Proceedings of ISA-18 Workshop at LREC2022, pages 71–75
Marseille, 20 June 2022

Do machines dream of artificial agreement?

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

Abstract

In this paper the (assumed) inconsistency between F1-scores and annotator agreement measures is discussed. This is exemplified in five corpora from the field of argumentation mining. High agreement is important in most annotation tasks and also often deemed important for an annotated dataset to be useful for machine learning. However, depending on the annotation task, achieving high agreement is not always easy. This is especially true in the field of argumentation mining, because argumentation can be complex as well as implicit. There are also many different models of argumentation, which can be seen in the increasing number of argumentation annotated corpora. Many of these reach moderate agreement but are still used in machine learning tasks, reaching high F1-score. In this paper we describe five corpora, in particular how they have been created and used, to see how they have handled disagreement. We find that agreement can be raised post-production, but that more discussion regarding evaluating and calculating agreement is needed. We conclude that standardisation of the models and the evaluation methods could help such discussions.

Keywords: annotation, inter-annotator agreement, argumentation mining, machine learning

1. Introduction

Most tasks in natural language processing require datasets annotated with some information, preferably of high quality, to learn from. The quality of such datasets is often measured by how well the annotators agree on the phenomenon being annotated: an inter-annotator agreement (IAA). The intuition behind this is that if a certain number of people agree upon something then the annotations represent some knowledge which can deemed more reliable, and thus it will be easier for a machine-learning algorithm to learn from the data.

However, in many tasks reaching high IAA is difficult, especially in more complex and possibly more subjective areas. In the field of argumentation mining, which aims to automatically identify and analyze argumentation, this is especially true. Many datasets annotated with argumentation report lower IAA than other tasks in natural language processing. This raises the question of what do with datasets in which the agreement is lower, can this be solved and will they still be useful?

In argumentation mining, there are several examples of corpora which have an IAA on the lower side, but still have been proven useful (that is good results[1]) in machine learning tasks. These results might also indicate that the current measurements of agreement might not be suitable for our tasks and that the agreement measures themselves can be difficult to interpret, something which has been discussed in Artstein and Poesio (2008).

Therefore, in this paper we describe some of these argumentation corpora, in order to explore how the agreement in these corpora has been tackled or how the (assumed) inconsistency can be explained. These corpora where selected as they are all within the same task in argumentation mining and all report moderate agreement but high F1-scores.

First, we give the field of argumentation mining a short introduction. Then, the argumentation corpora are described followed by discussion.

2. Argumentation mining & annotation

Argumentation mining is a relatively young field which aims to develop methods and datasets for automatically identifying argumentation. This is a challenging task, as argumentation can be complex and often implicit. How to annotate argumentation is also a challenge in itself, because argumentation does not have a unified definition which can be applied in all cases or an agreed upon way of modelling it (Van Eemeren et al., 2019; Habernal and Gurevych, 2017).

Nonetheless, the argumentation mining process is often described similarly – first identify the argumentative text, then the argumentation components such as claims and premises. After this step, relations between components and the arguments themselves can be annotated (for example attack or support) (Palau and Moens, 2009; Peldszus and Stede, 2013; Stab and Gurevych, 2017). There are also approaches focusing on argument quality (El Baff et al., 2018) or inferences in argumentation (Visser et al., 2018).

The agreement is often calculated using Cohen's κ, Fleiss' κ or Krippendorff's α, all of them measuring agreement (disagreement) by taking into account agreement (or disagreement) by chance. Values below 0 indicates agreement less than the chance agreement, and 1 indicates perfect agreement. Values be-

[1] What is "good" machine learning results can of course also be up for discussion but we leave that for another paper.

tween 0–1 are usually interpreted using the Landis & Koch scale, which says that results between 0.41–0.60 are moderate and 0.61–0.80 are substantial. As discussed in Artstein and Poesio (2008), the suitability of these measurements for linguistic annotation is not always clear. Duthie et al. (2016) raise the issue of using Cohen's κ when evaluating argumentation, and suggest the CASS-κ technique, however this has not been widely adopted.

The variety in how argumentation is modelled means there is also a great variety in how argumentation is annotated and how it is evaluated (see for example Lawrence and Reed (2020) or Habernal and Gurevych (2017)). This can make it difficult to compare results and datasets, even within similar tasks.

