Ranking the annotators: An agreement study on argumentation structure

Andreas Peldszus Manfred Stede

Applied Computational Linguistics, University of Potsdam

The 7th Linguistic Annotation Workshop Interoperability with Discourse
ACL Workshop, Sofia, August 8-9, 2013

Introduction

classic reliability study

- 2 or 3 annotators
- authors, field experts, at least motivated and experienced annotators
- measure agreement, identify sources of disagreement

Introduction

classic reliability study

- 2 or 3 annotators
- authors, field experts, at least motivated and experienced annotators
- measure agreement, identify sources of disagreement

crowd-sourced corpus

- 100-x annotators
- crowd
- bias correction [Snow et al., 2008]
  outlier identification, find systematic differences [Bhardwaj et al., 2010]
  spammer detection [Raykar and Yu, 2012]

Introduction

classic reliability study

- 2 or 3 annotators
- authors, field experts, at least motivated and experienced annotators
- measure agreement, identify sources of disagreement

classroom annotation

- 20-30 annotators
- students with different ability and motivation, obligatory participation
- do both: test reliabilty & identify and group characteristic annotation behaviour

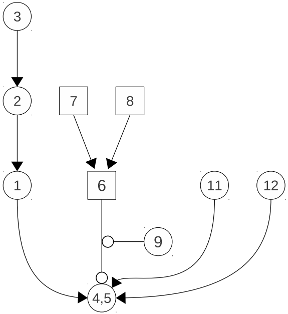
crowd-sourced corpus

- 100-x annotators
- crowd
- bias correction [Snow et al., 2008]
  outlier identification, find systematic differences [Bhardwaj et al., 2010]
  spammer detection [Raykar and Yu, 2012]

Outline

1. Introduction
2. Experiment
3. Evaluation
4. Ranking and clustering the annotators

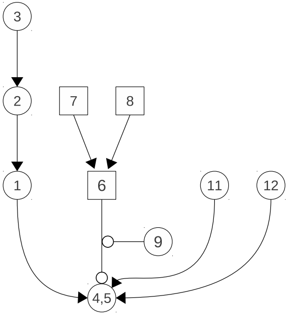
Experiment Task: Argumentation Structure



Scheme based on Freeman [1991, 2011]

- node types = *argumentative role*
  proponent (presents and defends claims)
  opponent (critically questions)
- link types = *argumentative function*
  support own claims (normally, by example)
  attack other's claims (rebut, undercut)

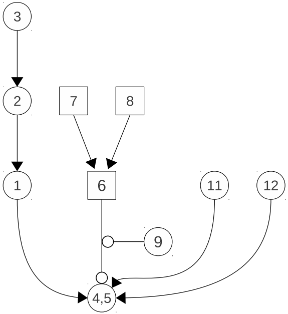
Experiment Task: Argumentation Structure



Scheme based on Freeman [1991, 2011]

- node types = *argumentative role*
  proponent (presents and defends claims)
  opponent (critically questions)
- link types = *argumentative function*
  support own claims (normally, by example)
  attack other's claims (rebut, undercut)

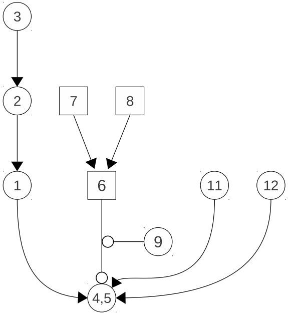
Experiment Task: Argumentation Structure



Scheme based on Freeman [1991, 2011]

- node types = *argumentative role*
  proponent (presents and defends claims)
  opponent (critically questions)
- link types = *argumentative function*
  support own claims (normally, by example)
  attack other's claims (rebut, undercut)

Experiment Task: Argumentation Structure



Scheme based on Freeman [1991, 2011]

- node types = *argumentative role*
  proponent (presents and defends claims)
  opponent (critically questions)
- link types = *argumentative function*
  support own claims (normally, by example)
  attack other's claims (rebut, undercut)

This annotation is tough!

- fully connected discourse structure
- unitizing ADUs from EDUs is already a complex text-understanding task

Experiment Data: Micro-Texts

Thus, we use micro-texts:

- 23 short, constructed, German texts
- each text exactly 5 segments long
- each segment is argumentatively relevant
- covering different argumentative configurations

A (translated) example

[*Energy-saving light bulbs contain a considerable amount of toxic substances.* $]_1$ [*A customary lamp can for instance contain up to five milligrams of quicksilver.* $]_2$ [*For this reason, they should be taken off the market,* $]_3$ [*unless they are virtually unbreakable.* $]_4$ [*This, however, is simply not case.* $]_5$

Experiment Data: Micro-Texts

Thus, we use micro-texts:

- 23 short, constructed, German texts
- each text exactly 5 segments long
- each segment is argumentatively relevant
- covering different argumentative configurations

A (translated) example

[*Energy-saving light bulbs contain a considerable amount of toxic substances.*]$_1$ [*A customary lamp can for instance contain up to five milligrams of quicksilver.*]$_2$ [*For this reason, they should be taken off the market,*]$_3$ [*unless they are virtually unbreakable.*]$_4$ [*This, however, is simply not case.*]$_5$

