Disagreement in Argumentation Annotation

Anna Lindahl
Språkbanken Text,
University of Gothenburg
Sweden
anna.lindahl@svenska.gu.se

Abstract

Disagreement, perspective or error? There is a growing discussion against the idea of a unified ground truth in annotated data, as well as the usefulness of such a ground truth and resulting gold standard. In data perspectivism, this issue is exemplified with tasks such as hate speech or sentiment classification in which annotators' different perspectives are important to include. In this paper we turn to argumentation, a related field which has had less focus from this point of view. Argumentation is difficult to annotate for several reasons, from the more practical parts of deciding where the argumentation begins and ends to questions of how argumentation is defined and what it consists of. Learning more about disagreement is therefore important in order to improve argument annotation and to better utilize argument annotated data. Because of this, we examine disagreement in two corpora annotated with argumentation both manually and computationally. We find that disagreement is often not because of annotation errors or mistakes but due to the possibility of multiple possible interpretations. More specifically, these interpretations can be over boundaries, label or existence of argumentation. These results emphasize the need for more thorough analysis of disagreement in data, outside of the more common inter-annotator agreement measures.

Keywords: annotation, disagreement, argumentation, aggregation, gold standard, inter-annotator agreement, argumentation mining

1. Introduction

Annotated data is needed in most NLP and machine learning tasks, often building upon the idea that phenomena can be consistently and uniformly labeled (Plank, 2022). However, annotation can be a complex task with several steps (Krippendorff, 2018; Artstein and Poesio, 2008) and it is often the case, especially the more subjective the task, that the annotators do not agree. Annotation disagreements or variation can be due to several reasons, such as an unclear or ambiguous task or annotator errors, but they can also be due to diverging opinions (Dumitrache, 2015; Uma et al., 2021b). Usually, these disagreements are disregarded, no matter their reason, and the annotations are aggregated using the majority vote for each annotation into a gold standard.

There is however a growing discussion concerning this practice, which argues that disagreements contain information which could (and should) be utilized (Uma et al., 2021b). For example, Plank et al. (2014) show that disagreement can be systematic and due to lingustically debatable cases rather than annotation error. Plank (2022) further argues that by assuming there exists a ground truth one misses information from disagreements, which can be due to subjectivity or multiple plausible answers. Mostafazadeh Davani et al. (2022) also discuss the issue of only using majority vote and present a model which learns from all annotations.[1]

A central concept in this discussion is data *perspectivism*,[2] (Cabitza et al., 2023; Basile et al., 2020), which argues that in highly subjective tasks (and many others) there isn't always one single truth or interpretation to be found in the data. For example, in tasks such as sentiment or hate speech classification, an annotator's ethnicity or social background might result in variation or disagreement between annotators (Akhtar et al., 2020). Disagreement or different perspectives could also arise due to ambiguity in language or to context Basile et al. (2021). Therefore, in order not to lose important information, all perspectives should be included in all steps when learning from (annotated) data, from using and sharing non-aggregated datasets to taking in multiple perspectives when evaluating (Basile et al., 2020, 2021).

An interesting example in this discussion is argumentation (annotation). Argumentation in itself is naturally full of perspectives and disagreement, which can spill over into the annotation and corpus creation process. In NLP, argumentation is often annotated with the intent of using it for argumentation mining, which aims to automatically identify and analyse argumentation (Lindahl and Borin, 2023). Considering this aim, including and representing all perspectives in argumentation should be relevant.

Annotating argumentation is challenging and time consuming. There is no uniform or widely accepted definition of argumentation(van Eemeren, 2017) which can make designing an annotation task non-trivial. Argumentation can also be context-

[1] It has also been the focus of two recent Semeval tasks (Leonardelli et al., 2023; Uma et al., 2021a).

[2] https://pdai.info/

dependent, ambiguous and complicated(Stede and Schneider, 2018), which can make reaching high agreement between annotators difficult. Identifying and analysing disagreements will thus not only help identify different perspectives but it will also be useful for developing better guidelines and tasks in the challenging field of argumentation annotation.