In this paper we focus on corpora annotated with the argumentation components claims and premises, but as we shall see there are variations of how to describe these components. There are also other examples of moderate IAA and higher machine learning results in other areas of argumentation mining Ajjour et al. (2017; Boltužić and Šnajder (2014).

3. Corpora annotated with claims and premises

In this section we will describe examples of argumentation annotated datasets. For each dataset, we will describe the data, annotation scheme and evaluation. Then, the results from a machine learning experiment using the same corpus as training data will be described. The datasets are also described in table 1.

A relatively early (with respect to the field of argumentation mining) argumentation annotated corpus was created by (Rosenthal and McKeown, 2012). This corpus consists of 4,000 sentences, half taken from blogposts from LiveJournal and half from discussions from Wikipedia debate forums. These sentences were annotated, without context, for presence of an opinionated claim. The definition of a claim was that "a claim is a statement that is a belief that can be justified". Two annotators annotated 2,000 sentences from each source, and the agreement is reported as 0.5 Cohen's κ for 633 blogpost sentences and 0.56 Cohen's κ for 997 Wikipedia sentences. The final gold standard was created by the annotators discussing and resolving all their disagreements. The final corpus has a ratio of 60–40% claims–non-claims for the blogposts and 64–36% for Wikipedia.

(Rosenthal and McKeown, 2012) then use logistic regression together with various features such as part of speech, sentiment and punctuation. They run experiments on both balanced and unbalanced versions of the two corpora. The best results on a balanced, combined, version of the corpus is 68.8% accuracy. Interestingly, when training on one domain and applying it to the other, the highest accuracy is achieved, between 74–76% for balanced, and 75–83% for unbalanced datasets.

(Teruel et al., 2018) perform an annotation study in which two annotators annotate major claims, claims and premises, and relations (attack or support) between them, in 7 judgments from the European Court of Human Rights (28,000 words). In their annotation study, they present a methodology for improving annotation guidelines. They loosely follow Toulmin (2003) when defining their components, adapting them according to Stab and Gurevych (2015). They define major claim as: "a general statement expressing the author's stance with respect to the topic under discussion", claim as: "a controversial statement whose acceptance depends on premises that support or attack it" and premise as: "reasons given by the author for supporting or attacking the claims"

While they find that annotators agree whether a sentence contains a span which represents an argumentation component or not (0.77–0.84 Cohen's κ), they agree less on which component is in the span (0.48–0.56 Cohen's κ). The corpus contained 3.0% major claims, 18.2% claims, 26.6% premises and 52.2% non-argumentative components. Noting that there is high disagreement in the major claim category, the authors merge these categories. This increases the agreement to 0.51–0.64 Cohen's κ.

There is no report on how the gold standard was created, but they report that using the system developed by Eger et al. (2017), the automatic classifier makes more mistakes in the categories in which humans disagree more. The corpus is also used for classification by Frau et al. (2019), who report that they use a version of the corpus with only claims and premises. They use a BiLSTM architecture with attention for two tasks: in a paragraph, detect which tokens are part of a claim or not, and likewise, in a paragraph, detect claims and premises. For the first task, they reach an F1-score of 0.82 and the latter 0.68.

In (Haddadan et al., 2019), a corpus of transcripts from US presidential debates between 1960–2016 is presented. The corpus consists of 6601 turns of dialogue, made up of about 34,000 sentences. The debates are annotated with claims and premises, where examples of claims are policies advocated for, judgments about other parties and candidates, stances on controversial subjects or opinions on issues, and premises are "are assertions made by the debaters for supporting their claims (i.e., reasons or justifications)". Three non-expert annotators annotated the debates, and the IAA was determined on a subset of 19 debates which were annotated by all three annotators. IAA was 0.57 κ[2] for sentences containing an argumentation component or not, and 0.4 κ for argumentation components. In order to create a gold-standard, two expert annotators annotated a subset of 6 debates. When resolving disagreement between the non-expert annotators, the annotator which had the highest agreement with the experts were chosen. The resulting corpus has 16,087 claims and

[2]The kind of κ is not reported.