Experiment Setup: Classroom Annotation

Obligatory annotation in class with 26 undergraduate students:

- minimal training
  - 5 min. introduction
  - 30 min. reading guidelines (6p.)
  - very brief question answering
- 45 min. annotation

Annotation in three steps:

- identify central claim / thesis
- decide on argumentative role for each segment
- decide on argumentative function for each segment

Experiment Setup: Classroom Annotation

Obligatory annotation in class with 26 undergraduate students:

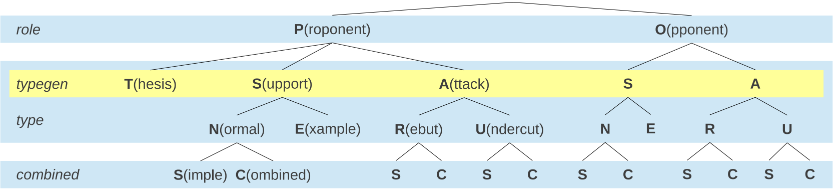
- minimal training
  - 5 min. introduction
  - 30 min. reading guidelines (6p.)
  - very brief question answering
- 45 min. annotation

Annotation in three steps:

- identify central claim / thesis
- decide on argumentative role for each segment
- decide on argumentative function for each segment

Evaluation: Preparation

Rewrite graphs as a list of (relational) segment labels

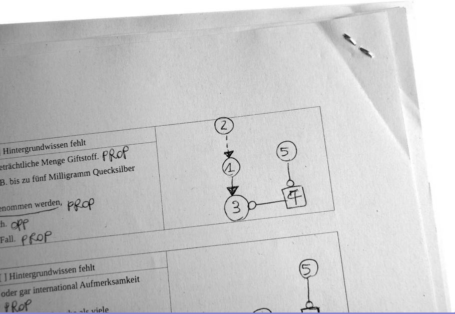


role **P**(roponent) **O**(pponent)

typegen **T**(hesis) **S**(upport) **A**(ttack) **S** **A**

type **N**(ormal) **E**(xample) **R**(ebut) **U**(ndercut) **N** **E** **R** **U**

combined **S**(imple) **C**(ombined) **S** **C** **S** **C** **S** **C** **S** **C** **S** **C**



Hintergrundwissen fehlt

...trächtliche Menge Giftstoff. PROP

...B. bis zu fünf Milligramm Quecksilber

...enommen werden, PROP

...h. OPP

Fall. PROP

[ ] Hintergrundwissen fehlt

oder gar international Aufmerksamkeit

PROP

...als viele

```
1:PSNS(3)
2:PSES(1)
3:PT()
4:OARS(3)
5:PARS(4)
```

Evaluation: Results

level	#cats	κ	A_O	A_E	α	D_O	D_E
role+type+comb+target	(71)	**0.384**	0.44	0.08			

unweighted scores in κ [Fleiss, 1971], weighted scores in α [Krippendorff, 1980]

- low agreement for the full task
- varying difficulty on the simple levels
- other complex levels: target identification has only small impact
- hierarchically weighted IAA yields slightly better results

Evaluation: Results

level	#cats	κ	A_O	A_E	α	D_O	D_E
role	2	**0.521**	0.78	0.55			
typegen	3	**0.579**	0.72	0.33			
type	5	**0.469**	0.61	0.26			
comb	2	**0.458**	0.73	0.50			
target	(9)	**0.490**	0.58	0.17			
role+type+comb+target	(71)	0.384	0.44	0.08			

unweighted scores in κ [Fleiss, 1971], weighted scores in α [Krippendorff, 1980]

- low agreement for the full task
- varying difficulty on the simple levels
- other complex levels: target identification has only small impact
- hierarchically weighted IAA yields slightly better results

Evaluation: Results

level	#cats	κ	A_O	A_E	α	D_O	D_E
role	2	0.521	0.78	0.55			
typegen	3	0.579	0.72	0.33			
type	5	0.469	0.61	0.26			
comb	2	0.458	0.73	0.50			
target	(9)	0.490	0.58	0.17			
role+typegen	5	**0.541**	0.66	0.25			
role+type	9	**0.450**	0.56	0.20			
role+type+comb	15	**0.392**	0.49	0.16			
role+type+comb+target	(71)	0.384	0.44	0.08			

unweighted scores in κ [Fleiss, 1971], weighted scores in α [Krippendorff, 1980]

- low agreement for the full task
- varying difficulty on the simple levels
- other complex levels: target identification has only small impact
- hierarchically weighted IAA yields slightly better results

Evaluation: Results

level	#cats	κ	A$_O$	A$_E$	α	D$_O$	D$_E$
role	2	0.521	0.78	0.55			
typegen	3	0.579	0.72	0.33			
type	5	0.469	0.61	0.26			
comb	2	0.458	0.73	0.50			
target	(9)	0.490	0.58	0.17			
role+typegen	5	0.541	0.66	0.25	**0.534**	0.28	0.60
role+type	9	0.450	0.56	0.20	**0.500**	0.33	0.67
role+type+comb	15	0.392	0.49	0.16	**0.469**	0.38	0.71
role+type+comb+target	(71)	0.384	0.44	0.08	**0.425**	0.45	0.79

unweighted scores in κ [Fleiss, 1971], weighted scores in α [Krippendorff, 1980]

