Ranking the annotators: An agreement study on argumentation structure

Andreas Peldszus
Applied Computational Linguistics
University of Potsdam
peldszus@uni-potsdam.de

Manfred Stede
Applied Computational Linguistics
University of Potsdam
stede@uni-potsdam.de

Abstract

We investigate methods for evaluating agreement among a relatively large group of annotators who have not received extensive training and differ in terms of ability and motivation. We show that it is possible to isolate a reliable subgroup of annotators, so that aspects of the difficulty of the underlying task can be studied. Our task is to annotate the argumentative structure of short texts.

1 Introduction

Scenarios for evaluating annotation experiments differ in terms of the difficulty of the task, the number of annotators, and the amount of training that annotators receive. For simple tasks, crowdsourcing involving very many annotators has recently attracted attention.[1] For more difficult tasks, the standard setting still is to work with two or a few more annotators, train them well, and compute agreement, usually in terms of the kappa measure. In this paper, we study a different scenario, which may be called 'classroom annotation': The group of annotators is bigger (in our example, 26), and there are no extensive training sessions: Students receive detailed written guidelines, there is a brief QA period, and annotation starts. In such a setting, one has to expect some agreement problems that are due to different abilities and different motivation of the students. Our goal is to develop methods for systematically studying the annotation results in such groups, to identify more or less competent subgroups, yet at the same time also learn about the difficulty of various aspects of the underlying annotation task. To this end, we investigate ways of ranking and clustering annotators.

[1] See, for instance, Snow et al. (2008) or Bhardwaj et al. (2010) for strategies to analyse and cope with diverging performance of annotators in that scenario.

Our task is the annotation of argumentation in short texts, which is somewhat similar to marking the rhetorical structure, e.g. in terms of RST (Mann and Thompson, 1988; Carlson et al., 2003). Thus we are dealing with a relatively difficult task involving text interpretation. We devised an annotation scheme (which is more fully described elsewhere), and in order to study the feasibility, first ran experiments with short hand-crafted texts that collectively cover all the relevant phenomena. This is the setting we report in this paper. A separate step for future work is guideline revision on the basis of the results, and then applying the scheme to authentic argumentative text (e.g., user generated content on various websites).

2 A theory of argumentation structure

Following up on Toulmin's (1958) influential analysis of argument, Freeman (1991; 2011) worked on integrating those ideas into the argument diagraming techniques of the informal logic tradition. Freeman's central idea is to model argumentation as a hypothetical dialectical exchange between a proponent, who presents and defends claims, and a challenger (the 'opponent'), who critically questions them in a regimented fashion. Every move in such a *basic dialectical situation* corresponds to a structural element in the argument diagram. The analysis of an argumentative text is thus conceived as finding the corresponding critical question of the challenger that is answered by a particular segment of the text.

Since the focus of this paper is on the evaluation methodology, we provide here only a brief sketch of the scheme; for a detailed description with many examples, see Peldszus and Stede (to appear). Premises and conclusions are propositions expressed in the text segments. We can graphically present an argument as an argument diagram, with propositions as nodes and the various relations as arrows linking either two nodes or

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 196–204,
Sofia, Bulgaria, August 8-9, 2013.

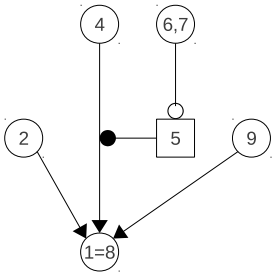

Figure 1: Example of an argumentation structure annotation for a short text

a node and a link[2]. See figure 1 for an example. Notice that segments in favor of the proponent's position are drawn in circles, whereas the challenger's perspective is given in boxes. The root of an argument tree is the central statement made in the text. In the example, it is expressed both in segment 1 and in segment 8; the = indicates that the annotator judges the contributions of the two segments as equivalent, which can happen for any node in the tree. Segments 2, 4, and 9 provide *support* to the central statement, which is the most simple configuration.

(1) [*We should tear the building down.*]$_1$ [*It is full of asbestos.*]$_2$

Support can be serial (transitive), when a supporting statement in turn receives support from another one. E.g., example (1) could be continued with ...[*The report of the commission made that very clear.*]$_3$.