13,434 premises.
This corpus was then used for two classification tasks – argumentation detection and argumentation component detection. The best results for both tasks came from using an LSTM, 0.84 for the first task and 0.67 F1 for the second.
(Schaefer and Stede, 2020) annotate 300 tweet-reply pairs, where the first tweet is seen as a context to the tweet which replies to it. The tweets are in German and all contain the German word for climate. The tweets were annotated with claim and evidence. A claim is described as a standpoint to a topic being discussed, while evidence is a statement which is used to support a standpoint. There were two annotators, and they reached an IAA of 0.55 Cohen's kappa for if a tweet contained a claim and 0.44 for if a tweet contained an evidence. They found that 14% of the tweets contained no argumentation component, 27% contained one argumentation component and 59% more than one component. Of the ones that contained one component, claim was the dominating component. How the gold standard is created is not reported, but the corpus is used for classification. Using different models, they achieve an F1-score of 0.82 for determining if a tweet contains an argument component, and 0.82 and 0.67 for if a tweet contained a claim or a premise, respectively. When classifying spans in tweets in a sequence labelling approach, the F1 for argumentation is 0.72, 0.59 for claims and 0.75 for evidence. Despite evidence having the lowest IAA, the sequence labeling approach worked best for that category.
(Wührl and Klinger, 2021) also annotate tweets, but in English and in the biomedical domain. Their corpus consists of 1200 tweets collected based on keywords from the medical domain. They annotate claims in the tweets following Stab and Gurevych (2017), describing claims as the central component of an argument in which the arguer expresses their conclusion. The claims are further annotated as explicit or implicit. Two annotators annotated the tweets, with 100 of the tweets being annotated by both annotators. The IAA was 0.56 κ for claim or not claim, and 0.48 κ for claims as implicit or explicit. About 44% of the tweets contained a claim.
There is no mention of how the gold standard was produced, but the corpus is used for claim classification. The best macro F1 results are reached with logistic regression for claims–non-claims and is 0.73 for explicit and implicit claims, or no claim, the macro F1 is 0.54 using a pipeline approach. They also report using their twitter corpus as training data and test it on a persuasive essays corpus (Stab and Gurevych, 2017), reaching 0.83 for the claims class.

4. Discussion and Outlook

As we have seen there are several ways the annotation and evaluation of argumentation components can be carried out, as well as machine learning applications. There are also different strategies in solving disagreement.
Most of the above studies revised their annotation and discussed the guidelines in order to increase the agreement and three of the mentioned studies took measures in order to increase the IAA after the annotation. (Rosenthal and McKeown, 2012) had their annotators solve their disagreement between themselves, which resulted in good machine-learning results. Likewise, (Haddadan et al., 2019) solved the disagreement using expert annotators, also reporting good machine learning results. However, while we can assume the expert annotators to be more in agreement with each other (Bayerl and Paul, 2011), we do not know by how much.
(Teruel et al., 2018) merge their annotation categories and increase their IAA, but not to a substantial level. Still, their machine learning results show that their corpus can be used for learning. Finally, for the two last studies we are not told how a gold standard was reached, but the machine learning task show that it is possible to learn from the data. Indeed, all studies described above show that it is possible to either solve disagreement in data or to learn from it anyway.
However, as previously mentioned, the measures of agreement can be difficult to interpret. But if we assume that the mentioned datasets all have not good agreement, we can think of a few, not mutually exclusive, explanations for the good machine learning results:

1. The corpus has been curated in such a way that the agreement has been raised.

2. What the machine-learning learns does not correspond to the original intention of the annotation.

3. The agreement measure is not representative of the "true" agreement.

4. High agreement is not needed in order to learn the task.

We have seen that 1. is indeed possible. Number 2 as an explanation could be due to anything between an unbalanced dataset (although all mentioned datasets here are fairly balanced.) to the machine learning algorithm picking up spurious cues (which relates to the whole field of blackbox nlp, see for example Niven and Kao (2019)). Number 3 would mean that there is agreement in the data which is not captured by the chosen IAA measure or that the scale for judging the IAA is not suitable. As previously stated, the suitability and interpretability of the different agreement measures has been discussed. However, they are still widely used as measure of quality, instead of for example percentage agreement. Perhaps, if the goal is to use the dataset for machine learning, the machine learning results could be included in evaluating the quality of a dataset.