- low agreement for the full task
- varying difficulty on the simple levels
- other complex levels: target identification has only small impact
- hierarchically weighted IAA yields slightly better results

Evaluation: Category confusions

- studying all individual confusion matrices not feasible: 26 annotators, 325 different pairs of annotators
- Cinková et al. [2012]: sum up all confusion matrices and build a **probabilistic confusion matrix**

Evaluation: Category confusions

- studying all individual confusion matrices not feasible:
  26 annotators, 325 different pairs of annotators
- Cinková et al. [2012]: sum up all confusion matrices and build a **probabilistic confusion matrix**

Evaluation: Category confusions

- studying all individual confusion matrices not feasible:
  26 annotators, 325 different pairs of annotators
- Cinková et al. [2012]: sum up all confusion matrices and build a **probabilistic confusion matrix**

Evaluation: Category confusions

- studying all individual confusion matrices not feasible: 26 annotators, 325 different pairs of annotators
- Cinková et al. [2012]: sum up all confusion matrices and build a **probabilistic confusion matrix**

	PT	PSN	PSE	PAR	PAU	OSN	OSE	OAR	OAU	?
PT	0.625	0.243	0.005	0.003	0.002	0.006	0.000	0.030	0.007	0.078
PSN	0.123	0.539	0.052	0.034	0.046	0.055	0.001	0.052	0.021	0.078
PSE	0.024	0.462	0.422	0.007	0.008	0.000	0.000	0.015	0.001	0.061
PAR	0.007	0.164	0.004	0.207	0.245	0.074	0.000	0.156	0.072	0.071
PAU	0.007	0.264	0.005	0.290	0.141	0.049	0.000	0.117	0.075	0.052
OSN	0.016	0.292	0.000	0.081	0.046	0.170	0.004	0.251	0.075	0.065
OSE	0.000	0.260	0.000	0.000	0.000	0.260	0.000	0.240	0.140	0.100
OAR	0.033	0.114	0.004	0.070	0.044	0.102	0.001	0.339	0.218	0.076
OAU	0.017	0.101	0.000	0.069	0.061	0.066	0.002	0.469	0.153	0.063
?	0.179	0.351	0.031	0.066	0.041	0.055	0.001	0.157	0.061	0.057

for the ‘role+type’-level; ‘?’ = missing annotations

Evaluation: Category confusions

- studying all individual confusion matrices not feasible:
  26 annotators, 325 different pairs of annotators
- Cinková et al. [2012]: sum up all confusion matrices and build a **probabilistic confusion matrix**

	PT	PSN	PSE	PAR	PAU	OSN	OSE	OAR	OAU	?
PT	**0.625**	0.243	0.005	0.003	0.002	0.006	0.000	0.030	0.007	0.078
PSN	0.123	**0.539**	0.052	0.034	0.046	0.055	0.001	0.052	0.021	0.078
PSE	0.024	0.462	**0.422**	0.007	0.008	0.000	0.000	0.015	0.001	0.061
PAR	0.007	0.164	0.004	**0.207**	0.245	0.074	0.000	0.156	0.072	0.071
PAU	0.007	0.264	0.005	0.290	**0.141**	0.049	0.000	0.117	0.075	0.052
OSN	0.016	0.292	0.000	0.081	0.046	**0.170**	0.004	0.251	0.075	0.065
OSE	0.000	0.260	0.000	0.000	0.000	0.260	**0.000**	0.240	0.140	0.100
OAR	0.033	0.114	0.004	0.070	0.044	0.102	0.001	**0.339**	0.218	0.076
OAU	0.017	0.101	0.000	0.069	0.061	0.066	0.002	0.469	**0.153**	0.063
?	0.179	0.351	0.031	0.066	0.041	0.055	0.001	0.157	0.061	**0.057**

for the ‘role+type’-level; ‘?’ = missing annotations

Evaluation: Category confusions

- studying all individual confusion matrices not feasible:
  26 annotators, 325 different pairs of annotators
- Cinková et al. [2012]: sum up all confusion matrices and build a **probabilistic confusion matrix**