Despite these challenges, not much work has looked at disagreement in argumentation annotation in detail, or from the perspectivist point of view. Any study about argumentation annotation deals to some extent with disagreement in data, but usually with the purpose of finding a single ground truth or at least a way of creating a gold standard. An exception to this is the study by Hautli-Janisz et al. (2022), which presents a taxonomy of disagreement in their political debates corpus. Their corpus is annotated with argumentation, using argumentation graphs.

Because of the above mentioned challenges, in this paper we present further data on disagreement in argumentation annotation. Compared to Hautli-Janisz et al. (2022), our corpora is in the domain of social media, in the Swedish language. Our analysis and annotation schemes also differ. The contributions of this paper are:

- Our data add to the knowledge of disagreement in argumentation annotation, more specifically:
  - Examples of disagreement from social media
  - Examples of disagreement from Swedish language data
- A comparison of annotation disagreements to quantitative measurements

We do the above by showing a range of examples of (presumed) disagreement from two Swedish corpora annotated with argumentation. In our examples, we show that in most cases disagreement do not stem from one right and one wrong interpretation. Instead, much of the disagreement could be considered different variations of the same argument or different, but equally plausible, interpretations. This is followed by various measures examining the disagreement in the two corpora contrasting it to the quantitative analysis. The data presented is also followed by a short discussion of what these disagreements could mean for argumentation annotation.

2. Argumentation annotation

Argumentation is often annotated for the reason of argumentation mining or the related field of stance detection. Argumentation mining aims to identify not just our opinions but how we argue for them, and can include everything from classifying argumentation and its components to analysing argumentation strategies or and inferences (Lawrence and Reed, 2020; Stede and Schneider, 2018).

Argumentation is difficult to annotate for several reasons, as mentioned in the previous section. One reason for this is because there is no single definition of argumentation, and there might not be a definition which covers all purposes (van Eemeren, 2017). There are also several different argumentation models (Bentahar et al., 2010; Toulmin, 1958; Walton et al., 2008). Regardless of theoretical foundation, argumentation is complex and context-dependent, and often implicit (Lawrence and Reed, 2020; Lindahl and Borin, 2023). Annotators are commonly told to disregard their own opinions when annotating argumentation, but some argumentation might need domain-knowledge or expertise, and it might even be up to personal opinion. There can also be cases where there is more than one possible interpretation. Choosing what unit to annotate is also not straightforward - argumentation can stretch over several sentences or be contained in one phrase.

These difficulties are reflected in argumentation annotated corpora - many are not very large with moderate IAA [3] (Rosenthal and McKeown, 2012; Lippi and Torroni, 2016; Torsi and Morante, 2018; Wührl and Klinger, 2021). The many variants of annotation models, schemes and methods also make it difficult to compare corpora and studies, especially when datasets are often already curated and aggregated into a gold standard (Lindahl and Borin, 2023).

2.1. Disagreements in argumentation annotation

Although many works on argumentation annotation discuss (dis)agreement to some extent, they usually only report some IAA measure. The annotations are then aggregated using majority vote, or it might not even be reported how the gold standard was created.

However, there are examples of disagreement being treated differently. For example, Rosenthal and McKeown (2012) have their two annotators resolve their differences together when creating the gold standard. Haddadan et al. (2019) resolve differences in their annotations by having experts annotate a subset of their data. When curating the data the annotators who where agreeing the most with the experts were chosen in cases of disagreement. Toledo et al. (2019) remove judgments by annotators who have an average low agreement with the other annotators or have failed hidden test

[3] IAA should however not be seen as the only measure of quality.

questions (with predefined answers) in the annotation task. They also motivate the usefulness of their data despite the moderate agreement by showing it can be used for prediction successfully.

While there are alternative approaches to curating data, not as much work exists which analyse and discuss disagreement. Stab and Gurevych (2014) annotate the argumentation components claims and premises, and find that the most disagreement occurs between the two (as compared to occurrence of components). They find that this could be because some components can function both as claim and premise, depending on which argumentation the component belongs to. In Lindahl et al. (2019) similar patterns are found, where a component can be both a conclusion and a premise depending on the context. Teruel et al. (2018) analyse their annotation of ECHR judgments. They find agreement on what is argumentative but not on the components claims, premises and major claims. When analysing the disagreements they find, in short, that claims and premises presented as facts is the reason for some disagreements.