If an argument involves multiple premises that support the conclusion only if they are taken together, we have a *linked* structure in Freeman's terminology. On its own none of the linked premises would be able to support the conclusion. In the basic dialectical situation, a linked structure is induced by the challenger's question as to why a premise is relevant to the claim. The proponent then answers by presenting another premise explicating the connection. Building linked structure is thus to be conceived as completing an argument. As an example, consider the following continuation of example (1) ...[*All buildings with hazardous materials should be demolished.*]$_3$. Linked support is shown in the diagram by connecting the premises before they link to the conclusion.

Two more configurations, which turn up in Figure 1, are the attacking relations (all with a circled arrowhead): *undercut* and *rebuttal*. The former (segment 5) denies the relevance of a stated relation, here: the support that 4 lends to 1=8. The opponent does not dispute the truth of 4 itself but challenges the idea that it can in fact lend support to 1=8. We draw it as an attack arrow pointing at the relation in question. In contrast, a rebuttal directly challenges the truth of a statement. In the example, the annotator first decided that segments 6 and 7 play a joint role for the argumentation (this is the step of *merging* two segments) and then marked them as the proponent's rebuttal of the challenger's statement 5.

3 Annotation Experiment

3.1 Guidelines

We developed annotation guidelines based on the theory presented in Section 2. The guidelines (6 pages) contain text examples and the corresponding graphs for all basic structures, and they present different combinations of attack and counter-attack. The annotation process is divided into three steps: First, one segment is identified as the central claim of the text. The annotator then chooses the dialectical role (proponent or opponent) for all remaining segments. Finally, the argumentative function of each segment (is it supporting or attacking) and the corresponding subtypes have to be determined, as well as the targeted segment.

3.2 Data

Applying the scheme demands a detailed, deep understanding of the text, which is why we choose to first evaluate this task on short and controlled instances of argumentation. For this purpose we built a set of 23 constructed German texts, where each text consists of only five discourse segments. While argumentative moves in authentic texts are often surrounded by material that is not directly relevant to the argumentation, such as factual background information, elaborations or rhetorical decoration, in the constructed texts all segments are clearly argumentative, i.e. they either presents the central claim, a reason, an objection or a counter-attack. Merging segments and identifying restatements is thus not necessary. The texts cover several combinations of the basic constructs in different linearisations, typically one central claim, two (simple, combined or exemplifying) premises, one objection (rebutting a premise, rebutting the conclusion or undercutting the link be-

[2]When an artificial node is introduced in such places, a standard tree representation results.

tween them) and a possible reaction (rebutting or undercutting counter-attacks, or a new reason that renders the objection uncountered). A (translated) example of a micro text is given in (2). In the questionaire the order of the texts has been randomized.

(2) [*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.*]$_5$

3.3 Procedure

The annotation experiment was carried out in the context of an undergraduate university course with 26 students, participation was obligatory. The annotators only received minimal training: A short introduction (5 min.) was given to set the topic. After studying the guidelines (∼30 min.) and a very brief question-answering, the subjects annotated the 23 texts (∼45 min.), writing their analysis as an argumentative graph in designated areas of the questionaire.

4 Evaluation

4.1 Preparations

Since the annotators were asked to assign one and only one function to each segment, every node in the argumentative graph has exactly one out-going arc. The graph can thus be reinterpreted as a list of segment labels.

Every segment is labeled on different levels: The 'role'-level specifies the dialectical role (proponent or opponent). The 'typegen'-level specifies the general type, i.e. whether the segment presents the central claim (thesis) of the text, supports or attacks another segment. The 'type'-level additionally specifies the kind of support (normal or example) and the kind of attack (rebutter or undercutter). Whether a segment's function holds only in combination with that of another segment (combined) or not (simple) is represented on the 'combined'-level.[3] The target is finally specified by the segment identifier ($1 \ldots 5$) or relation identifier ($a \ldots d$) on the 'target'-level.

The labels of each separate level can be merged to form a complex tagset. We interpret the result as a hierarchical tagset as it is presented in Figure 2.[4] The label 'PSNC(3)' for example stands for a proponent's segment, giving normal support to segment 3 in combination with another segment, while 'OAUS(b)' represents an opponent's segment, undercutting a relation b, not combined.