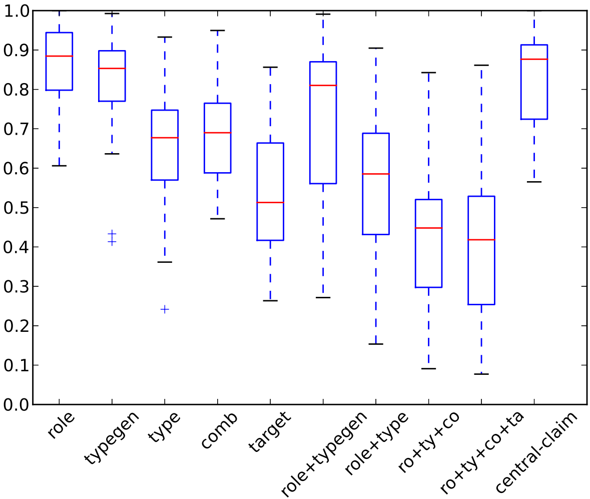
	PT	PSN	PSE	PAR	PAU	OSN	OSE	OAR	OAU	?
PT	**0.625**	0.243	0.005	0.003	0.002	0.006	0.000	0.030	0.007	0.078
PSN	0.123	**0.539**	0.052	0.034	0.046	0.055	0.001	0.052	0.021	0.078
PSE	0.024	0.462	**0.422**	0.007	0.008	0.000	0.000	0.015	0.001	0.061
PAR	0.007	0.164	0.004	**0.207**	0.245	0.074	0.000	0.156	0.072	0.071
PAU	0.007	0.264	0.005	0.290	**0.141**	0.049	0.000	0.117	0.075	0.052
OSN	0.016	0.292	0.000	0.081	0.046	**0.170**	0.004	0.251	0.075	0.065
OSE	0.000	0.260	0.000	0.000	0.000	0.260	**0.000**	0.240	0.140	0.100
OAR	0.033	0.114	0.004	0.070	0.044	0.102	0.001	**0.339**	0.218	0.076
OAU	0.017	0.101	0.000	0.069	0.061	0.066	0.002	0.469	**0.153**	0.063
?	0.179	0.351	0.031	0.066	0.041	0.055	0.001	0.157	0.061	**0.057**

for the ‘role+type’-level; ‘?’ = missing annotations

Evaluation: Comparison with Gold-Data

Evaluation: Comparison with Gold-Data

Distribution of annotator's F1 score per level, macro-averaged over categories



Ranking and clustering the annotators

Questions:

- What range of agreement is possible in this group of annotators?
- How to give structure to this inhomogenous group of annotators?
- How to identify subgroups of good annotators, how to sort out bad ones without too much gold data?

Ranking and clustering the annotators

Questions:

- What range of agreement is possible in this group of annotators?
- How to give structure to this inhomogenous group of annotators?
- How to identify subgroups of good annotators, how to sort out bad ones without too much gold data?

Ranking and clustering the annotators

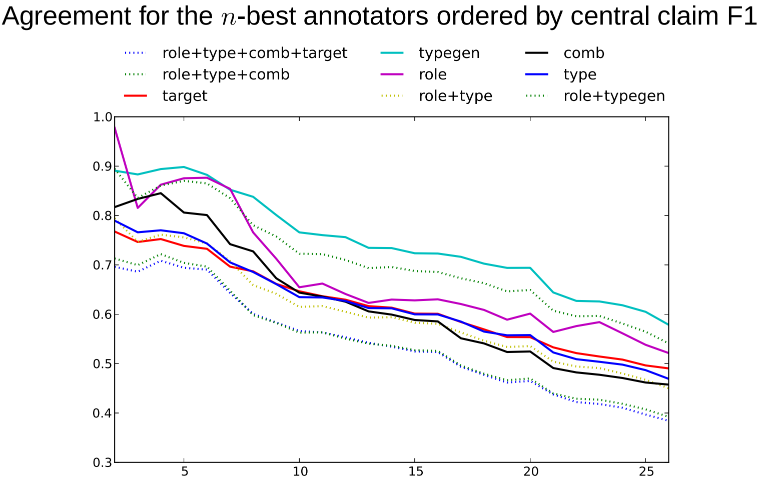
Questions:

- What range of agreement is possible in this group of annotators?
- How to give structure to this inhomogenous group of annotators?
- How to identify subgroups of good annotators, how to sort out bad ones without too much gold data?

Ranking by thesis F1

Ranking the annotators: by central claim F1

Ranking the annotators: by central claim F1



Ranking and clustering the annotators

Ranking by thesis F1

- still requires *some* gold data
- identifies bad annotators
- identifies good annotators

Ranking and clustering the annotators

Ranking by thesis F1

- still requires *some* gold data
- identifies bad annotators
- identifies good annotators

Ranking by $\Delta^{\varnothing}$ cat. distr.