The previously mentioned Hautli-Janisz et al. (2022) present similar work to what is presented here. They investigate annotation of political debates with Inference Anchoring Theory (IAT) (Budzynska et al., 2014, 2016). Their annotation/analysis of the debates is done within the IAT framework, which includes segmenting the text into appropriate argumentative discourse units (called locutions) and their propositional content. These units are then used to build a directed graph which shows the argumentation structure and relations. They present a taxonomy of disagreement with three main categories: annotation error, fuzzy language and ambiguity. Annotation errors are annotations that don't agree with the guidelines, fuzziness refers to examples which can be "semantically and pragmatically fuzzy" and different interpretations occur because of underspecified language . Ambiguity refers to "clearly separate interpretations based on syntactic, (lexical) semantic or pragmatic ambiguity" (clearly separate interpretations). The categories also include subcategories.

When it comes to incorporating the different annotators' views into the learning process, there are are some but not many examples of perspectives being used in argumentation mining.[4] Romberg (2022) predicts concreteness and subjectivity, using both the hard labels from the data and a subjectivity score from the annotations. Furthermore, Heinisch et al. (2023) explore different ways of representing perspectives (from majority vote to isolating annotators) in an argument quality task and Van Der Meer et al. (2024) evaluate diversity in an argument summarizing task.

3. Case studies: two argumentation corpora

Below the two corpora discussed in this paper are described. The two corpora are annotated for argumentation or stance. Both corpora have spans as unit of annotation, decided independently by the annotators. This presumably leads to more disagreement, but it also gives us the most information about the annotators' opinions and variation compared to annotating more discrete units.

3.1. Political tweets

This corpus consists of 4,028 tweets from Swedish political parties and party leaders (in preparation). The tweets are annotated for positive and negative stance by four annotators, with each tweet being annotated by at least three annotators.

The annotators were first asked to determine if there was a positive or negative attitude expressed in the tweet (also phrased as if the tweeter was for or against something). If so, they should mark the object the attitude is about. The unit of annotation is spans, as an object of attitude can range from a single word ("littering") or noun phrase ("the sale of diesel cars") to longer spans such as sentences or tweets. The annotators were however instructed to annotate the shorter interpretation if in doubt and if possible to avoid longer spans. They were also told to annotate all instances of an attitude.

3.2. Online forums

This corpus consists of 9 threads from two Swedish online forums, about 28,500 tokens, annotated by 8 annotators (Lindahl, 2020). The annotators were asked to annotate spans of argumentation, given a definition of argumentation. They did not annotate any argumentation components or structure. Half of the annotators also gave a summary of each argumentation span they annotated, providing valuable insight in their perspectives[5].

4. Examples of disagreements

In this section examples of disagreement from the two corpora are shown. All examples are originally in Swedish. In the examples from the political corpus, positive spans are shown in **bold** and negative spans in *italics*. In the examples from the online

[4] More examples will surely come as there is a shared task for perspective argument retrieval in the argumentation mining workshop 2024: `https://blubberli.github.io/perspective-argument-retrieval.github.io/`

[5] The summaries are not included in the original paper

forum corpus spans of argumentation are shown underlined.

In the first example below we can see how the four annotators have annotated a tweet in the political tweets corpus. There are several interesting things to notice here. While a token comparison would indicate a high level of disagreement, we can see that the four of them do agree on "Centerpartiet", 'The center party', being described with a positive attitude (in **bold**). However, one of the annotators have chosen to include the full sentence where the word occurs ("Centerpartiet 11th of September"), which leads to token disagreement.[6] Three of them have also chosen to annotate "compassion" as positive, with two of them including "always", which also increases token disagreement.

A. To not *discriminate between people, distinguish them based on origin or faith,* that is a matter of showing respect. It is really quite simple. Always compassion. Never racism. Vote for **Centerpartiet** 11th of September. For Sweden's sake.