Due to space and readability constraints, we focus the detailed discussion of the experiment's result on the 'role+type'-level. Still, general results will be reported for all levels.

Another question that arises before evaluation, especially in our setting, is how to deal with missing annotations, since measuring inter-annotator agreement with a κ-like coefficient requires a decision of every annotator (or at least the same number of annotators) on each item. One way to cope with this is to exclude annotators with missing annotations, another to exclude items that have not been annotated by every subject. In our experiment only 11 of the 26 subjects annotated every segment. Another 10 annotated at least 90% of the segments, five annotated less. Excluding some annotators would be possible in our setting, but keeping only 11 of 26 is unacceptable. Excluding items is also inconvenient given the small dataset. We thus chose to mark segments with missing annotations as such in the data, augmenting the tagset with the label '?' for missing annotations. We are aware of the undesired possibility that two annotators 'agree' on not assigning a category to a segment. Still, we can decide to only exclude those annotators who omitted many decisions, and to measure agreement for the remaining ones, thereby reducing the risk of false agreement.

4.2 IAA over all annotators

The agreement in terms of Fleiss's κ (Fleiss, 1971)[5] of all annotators on the different levels is shown in Table 1. For the complex levels we additionally report Krippendorff's α (Krippendorff, 1980) as a weighted measure of agreement. We use the distance between two tags in the tag hierarchy to weigh the confusion (similar to Geertzen and Bunt (2006)), in order to capture the intuition that confusing, e.g., PSNC with PSNS is less severe than confusing it with OAUS.

According to the scale of Krippendorff (1980),

[3] This is roughly equivalent to Freeman's 'linked premises'.

[4] Notice that this hierarchy is implicit in the annotation process, yet the annotators were neither confronted with a decision-tree version nor the labels of this tag hierarchy.

[5] A generalisation of Scott's π (Scott, 1955) for more than two annotators, as Artstein and Poesio (2008) pointed out.

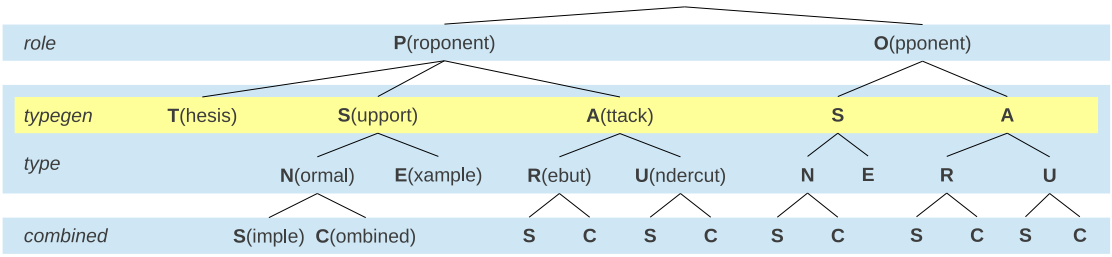

Figure 2: The hierarchy of segment labels.

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

Table 1: Agreement for all 26 annotators on 115 items for the different levels. The number of categories on each level (without '?') is shown in the second column (possible target categories depend on text length). We report Fleiss's κ with the associated observed (A_O) and expected agreement (A_E). Weighted scores were calculated using Krippendorff's α, with observed (D_O) and expected disagreement (D_E).

the annotators in our experiment did neither achieve reliable ($\kappa \geq 0.8$) nor marginally reliable ($0.67 \leq \kappa < 0.8$) agreement. On the scale of Landis and Koch (1977), most results can be interpreted to show moderate correlation ($0.4 < \kappa \leq 0.6$), only the two most complex levels fall out. Considering weighted scores for those complex levels, all fall into the window of moderate correlation.

While typical results in discourse structure tagging usually reach or exceed the 0.7 threshold[6], we expected lower results for three reasons: first the minimal training of the naive annotators only based on the guidelines, second the varying commitment to the task of the annotators in the constrained setting and finally the nature of the task, which requires a precise specification of the annotators interpretation of the texts.