Author	Size	IAA	F1-score
(Rosenthal and McKeown, 2012)	4,000 sent.	0.5–0.55 Cκ (subset)	68–80% accuracy (No F1 reported)
(Teruel et al., 2018)	28,000 words	arg sent.: 0.78–0.88 Cκ arg comp.: 0.48–0.56 Cκ	claim detection: 0.841 arg comp.: 0.704 (Frau et al., 2019)
(Haddadan et al., 2019)	34,013 sent.	arg. sent.: 0.57 κ arg. comp.: 0.4 κ	arg: 0.84 arg comp.: 0.67
(Schaefer and Stede, 2020)	300 tweets	claims: 0.55 Cκ evidence 0.37 Cκ	arg: 0.82 F1 claim detection: 0.82 premise detection: 0.67
(Wührl and Klinger, 2021)	1,200 tweets	claims: 0.56 Cκ explicit or implicit: 0.48 Cκ (subset)	claims: 0.70 non-claims: 0.76

Table 1: Argumentation corpora with moderate IAA.

Number 4 ties in to discussions in Uma et al. (2021) which discuss scenarios where there might be more that one possible interpretation of a gold label and how to learn from that. As two of the datasets provide the raw annotations, approaches mentioned Uma et al. (2021) would be also be a potential future research. Number 4 also raises the question of whether there is a lower limit of an IAA for the data to be useful.

All of above the explanations open up for more studies, but especially calls for more discussion of IAA in relation to machine learning results, such as the discussion in (Teruel et al., 2018).

To conclude, we have seen that it is possible to achieve good results on classification tasks even with lower IAA. This raises several interesting questions, such as what do the machine algorithm learn from or what is a sufficient IAA, but also highlights the need for discussing these issues. In particular, it calls for more discussion regarding agreement, how to calculate it and how to use it. It also shows the need for standardisation in many of the aspects in argumentation mining – the annotation, evaluation and use of datasets.

5. Bibliographical References

Ajjour, Y., Chen, W.-F., Kiesel, J., Wachsmuth, H., and Stein, B. (2017). Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark, September. Association for Computational Linguistics.

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

Bayerl, P. S. and Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Boltužić, F. and Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Duthie, R., Lawrence, J., Budzynska, K., and Reed, C. (2016). The cass technique for evaluating the performance of argument mining. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 40–49.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. In *Proceedings of ACL 2017 (Volume 1: Long Papers)*, pages 11–22, Vancouver. ACL.

El Baff, R., Wachsmuth, H., Al Khatib, K., and Stein, B. (2018). Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464.

Frau, J., Teruel, M., Alemany, L. A., and Villata, S. (2019). Different flavors of attention networks for argument mining. In *The Thirty-Second International Flairs Conference*.

Habernal, I. and Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, 4(1).

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

Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy, July. Association for Computational Linguistics.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of ICAIL 2009*, page 98, Barcelona. ACM Press.

Peldszus, A. and Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey.

Int. J. Cogn. Inform. Nat. Intell., 7(1):1–31.

Stab, C. and Gurevych, I. (2015). Guidelines for annotating argumentation structures in persuasive essays. *Ubiquitous Knowledge Processing Lab (UKP Lab) Computer Science Department, Technische Universität Darmstadt*.

Stab, C. and Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Van Eemeren, F. H., Grootendorst, R., and Kruiger, T. (2019). Handbook of argumentation theory. In *Handbook of Argumentation Theory*. De Gruyter Mouton.

Visser, J., John, L., Jean, W., and Chris, R. (2018). Revisiting computational models of argument schemes: Classification, annotation, comparison. In *Proceedings of the 7th International Conference on Computational Models of Argument (COMMA 2018)*.

6. Language Resource References

Haddadan, Shohreh and Cabrio, Elena and Villata, Serena. (2019). *Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates*. ACL.

Rosenthal, Sara and McKeown, Kathy. (2012). *Detecting Opinionated Claims in Online Discussions*.

Schaefer, Robin and Stede, Manfred. (2020). *Annotation and detection of arguments in tweets*.

Teruel, Milagro and Cardellino, Cristian and Cardellino, Fernando and Alonso Alemany, Laura and Villata, Serena. (2018). *Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines*. ELRA.

Wührl, Amelie and Klinger, Roman. (2021). *Claim Detection in Biomedical Twitter Posts*. Association for Computational Linguistics.