B. **To not discriminate between people, distinguish them based on origin or faith, that is a matter of showing respect.** It is really quite simple. **Always compassion.** Never *racism*. **Vote for Centerpartiet 11th of September.** For Sweden's sake.

C. To not *discriminate between people, distinguish them based on origin or faith,* that is a matter of **showing respect.** It is really quite simple. Always **compassion.** Never *racism.* Vote for **Centerpartiet** 11th of September. For Sweden's sake.

D. To not discriminate between people, distinguish them based on origin or faith, that is a matter of showing respect. It is really quite simple. Always **compassion.** Never *racism*. Vote for **Centerpartiet** 11th of September. For Sweden's sake.

The first sentence in the tweet displays a disagreement that might not be one. Annotator B has annotated "To not discriminate between people, distinguish them based on origin or faith, that is a matter of showing respect" as positive. Annotator A and C has instead chosen to exclude the initial "To not", resulting in a negative label. Both of these annotations could be considered correct as well as in some kind of agreement. This kind of issue also arises with terms such as "stop" ("stop the municipal crisis"), "prevent" ("prevent the climate crisis"). This was brought up before the main annotation round and the annotators were asked to not include the negative term in the annotation, but it might not have been easy to determine in some cases.

A shorter example of disagreement about what to include is seen below. All annotators agree that "solve the problems" is positive and two of them have annotated "not ignore them" as negative. Again, it is not obvious that either annotation is clearly wrong or right, or in conflict.

A. Let us **solve the problems.** *Not ignore them.*

B. Let us **solve the problems.** Not ignore them.

C. Let us **solve the problems.** Not ignore them.

D. Let us **solve the problems.** *Not ignore them.*

If we instead look at examples from the annotation of online forums, we can see similar examples of disagreement over boundaries, even if the task is slightly different. Spans annotated as argumentation are here marked in **bold**. In the example below, 7 out 8 annotators agree that "It is like encouraging a life as a housewife" is argumentation (the topic of the thread is home economics). Two of them have also included "And housewives do not belong in a society in the year 2020".

- 5 of 8: It is like encouraging a life as a housewife. And housewives don't belong in a society in the year 2020.

- 2 of 8: It is like encouraging a life as a housewife. And housewives don't belong in a society in the year 2020.

- 1 of 8: It is like encouraging a life as a housewife. And housewives don't belong in a society in the year 2020.

Three of the annotators wrote a summary for their annotations. One of them have chosen to motivate the argumentation using "it does not belong in the year 2020" even if the annotator did not include that in his or her span (this would maybe be a reconstruction error in Hautli-Janisz et al. (2022)).

In the next example we can also see that most annotators agree that the first sentence is argumentation, but only three of them have included the second sentence. Two have also chosen not to annotate at all.

- 3 of 8: Well Anders is an old man's name right now so I hardly think it would have been popular anyway. Today's celebrities will be long forgotten before it is popular again.

[6] One could of course argue that annotator B considers the positive attitude as referring to *voting* for the center party on the 11th of September, instead of the general positive attitude the other annotators presumably have inferred from the urging to vote message.

- 3 of 8: Well Anders is an old man's name right now so I hardly think it would have been popular anyway. Today's celebrities will be long forgotten before it is popular again.
- 2 of 8: Well Anders is an old man's name right now so I hardly think it would have been popular anyway. Today's celebrities will be long forgotten before it is popular again.

Below is another example of how a post was annotated. A,F,G,H annotated only the underlined part. B annotated the whole post. C annotated the first part as one argument, and the second underlined partas another argument. D also annotated the post as two arguments but split between the arguments at the last sentence. E did not annotate the post at all.

> I agree. Young kids can be a handful and tough on relations, yes. And to prefer one parent, is fully normal even if it of course is tough. What does the three-year old have to be thankful for? That he/she should be happy and grateful because you have "made a sacrifice" and moved to live with them is to complicated and too much to ask of a three-year old regardless if he/she likes to live with you.

F,G,H, who annotated the same span, summarised the argumentation:

- It's too much to ask to expect gratefulness because the child is three years old and has nothing to be grateful for.
- The three-year old can not be expected to be grateful because it is too complicated and to much to ask of a three-year old.
- A three-year old does not need to be grateful, he/she is too small to understand what you have "sacrificed".