When it comes to investigation of the reasons of disagreement, the informativeness of a single inter-annotator agreement value is limited. We want to identify sources of disagreement in both the set of annotators as well as the categories. To this end, contingency tables (confusion matrices) are studied, which show the number of category agreements and confusions for a pair of annotators. However, the high number of annotators in our study makes this strategy infeasible, as there are 325 different pairs of annotators. One solution to still get an overview of typical category confusions, is to build an aggregated confusion matrix, which sums up the values of category pairs across all 325 normal confusion matrices. As proposed in Cinková et al. (2012), we derive a confusion probability matrix from this aggregated matrix, which is shown in Table 2. It specifies the conditional probability that one annotator will annotate an item with $\text{category}_{\text{column}}$, given that another has chosen $\text{category}_{\text{row}}$, so the rows sum up to 1. The diagonal cells display the probability of agreement for each category.

cat.	$\Delta\kappa$	n	A_O	A_E
PT	+0.265	572	0.91	0.69
PSE	+0.128	112	0.97	0.93
PSN	+0.082	1075	0.79	0.54
OAR	−0.027	430	0.86	0.75
PAR	−0.148	173	0.92	0.89
OSN	−0.198	153	0.93	0.90
OAU	−0.229	172	0.92	0.89
PAU	−0.240	138	0.93	0.91
OSE	−0.451	2	0.99	0.99

Table 3: Krippendorff's category definition diagnostic for the level 'role+type', base κ=0.45.

[6]Agreement of professional annotators on 16 rhetorical relations was κ=0.64 in the beginning and 0.82 after extensive training (Carlson et al., 2003). Agreement on 'argumentative zones' is reported κ=0.71 for trained annotators with detailed guidelines, another study for untrained annotators with only minimalistic guidelines reported values varying between 0.35 and 0.72 (depending on the text), see Teufel (2010).

	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

Table 2: Confusion probability matrix over all 26 annotators for the level 'role+type'.

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

Table 4: Krippendorff's category distinction diagnostic for the level 'role+type', base κ=0.45.

Krippendorff (1980) proposed another way to investigate category confusions by systematically comparing the agreement on the original category set with the agreement on a reduced category set. There are two different methods to collapse categories: The first is the *category definition test*, where all but the one category of interest are collapsed together, yielding a binary category distinction. When measuring the agreement with this binary distinction only confusions between the category of interest and the rest count, but no confusions between the collapsed categories. If agreement increases for the reduced set compared to the original set, that category of interest is better distinguished than the rest of the categories. As Table 3 shows, the highest distinguishability is found for PT, PSN and PSE. Rebutters are better distinguished for the opponent role than for the proponent role. Undercutters seem equally problematic for both roles. The extreme value for OSE is not surprising, given that this category was not supposed to be found in the dataset and was only used twice. It shows, though, that the results of this test have to be interpreted with caution for rare categories, since in these cases the collapsed rest always leads to a very high chance agreement.

The other of Krippendorff's diagnostics is the *category distinction test*, where two categories are collapsed in order to measure the impact of confusions between them on the overall agreement value. The higher the difference, the greater the confusion between the two collapsed categories. Table 4 shows the result for some category pairs. The highest gain is found between rebutting and undercutting attacks on the opponents side: Given the base κ=0.45, the +0.048 increase means a potential improvement of 10% if these confusions could be reduced. However, distingishing rebutters and undercutters often depends on interpretation and we consider it unlikely to reach perfect agreement on that decision.

4.3 Comparison with gold data

We now compare the result of the annotation experiment with the gold annotation. For each annotator and for each level of annotation, we calculated the F1 score, macro-averaged over the categories of that level. Figure 3 shows the distribution of those values as boxplots. We observe varying degrees of difficulty on the basic levels: While the scores on the 'role' and 'typegen' are relatively dense between 0.8 and 0.9, the distribution is much wider and also generally lower for 'type', 'comb' and 'target'. Especially remarkable is the drop of the median when comparing 'typegen' with 'type': For the simpler level, all values of the better half of annotators lie above 0.85, but for the more complex level, which also requires the distinction between rebutters and undercutters, the median drops to 0.67. The figure also shows the pure F1 score for identifying the central claim (PT). While the larger part of the annotators performs well in this task, there are still some below 0.7. This is remarkable, since identifying one segment as the central claim of a five-segment text does not appear to be a challenging task.