Ranking the annotators: by $\Delta^{\varnothing}$ category distributions

Deviation from average category distribution: **no attacks, only support**

anno	PT	PSN	PSE	PAR	PAU	OSN	OSE	OAR	OAU	?	Δ^{gold}	$\Delta^{\varnothing}$
A01	23	40	5	13	0	6	0	24	0	4	17	15.6
A02	22	33	7	8	11	3	0	23	1	7	17	16.9
A03	23	40	6	4	12	5	0	16	9	0	7	11.8
A04	21	52	6	1	0	0	0	14	11	10	25	20.5
A05	23	42	5	15	2	5	0	20	3	0	10	14.2
A06	24	39	6	6	9	7	0	15	9	0	7	10.9
A07	22	41	1	12	8	5	0	13	8	5	13	9.4
A08	23	35	6	6	14	6	1	17	7	0	9	13.3
A09	23	43	2	6	7	7	0	15	12	0	9	10.8
A10	23	51	3	3	4	8	0	8	15	0	21	21.2
A11	21	41	3	2	1	1	0	22	9	15	21	16.6
A12	23	42	6	15	5	3	0	13	4	4	13	11.7
A13	23	40	4	16	0	7	0	17	8	0	14	13.3
A14	19	33	6	10	4	4	0	11	8	20	26	20.2
A15	19	37	2	6	7	3	0	18	3	20	20	16.9
A16	20	31	4	7	10	7	0	14	5	17	22	16.9
A17	22	53	2	4	3	0	0	20	6	5	17	15.1
A18	23	51	5	0	0	34	1	0	1	0	39	40.4
A19	24	41	7	13	2	5	0	20	3	0	10	14.5
A20	21	41	4	0	1	2	0	31	5	10	22	18.2
A21	16	40	0	1	0	20	0	0	1	37	52	44.8
A22	22	34	7	5	10	6	0	17	9	5	12	10.3
A23	23	52	0	1	0	0	0	32	6	1	24	27.1
A24	23	41	6	6	9	5	0	22	3	0	4	11.8
A25	23	38	4	5	15	0	0	7	23	0	24	27.1
A26	23	44	5	8	4	4	0	21	3	3	9	10.2
$\varnothing$	22.0	41.3	4.3	6.7	5.3	5.9	0.1	16.5	6.6	6.3		
gold	23	42	6	6	8	5	0	19	6	0		

Ranking the annotators: by $\Delta^{\varnothing}$ category distributions

Deviation from average category distribution: **no proponent attacks**

anno	PT	PSN	PSE	PAR	PAU	OSN	OSE	OAR	OAU	?	Δ^{gold}	$\Delta^{\varnothing}$
A01	23	40	5	13	0	6	0	24	0	4	17	15.6
A02	22	33	7	8	11	3	0	23	1	7	17	16.9
A03	23	40	6	4	12	5	0	16	9	0	7	11.8
A04	21	52	6	1	0	0	0	14	11	10	25	20.5
A05	23	42	5	15	2	5	0	20	3	0	10	14.2
A06	24	39	6	6	9	7	0	15	9	0	7	10.9
A07	22	41	1	12	8	5	0	13	8	5	13	9.4
A08	23	35	6	6	14	6	1	17	7	0	9	13.3
A09	23	43	2	6	7	7	0	15	12	0	9	10.8
A10	23	51	3	3	4	8	0	8	15	0	21	21.2
A11	21	41	3	2	1	1	0	22	9	15	21	16.6
A12	23	42	6	15	5	3	0	13	4	4	13	11.7
A13	23	40	4	16	0	7	0	17	8	0	14	13.3
A14	19	33	6	10	4	4	0	11	8	20	26	20.2
A15	19	37	2	6	7	3	0	18	3	20	20	16.9
A16	20	31	4	7	10	7	0	14	5	17	22	16.9
A17	22	53	2	4	3	0	0	20	6	5	17	15.1
A18	23	51	5	0	0	34	1	0	1	0	39	40.4
A19	24	41	7	13	2	5	0	20	3	0	10	14.5
A20	21	41	4	0	1	2	0	31	5	10	22	18.2
A21	16	40	0	1	0	20	0	0	1	37	52	44.8
A22	22	34	7	5	10	6	0	17	9	5	12	10.3
A23	23	52	0	1	0	0	0	32	6	1	24	27.1
A24	23	41	6	6	9	5	0	22	3	0	4	11.8
A25	23	38	4	5	15	0	0	7	23	0	24	27.1
A26	23	44	5	8	4	4	0	21	3	3	9	10.2
∅	22.0	41.3	4.3	6.7	5.3	5.9	0.1	16.5	6.6	6.3		
gold	23	42	6	6	8	5	0	19	6	0		

Ranking the annotators: by $\triangle^{\varnothing}$ category distributions

Deviation from average category distribution: **missing annotations**

anno	PT	PSN	PSE	PAR	PAU	OSN	OSE	OAR	OAU	?	$\triangle^{gold}$	$\triangle^{\varnothing}$
A01	23	40	5	13	0	6	0	24	0	4	17	15.6
A02	22	33	7	8	11	3	0	23	1	7	17	16.9
A03	23	40	6	4	12	5	0	16	9	0	7	11.8
A04	21	52	6	1	0	0	0	14	11	10	25	20.5
A05	23	42	5	15	2	5	0	20	3	0	10	14.2
A06	24	39	6	6	9	7	0	15	9	0	7	10.9
A07	22	41	1	12	8	5	0	13	8	5	13	9.4
A08	23	35	6	6	14	6	1	17	7	0	9	13.3
A09	23	43	2	6	7	7	0	15	12	0	9	10.8
A10	23	51	3	3	4	8	0	8	15	0	21	21.2
A11	21	41	3	2	1	1	0	22	9	15	21	16.6
A12	23	42	6	15	5	3	0	13	4	4	13	11.7
A13	23	40	4	16	0	7	0	17	8	0	14	13.3
A14	19	33	6	10	4	4	0	11	8	20	26	20.2
A15	19	37	2	6	7	3	0	18	3	20	20	16.9
A16	20	31	4	7	10	7	0	14	5	17	22	16.9
A17	22	53	2	4	3	0	0	20	6	5	17	15.1
A18	23	51	5	0	0	34	1	0	1	0	39	40.4
A19	24	41	7	13	2	5	0	20	3	0	10	14.5
A20	21	41	4	0	1	2	0	31	5	10	22	18.2
A21	16	40	0	1	0	20	0	0	1	37	52	44.8
A22	22	34	7	5	10	6	0	17	9	5	12	10.3
A23	23	52	0	1	0	0	0	32	6	1	24	27.1
A24	23	41	6	6	9	5	0	22	3	0	4	11.8
A25	23	38	4	5	15	0	0	7	23	0	24	27.1
A26	23	44	5	8	4	4	0	21	3	3	9	10.2
∅	22.0	41.3	4.3	6.7	5.3	5.9	0.1	16.5	6.6	6.3		
gold	23	42	6	6	8	5	0	19	6	0		