In the summaries we can see that even if the annotators have annotated the exact same parts, they interpret the argumentation slightly differently - there is no reason for the three year old to be grateful compared to that there is a reason to be grateful, but the three year old cannot understand it. The variation in the summaries is similar to the "fuzziness" disagreement in Hautli-Janisz et al. (2022), more specifically the subcategory "fuzzy reconstruction".

These examples show some broad trends in disagreements (disregarding disagreements from errors). These are :

1. Disagreement over boundaries – what to include
2. Disagreement over what to annotate – existence of argumentation
3. Disagreement over positive or negative label

We have seen examples of 1 in both corpora. This might indicate that there is some agreement over some minimal unit of argumentation, but not where it starts or ends. Examples of annotators summarising the annotations including parts they did not mark in their spans might also indicate that these boundaries are not set in stone. There are however examples where different boundaries could result in slightly or very different interpretations, even if no example of the latter was shown here.

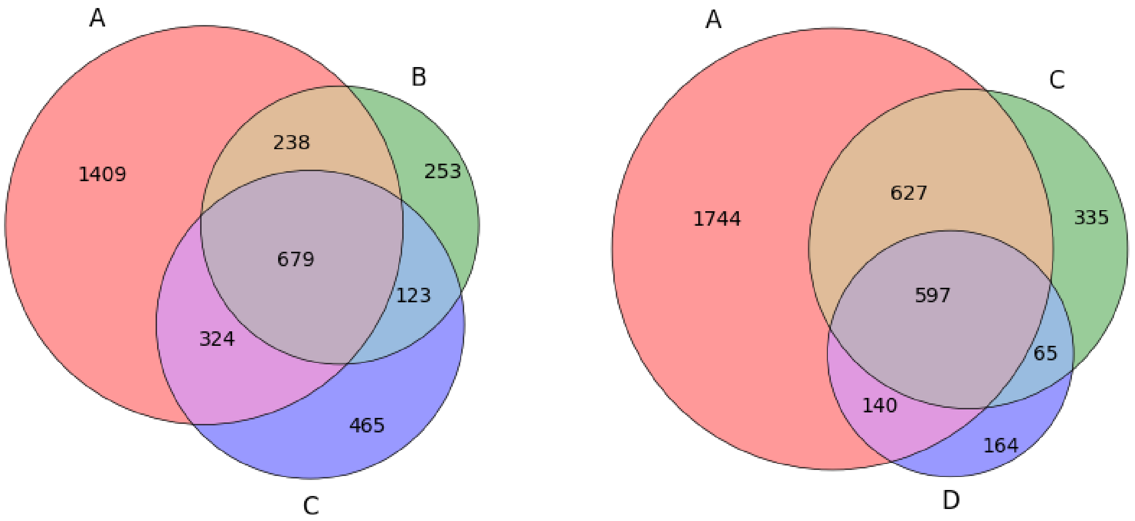
We can see an example of 2 in the first example. This might be due to different viewpoints or perspectives in the annotators. In the absence of annotation it is difficult to make any conclusions about why an annotator has chosen or not chosen to annotate, expect that an annotator has not considered the text argumentation. However, during discussions with the annotators, examples which one annotator had annotated as argumentative and the others had not were brought up. The divergent annotator would often have the others agree with him or her. It might not be the case that they strictly don't agree on argumentation they have left out to annotate but instead that they focus on different things in the text.

The third disagreement category, disagreement over positive or negative label, can be a "real" disagreement. But it can also depend on what was included in the annotated span, as we have seen. All three disagreements could also of course indicate some problem in the annotator guidelines.

5. Disagreement in numbers

Can one assume that these examples of disagreements are representative for all the annotations? Is it possible to find these kinds of disagreements computationally? We can find some clues if we look at the annotators. We can see differences in how much the annotators have annotated. Table 1 shows annotator statistics from the political tweets corpus. A has annotated more, both in spans and tokens, meaning A probably disagrees with the others over existence of argumentation. However, the proportion between negative and positive spans is similar to the other annotators. A has also shorter spans on average than the others, something which could indicate differences in splitting up argumentation as shown the previous section (disagreement over boundaries).