4.4 Ranking and clustering the annotators

Until now we have mainly investigated the tagset as a factor in measuring agreement. The widespread distribution of annotator scores in the comparison with gold data however showed that

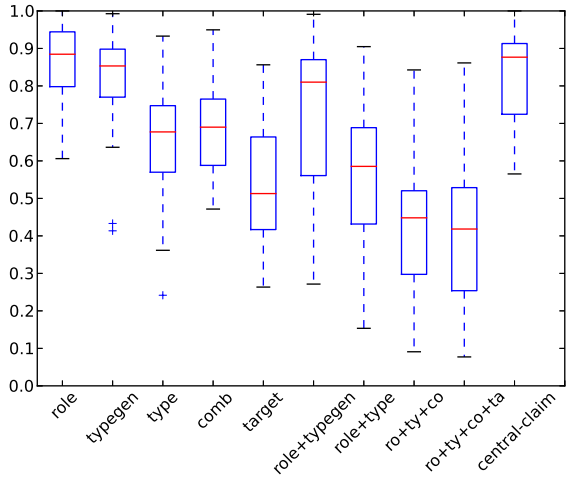

Figure 3: Comparison with gold annotation: For each level we show a boxplot of the F1 scores of all annotators (each score macro-averaged over categories of that level). Also, we present the F1 score for the recognition of the central claim.

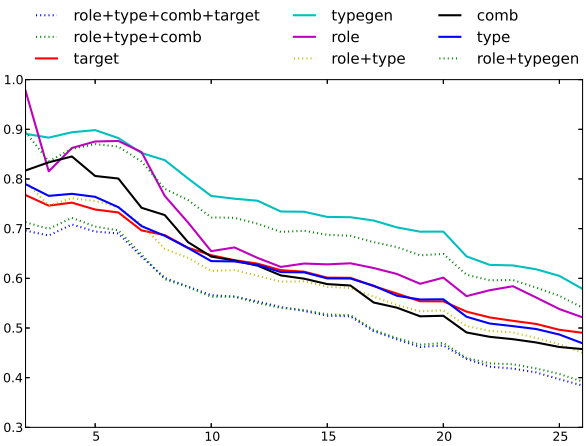

Figure 4: Agreement in κ on the different levels for the n-best annotators ordered by their F1 score in identifying the central claim.

their performance differs greatly. As described in Section 3.3, participation in the study was obligatory for our subjects (students in class). We thus want to make sure that the differences in performance are a result of the annotator's varying commitment to the task, rather than a result of possible ambiguities or flaws of the guidelines. The inter-annotator agreement values presented in Table 1 are not so helpful for answering this question, as they only provide us with an average measure, but not with an upper and lower bound of what is achievable with our annotators. Consequently, the goal of this section is to give structure to the set of annotators, to impose a (partial) order on it or even divide it into different groups and investigate their characteristic confusions.

Central claim: During the conversion of the written graphs into segment label squences, it became obvious that certain annotators nearly always chose the first segment of the text as the central claim, even in cases where it was followed by a consecutive clause with a discourse marker. Therefore, our first heuristic was to impose an order on the set of annotators according to their F1 score in identifying the central claim. This not only identifies those outliers but can additionally serve as a rough indicator of text understanding. Although this ordering requires gold data, producing gold data for the central claim of a text is relatively simple and using them only gives minimal bias in the evaluation (in contrast to e.g. 'role+type' F1 score as a sorting criterion). With this ordering we can then calculate agreement on different subsets of the annotators, e.g. only for the two best annotators, for the ten best or for all. Figure 4 shows κ on the different levels for all n-best groups of annotators: From the two best to the six best annotators the results are quite stable. The six best annotators achieve an encouraging κ=0.74 on the 'role+type' level and likewise satisfactory κ=0.69 for the full task, i.e. on the maximally complex 'role+type+comb+target' level. For increasingly larger n-best groups, the agreement decreases steadily with only minor fluctuations. Although the central claim F1 score proves to be a useful sorting criterion here, it might not work as well for authentic texts, due to the possibility of restated, or even implicit central claims.