Ranking and clustering the annotators

Ranking by thesis F1

- still requires *some* gold data
- identifies bad annotators
- identifies good annotators

Ranking by $\Delta^{\varnothing}$ cat. distr.

- no gold data required
- identifies outliers
- but beware: outliers could also be above average good annotators

Ranking and clustering the annotators

Ranking by thesis F1

- still requires *some* gold data
- identifies bad annotators
- identifies good annotators

Ranking by $\Delta^{\varnothing}$ cat. distr.

- no gold data required
- identifies outliers
- but beware: outliers could also be above average good annotators

Clustering by agreement

Clustering the annotators

Agglomerative hierarchical clustering:

- initialize clusters as singletons for each annotator
- while $|\text{clusters}| > 1$ do:
  - calc κ for all pairs of clusters
  - merge cluster pair with highest agreement

Clustering the annotators

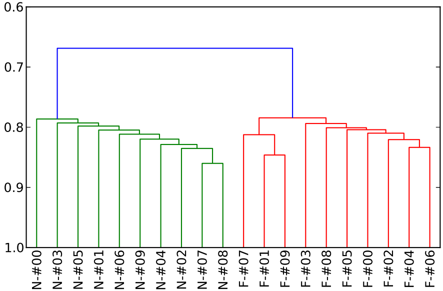
Agglomerative hierarchical clustering:

- initialize clusters as singletons for each annotator
- while $|\text{clusters}| > 1$ do:
    - calc κ for all pairs of clusters
    - merge cluster pair with highest agreement

Clustering the annotators

Agglomerative hierarchical clustering:

- initialize clusters as singletons for each annotator
- while $|\text{clusters}| > 1$ do:
  - calc κ for all pairs of clusters
  - merge cluster pair with highest agreement

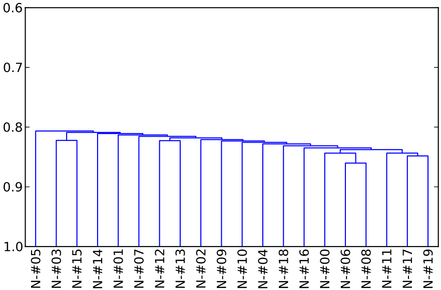


simulation: noise and systematic differences

Clustering the annotators

Agglomerative hierarchical clustering:

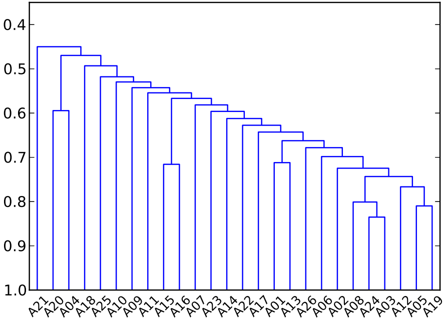
- initialize clusters as singletons for each annotator
- while $|\text{clusters}| > 1$ do:
    - calc κ for all pairs of clusters
    - merge cluster pair with highest agreement



simulation: noise but no systematic differences

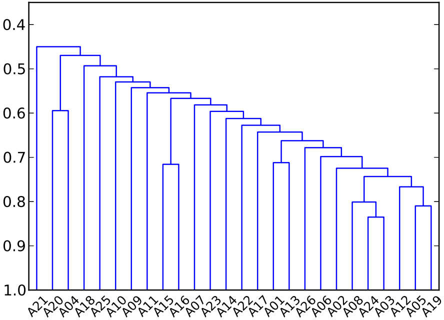
Clustering the annotators: Results for ‘role+type’

- linear growth, no strong clusters
- range from κ=0.45 to κ=0.84
- conforms with central claim ranking in picking out the same set of reliable and good annotators
- conforms with both rankings in picking out similar sets of worst annotators



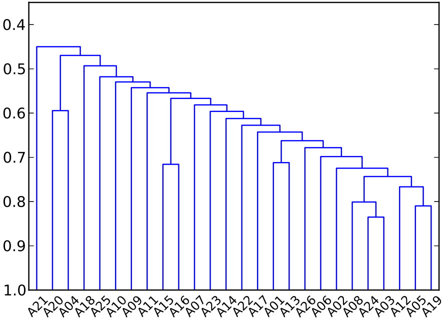
Clustering the annotators: Results for ‘role+type’