We can see differences between the annotators in the online forum corpus as well (table 2), with the number of annotated tokens ranging between



(a) ABC

(b) ADC

Figure 1: Overlapping spans for annotators ABC and ACD

	A	B	C	D
annot. spans	10304	5098	5254	3600
avg spans/tweet	3.2	1.9	2	1.3
avg span length	4	6	6	6
nr of pos spans	7185	3450	3735	2384
nr of neg spans	3119	1648	1519	1216
% tokens annot.	42%	29%	31%	21%
% tweets annot.	95%	81%	80%	84%

Table 1: Annotator statistics - political tweets

Annotator	no. arg. spans	no. arg. tokens	% of tokens anno.	avg. no. sent/arg span
A	135	9346	46%	4.45
B	174	11721	57%	4.40
C	81	6049	30%	5.11
D	109	6755	33%	4.14
E	75	2094	10%	1.87
F	141	5704	28%	2.60
G	167	1257	61%	4.92
H	134	7118	35%	3.39

Table 2: Annotator statistics - online forum (Lindahl, 2020)

10 to 57%. Annotator E has annotated a lot less than the others, which might indicate actual error or misunderstanding of the task. Note also that C and D have annotated roughly the same number of tokens but not the same number of spans, which might indicate more agreement than seen in the numbers. Thus, comparing number of tokens and units annotated between annotators might hint that the disagreement is over boundaries or over existence of argumentation.

With the differences in amount of tokens annotated, the IAA measures (table 3 and 4) are, as expected, low to moderate (Krippendorff's α, K-α)(Landis and Koch, 1977).

	K α	% agreement
	Tokens	Tokens
All	0.4	0.57
ABCD	0.36	0.46
ABC	0.46	0.63
ABD	0.39	0.58
ACD	0.36	0.53
BCD	0.42	0.6

Table 3: IAA for tweets

	K-α	% agreement
Tokens	0.30	25
Sents	0.36	40

Table 4: IAA for online forum

There are however differences between the annotators - some agree more than others. In table 3, we can see that the 'ABC' combination agree more than 'ACD'. Likewise, Cohen's κ pairwise between the annotators (tokens) vary from 0.49 (A &B) to 0.30 (A & D). In the online forum it varies from 0.57 (A & B) (or 0.55 B & H) to 0.14 (D & E). Note that using tokens or sentences for IAA is only one way of measuring agreement, as shown in the examples in the previous section, where the annotators sometimes agree on a part of the same span.

This partial agreement might indicate that there is some consistent overlap between some of the annotators even if they don't agree on the boundaries. In the political tweets corpus, we find that the majority of the spans overlap with at least one other annotator. In figure 1a, overlaps between spans among the three annotators with the highest K-alpha is shown (ABC). Annotator A has annotated the most spans, and most of the spans from the other two annotators overlap with A's. B and C do not overlap as much with each other. The overlaps between the annotator combination (ACD) with the lowest k-alpha is shown in figure 1b. Although the number of overlapping spans between all annotators is greater in figure 1a than in figure 1b, annotator A's spans overlap with more spans individually in figure 2. The other annotator combinations show similar patterns (see appendix A).

Tag combination	% of total tokens
O,O,O	48
O,O,POS	16
O,POS,POS	11
POS,POS,POS	8
O,O,NEG	8
NEG,NEG,O	5
NEG,NEG,NEG	3
NEG,POS,O	1
NEG,POS,POS	1
NEG,NEG,POS	0.3

Table 5: Distribution of tag combinations

If we instead look at the labels in the political tweets corpus, we can see that despite the example of the label changing depending on span length, the agreement is high. About 10% of tokens were annotated with either a positive or negative label, and the observed agreement is 92% and K-α is 0.86. This indicates that the annotators agree on what is negative and positive. The most common disagreement is instead between no label and the positive label, followed by no label and negative. Disagreement over existence or boundaries of positive spans seems to be more difficult than negative spans. This can be seen in table 5. This table shows the distribution of the tag combinations for all tweets which has been seen by three annotators, regardless of annotator identity.