Category distributions: Investigating the annotator bias is also a promising way to impose structure onto the group of annotators. A look on the individual distribution of categories per annotator quickly reveals that there are some deviations. Table 5 shows the individual distributions for the 'role+type'-level, as well as the average annotator distribution and that found in the gold data. We focus on three peculiarities here. First, both annotators A18 and A21 refrain from classifying segments as attacking. Although they make the distinction between the roles, they give only supporting segments. Checking the annotations shows that they must have mixed the concepts of dialectical role and argumentative function. Another example is the group of A04, A20 and A23, who refrain from using proponent attacks. Al-

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		

Table 5: Distribution of categories for each annotator in absolute numbers for the 'role+type' level. The last two rows display gold and average annotator distribution for comparison. The two rightmost columns specify for each annotator the total difference to gold or average distribution $\Delta^{gold/\varnothing} = \frac{1}{2}\sum_c \Delta_c^{gold/\varnothing}$.

though they make the distinction between the argumentative functions of supporting and attacking, they do not systematically attribute counter-attacks to the proponent. Finally, as pointed out before, there are several annotators with a different amount of missing annotations. Note, that missing annotations must not necessarily signal an unmotivated annotator (who skips an item if deciding on it is too tedious). It could very well also be a diligent but slow annotator. Still, missing annotations lead to lower agreement in most cases, so filtering out the severe cases might be a good idea. Most of the annotators showing deviations in category distribution could be identified, if annotators are sorted by deviation from average distribution $\Delta^{\varnothing}$, which is shown in the last column of Table 5. Filtering out the 7 worst annotators in terms of $\Delta^{\varnothing}$, the resulting κ increases from 0.45 to 0.54 on the 'role+type'-level, which is nearly equal to the 0.53 achieved when using the same size of annotator set in the central claim ordering. Although this ordering suffices to detect outliers in the set of annotators without relying on gold data, it still has two drawbacks: It only maximizes to the average and will thus not garantuee best agreement scores for the smaller n-best sets. Furthermore a more general critique on total orders of annotators: There are various ways in which a group agrees or disagrees simultaneously that might not be linearized this way. Luckily, a better solution is at hand.

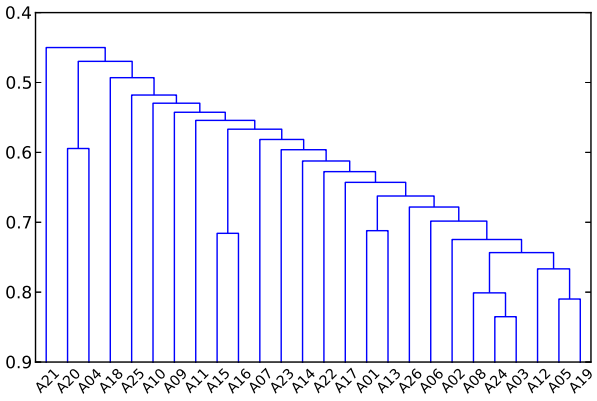

Figure 5: Clustering of the annotators (on the x-axis) for the 'role+type' level. The y-axis specifies the distance between the clusters, i.e. the κ reached by the annotators of both clusters.

Agglomerative hierarchical clustering: We apply hierarchical clustering in order to investigate the structure of agreement in the set of annotators. The clusters are initialized as singletons for each annotator. Then agreement is calculated for all possible pairs of those clusters. The pair of clusters with highest agreement is merged. This procedure is iterated until there is only one cluster left. In contrast to normal clustering, the linkage

criterion does not determine the distance between complex clusters indirectly as function of the distance between singleton clusters, but directly measures agreement for the unified set of annotators of both clusters. Figure 5 shows the clustering on the 'role+type'-level. It not only gives an impression of the possible range of agreement, but also allows us to check for ambiguities in the guidelines: If there were stable alternative readings in the guidelines, we would expect multiple larger clusters that can only be merged at a lower level of κ. As the Figure shows, the clustering grows steadily, maximally incorporating clusters of two annotators, so we do not see the threat of ambiguity in the guidelines. Furthermore, the clustering conforms with central claim ordering in picking out the same set of six reliable and good annotators (with an average F1 of 0.76 for 'role+type' and of 0.67 for the full task compared to gold) and it conforms with both orderings in picking out similar sets of worst annotators.