- linear growth, no strong clusters
- range from κ=0.45 to κ=0.84
- conforms with central claim ranking in picking out the same set of reliable and good annotators
- conforms with both rankings in picking out similar sets of worst annotators



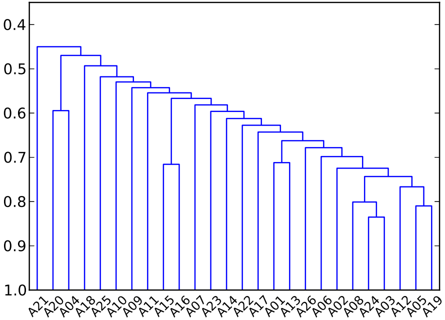
Clustering the annotators: Results for ‘role+type’

- linear growth, no strong clusters
- range from κ=0.45 to κ=0.84
- conforms with central claim ranking in picking out the same set of reliable and good annotators
- conforms with both rankings in picking out similar sets of worst annotators

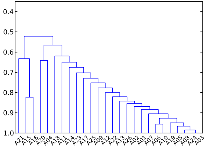


Clustering the annotators: Results for ‘role+type’

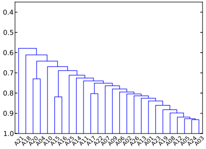
- linear growth, no strong clusters
- range from κ=0.45 to κ=0.84
- conforms with central claim ranking in picking out the same set of reliable and good annotators
- conforms with both rankings in picking out similar sets of worst annotators



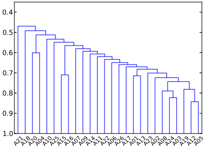
Clustering the annotators: Results for all levels



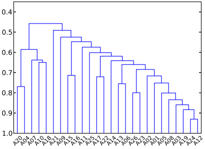
role



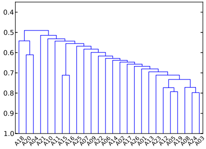
typegen



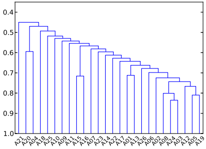
type



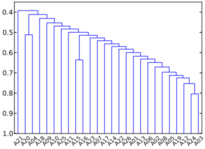
comb



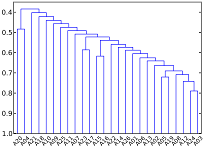
target



role+type



ro+ty+co



ro+ty+co+ta

Ranking and clustering the annotators

Ranking by thesis F1

- still requires *some* gold data
- identifies bad annotators
- identifies good annotators

Ranking by $\Delta^{\varnothing}$ cat. distr.

- no gold data required
- identifies outliers
- but beware: outliers could also be above average good annotators

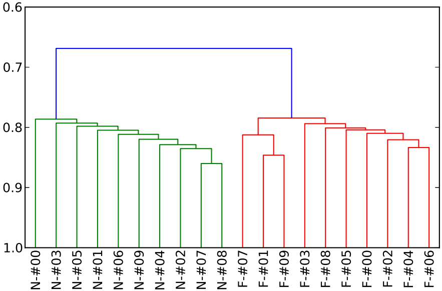
Clustering by agreement

- no gold data required
- identifies subgroups with characteristic annotation behaviour
- identifies good & bad annotators
- but beware: high agreement $\neq$ best annotators

Clustering the annotators: And then?

For ‘strong’ clusters pairs, investigate what makes them so different:

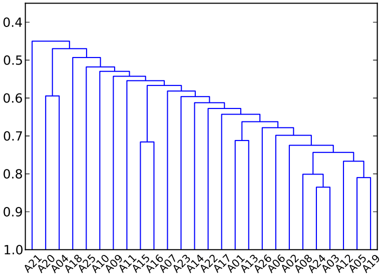
- compare their category distribution
- compare their typical confusions
- compare their Krippendorff diagnostics
- . . .



Clustering the annotators: And then?

For ‘steadily growing’ clusters:

- partial order on annotators by path from best to maximum cluster
- investigate confusion rate on the growing cluster path



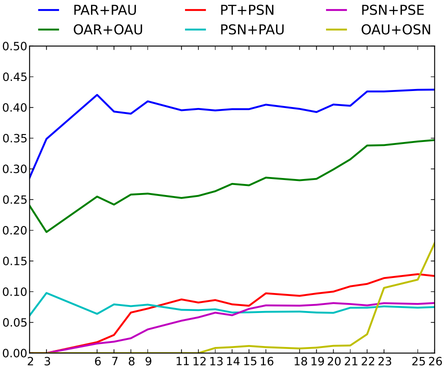
$$\text{conf}_{c_1,c_2} = \frac{|c_1 \circ c_2|}{|c_1 \circ c_1| + |c_1 \circ c_2| + |c_2 \circ c_2|}$$

Clustering the annotators: And then?

For ‘steadily growing’ clusters:

- partial order on annotators by path from best to maximum cluster
- investigate confusion rate on the growing cluster path

$$\text{conf}_{c_1,c_2} = \frac{|c_1 \circ c_2|}{|c_1 \circ c_1| + |c_1 \circ c_2| + |c_2 \circ c_2|}$$



Conclusions

- analyse the possible interpretations of the guidelines in a fine-grained manner by using more annotators
- learn about the task difficulty
- identify subgroups of good & reliable annotators, even if overall agreement is dissatisfactory

Thank You!