6. Discussion

In comparing our disagreement categories to the categories in Hautli-Janisz et al. (2022) we can find both similarities and differences. Their first category, annotation errors w.r.t. the guidelines is difficult to compare against since our annotation schemes differ (annotation of spans compared to construction of argumentation graphs). As our guidelines allowed for any span length, we can't consider boundary disagreement as errors. While we do find some annotation errors in our data, they do not seem to be behind the disagreement examined so far. Annotation errors make up most of the disagreement in Hautli-Janisz et al. (2022). Our manual analysis does not look at as many examples as theirs, but it seems like disagreement over boundaries are more frequent.

Our first disagreement category, boundary disagreement, is similar both to the 'fuzziness' and 'ambiguity' category. Hautli-Janisz et al. (2022) distinguishes between the two by defining ambiguity as "those instances where a string yields two fully discrete discourse or argumentative structures" whereas fuzziness relates to language patterns common in natural language such as vagueness which "therefore result in different analyses which themselves are valid, but illustrate the uncertainty in representing partially underspecified or vague language." A disagreement in boundary could result in both separate and similar interpretations. Looking at the reformulations made by the annotators in the online forums corpus, it seems that they do interpret the argumentation similar but slightly different. This would mean that we found more fuzziness than ambiguity.

No matter the type of disagreement, dealing with disagreements require some kind of strategy. As mentioned in section 2.1, analysing and utilizing disagreements in argumentation corpora is usually disregarded in favor for majority vote, or some other aggregation method is used. It would perhaps make more sense, that in order to deal with disagreements one must first know what kind of disagreements there are. If the disagreements are actual annotation errors these should be dealt with accordingly. For example, there are methods for finding unreliable annotators (Hovy et al., 2013; Simpson and Gurevych, 2019).

However, as we have shown examples of here, disagreement in argumentation annotation is not always because of annotation errors but can be due to the possibility of several interpretations or boundaries. A more thorough analysis of the annotations, including both quantitative and qualitative aspects, instead of only relying on standard IAA measures could help identify disagreements. For example, the manual analyses we have shown here found that boundary disagreement wasn't necessarily wrong. A more liberal matching approach in combination with agreement measures could help with resolving and measuring such disagreement. Manual analysis could also identify specific disagreements like the effect inclusion of negation in a span has on disagreement. This possibly could be solved (or identified) by automatically inverting the negation in the text.

This still leaves the cases where there are different interpretations of the same argumentation, or cases where annotators have annotated varying number of argumentation. Assuming we want to keep all perspectives, we could resolve this by either *weak perspectivism*: creating a gold standard combining all voices in some way, or *strong perspectivism*: using the data from the annotators individually (Cabitza et al., 2023).

7. Conclusion and Outlook

In our examples, we have shown that not all disagreements in argumentation corpora are the same, and that not all of them should be considered disagreements but rather variation or perspectives. In order to determine what kinds of disagreement there are, IAA measures are not enough and a thorough look at the data is needed. This requires methodologies and research about disagreement in argumentation annotation. The development of taxonomies of disagreement specific to argumentation annotation, as in Hautli-Janisz et al. (2022), will also help categorizing disagreement. More research is needed on disagreement in argumentation corpora in order to find further patterns of disagreement or perspectives. An important part of this would be access to more non-aggregated datasets, which would enable more studies across argumentation domains and models. And finally, methods for learning from disagreement, such as soft loss (Uma et al., 2020) or labels (Fornaciari et al., 2021; Wu et al., 2023), is as far as we know a relatively unexplored area for argumentation annotated data and will surely give interesting results when applied.

8. Acknowledgements

The research presented here has been enabled by the Swedish national research infrastructure Nationella språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

9. Bibliographical References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

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

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Valerio Basile et al. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *Fifth International Conference on Computational Models of Argument*, pages 185–196. IOS Press.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint-Dizier. 2016. Theoretical foundations for illocutionary structure parsing. *Argument & Computation*, (1):91–108.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31–June 4, 2015. Proceedings 12*, pages 701–710. Springer.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of ACL 2019*, pages 4684–4690, Florence. ACL.