With this clustering we now have the possibility to investigate the agreement for subgroups of annotators. Since the growth of the clusters is rather linear, we choose to track the confusion over the best path of growing clusters, i.e. starting from the best scoring {A24,A03} cluster to the maximal cluster. It would be interesting to see the change in Krippendorff's category distinction diagnostic for selected confusion pairs. However, this value not only depends on the amount of confusion but also on the frequency of that categories[7], which cannot be assume to be identical for different sets of annotators. We thus investigate the confusion rate conf_{c_1,c_2}, i.e. the ratio of confusing assigments pairs $|c_1 \circ c_2|$ in the total set of agreeing and confusing assignments pairs for these two categories:

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

Figure 6 shows the confusion rate for selected category pairs over the path from the best scoring to the maximal cluster. The confusion between rebutters and undercutters is already at a high level for the best six best annotators, but increases when worse annotators enter the cluster. A constant and relatively low confusion rate has PSN+PAU, which means that distinguishing counter-attacks from new premises is equally 'hard' for all annotators. Distinguishing normal and example support, as well as central claims and supporting segments is not a problem for the six best annotators. It becomes slightly more confusing for more annotators, yet ends at a relatively low level around 0.08 and 0.13 respectively. Confusing undercutters and support on the opponents side is only a problem of the low-agreeing annotators, the confusion rate is nearly 0 for the first 21 annotators on the cluster path. Finally note, that there is no confusion typical for the high-agreeing annotators only.

[7] 20% confusion of frequent categories have a larger impact on agreement than that of less frequent categories.

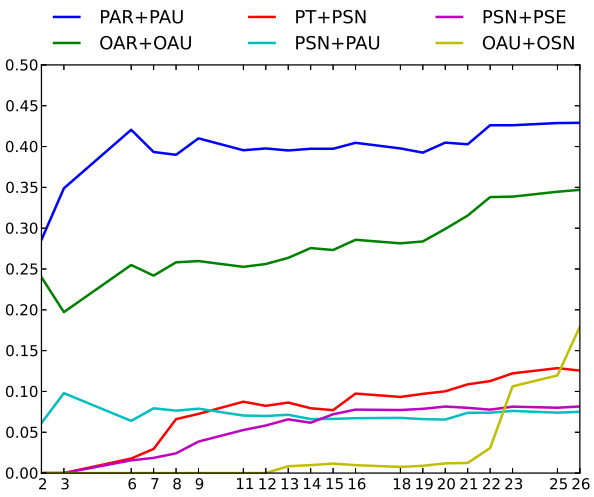

Figure 6: Confusion rate for selected category pairs in the growing clusters, with the numbers of annotators in the cluster on the x axis.

5 Conclusions

We presented methods to systematically study the agreement in a larger group of annotators. To this end, we evaluated an annotation study, where 26 untrained annotators marked the argumentation structure of small texts. While the overall agreement showed only moderate correlation (as one could expect from naive annotators in a text interpretation task) we could identify a subgroup of annotators reaching a reliable level of agreement and good F1 scores in comparison with gold data by different ranking and clustering approaches and investigated which category confusions were characteristic for the different subgroups.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. The first author was supported by a grant from Cusanuswerk and the second author by Deutsche Forschungsgemeinschaft (SFB 632).

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December.

Vikas Bhardwaj, Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. 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. Association for Computational Linguistics.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer, Dordrecht.

Silvie Cinková, Martin Holub, and Vincent Kríž. 2012. 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. Association for Computational Linguistics.

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

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

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

Jeroen Geertzen and Harry Bunt. 2006. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 126–133, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorial data. *Biometrics*, 33(1):159–174, March.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Andreas Peldszus and Manfred Stede. to appear. From argument diagrams to automatic argument mining: A survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1).

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. 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. Association for Computational Linguistics.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Studies in Computational Linguistics. CSLI Publications.

Stephen Toulmin. 1958. *The Uses of Argument*. Cambridge University Press, Cambridge.