Conclusions

- analyse the possible interpretations of the guidelines in a fine-grained manner by using more annotators
- learn about the task difficulty
- identify subgroups of good & reliable annotators, even if overall agreement is dissatisfactory

Thank You!

Literatur I

Vikas Bhardwaj, Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. Anveshan: a framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 47–55, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Silvie Cinková, Martin Holub, and Vincent Kríž. Managing uncertainty in semantic tagging. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 840–850, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

James B. Freeman. *Dialectics and the Macrostructure of Argument*. Foris, Berlin, 1991.

James B. Freeman. *Argument Structure: Representation and Theory*. Argumentation Library (18). Springer, 2011.

Klaus Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA, 1980.

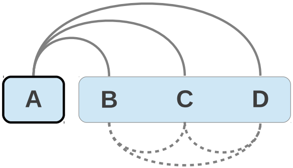
Vikas C. Raykar and Shipeng Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13:491–518, 2012.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Evaluation: Krippendorff’s Category Definition Test

Krippendorff [1980] diagnostics:

- systematically compare agreement on the original tagset with that on a reduced tagset
- **category definition test**: one category of interest against the rest
- compare the resulting $\Delta\kappa$ values to see which category is distinguished better from the rest



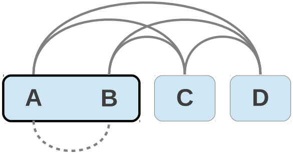
category	$\Delta\kappa$	A$_O$	A$_E$
PT	$+0.265$	0.91	0.69
PSE	$+0.128$	0.97	0.93
PSN	$+0.082$	0.79	0.54
OAR	-0.027	0.86	0.75
PAR	-0.148	0.92	0.89
OSN	-0.198	0.93	0.90
OAU	-0.229	0.92	0.89
PAU	-0.240	0.93	0.91

level ‘role+type’; base κ=0.45

Evaluation: Krippendorff's Category Distinction Test

Krippendorff [1980] diagnostics:

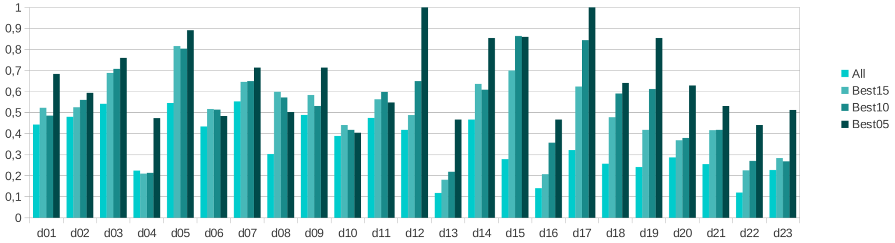
- systematically compare agreement on the original tagset with that on a reduced tagset
- **category distinction test**: only collapse one pair of categories
- $\Delta\kappa$ tells you how much you loose due to confusions between those two categories



category pair	$\Delta\kappa$	A_O	A_E
OAR+OAU	+0.048	0.61	0.22
PAR+PAU	+0.026	0.59	0.21
OAR+OSN	+0.018	0.58	0.22
PSN+PSE	+0.012	0.59	0.23
OAR+PAR	+0.007	0.58	0.22
PSN+OSN	+0.007	0.59	0.24
PAR+OSN	+0.005	0.57	0.21
...	...	...	...

level 'role+type'; base κ=0.45

Evaluation: Text-specific agreement



κ for the full task ('role+type+comb+target')

Scores for the 6-best annotators

	role+type	ro+ty+co+ta
∅F1	0.76	0.67
κ	0.74	0.69
α	0.83	0.73

	PT	PSN	PSE	PAR	PAU	OSN	OSE	OAR	OAU	?
PT	**0.915**	0.044	0.028	0.006	0.008	0.000	0.000	0.000	0.000	0.000
PSN	0.024	**0.843**	0.015	0.008	0.061	0.012	0.002	0.020	0.003	0.012
PSE	0.100	0.100	**0.800**	0.000	0.000	0.000	0.000	0.000	0.000	0.000
PAR	0.010	0.024	0.000	**0.432**	0.437	0.015	0.000	0.058	0.019	0.005
PAU	0.016	0.216	0.000	0.486	**0.189**	0.005	0.000	0.049	0.038	0.000
OSN	0.000	0.092	0.000	0.034	0.011	**0.667**	0.034	0.161	0.000	0.000
OSE	0.000	0.200	0.000	0.000	0.000	0.600	**0.000**	0.200	0.000	0.000
OAR	0.000	0.038	0.000	0.035	0.027	0.041	0.003	**0.593**	0.230	0.032
OAU	0.000	0.017	0.000	0.034	0.059	0.000	0.000	0.661	**0.229**	0.000
?	0.000	0.400	0.000	0.050	0.000	0.000	0.000	0.550	0.000	**0.000**

for the ‘role+type’-level