Annette Hautli-Janisz, Ella Schad, and Chris Reed. 2022. Disagreement space in argument analysis. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 1–9, Marseille, France. European Language Resources Association.

Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

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

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Anna Lindahl and Lars Borin. 2023. Annotation for computational argumentation analysis: Issues and perspectives. *Language and Linguistics Compass*, 18(1):e12505.

Anna Lindahl, Lars Borin, and Jacobo Rouces. 2019. Towards assessing argumentation annotation - a first step. In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186, Florence. ACL.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Sara Rosenthal and Kathy McKeown. 2012. Detecting opinionated claims in online discussions. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37.

Edwin Simpson and Iryna Gurevych. 2019. A Bayesian approach for sequence tagging with crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1093–1104, Hong Kong, China. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014*, pages 1501–1510. ACL.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*. Morgan & Claypool, San Rafael.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena

Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of LREC 2018*, pages 4061–4064, Miyazaki. ELRA.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment-new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635.

Benedetta Torsi and Roser Morante. 2018. Annotating claims in the vaccination debate. In *Proceedings of the 5th Workshop on Argument Mining*, pages 47–56, Brussels. ACL.

Stephen Edelston Toulmin. 1958. *The use of argument*. Cambridge University Press, Cambridge.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Michiel Van Der Meer, Piek Vossen, Catholijn Jonker, and Pradeep Murukannaiah. 2024. An empirical analysis of diversity in argument summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2028–2045, St. Julian's, Malta. Association for Computational Linguistics.

Frans H. van Eemeren. 2017. Rhetoric and argumentation. In Michael J. MacDonald, editor, *The Oxford Handbook of rhetorical Studies*, pages 661–672. Oxford University Press, Oxford.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.

Ben Wu, Yue Li, Yida Mu, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Don't waste a single annotation: improving single-label classifiers through soft labels. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5347–5355, Singapore. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. ACL.

10. Language Resource References

Lindahl, Anna. 2020. *Annotating argumentation in Swedish social media*. ACL.

A. Overlap between annotators

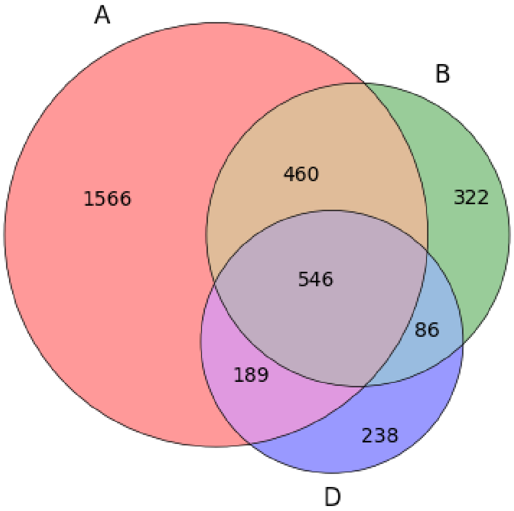


Figure 2: Overlapping spans for annotators ABD

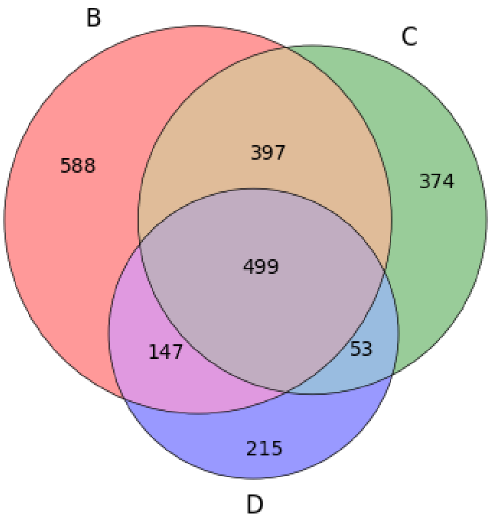


Figure 3: Overlapping spans for annotators BCD