelle: Beispiel – Berechnung Krippendorff's Alpha (Krippendorff 2011)

	1	.	k	.	.	
1	o_{11}	.	o_{1k}	.	.	n_1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
c	o_{c1}	.	.	.	.	$n_c = \sum_k o_{ck}$
.	.	.	.	.	.	.
	n_1	.	n_1	.	.	$n_c = \sum_c \sum_k n_{ck}$

Tabelle: Koinzidenz-Matrix (allg.)

	1	2	3	4	5	
1	7	4/3	1/3	1/3	.	9
2	4/3	10	4/3	1/3	.	13
3	1/3	1/3	8	1/3	.	10
4	1/3	1/3	1/3	4	.	5
5	.	.	.	.	3	3
	9	13	10	5	3	40

Tabelle: Koinzidenz-Matrix

Beispiel Krippendorf's Alpha

Ann.	0	1	2	3	4	5	6	7	8	9	10
A	pre	anam	diag	diag	ther	diag	ther	fut	fut	app	app
B	pre	anam	anam	diag	ther	ther	ther	ther	fut	app	app
C	pre	anam	diag	diag	pre	ther	ther	ther	fut	diag	app
D	pre	anam	diag	diag	ther	ther	ther	ther	fut	ther	app
E	pre	anam	diag	diag	pre	diag	ther	fut	fut	diag	app
F	pre	anam	anam	diag	ther	ther	ther	ther	fut	diag	app
G	pre	anam	anam	diag	ther	ther	ther	fut	fut	diag	app

Tabelle: Beispiel – Entscheidungen einer Annotationsaufgabe

$\alpha = 0.694$ (Beispiel)

	anam	app	diag	fut	pre	ther
anam	8	0	2	0	0	0
app	0	7.33	1.33	0	0	0.33
diag	2	1.33	11.33	0	0	2.33
fut	0	0	0	8	0	2
pre	0	0	0	0	7.33	1.66
ther	0	0.33	2.33	2	1.66	15.66

Tabelle: Beispiel – Koinzidenz-Matrix

Anwendung

- IAA für Sektionsannotation (relativ hoch)
- Übereinstimmung für einzelne Tasks (über alle Annotatoren) und gesamt
- Implementierung?!

Literatur

- Artstein, R. (2017) Inter-Annotator Agreement. In: Ide, Nancy, James Pustejovsky (eds.), Handbook of Linguistic Annotation, Springer Netherlands, 2017, 297-313.
- Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability.
- Powers, D.M.W. (2012). The Problem with Kappa. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 345–355.
- Blog von K. Gwet, Urheber von AC1 und AC2
- Joyce, M.(2013) Picking the Best Intercoder Reliability Statistic for Your Digital Activism Content Analysis