Annotating argumentation in Swedish social media

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

Abstract

This paper presents a small study of annotating argumentation in Swedish social media. Annotators were asked to annotate spans of argumentation in 9 threads from two discussion forums. At the post level, Cohen's κ and Krippendorff's α 0.48 was achieved. When manually inspecting the annotations the annotators seemed to agree when conditions in the guidelines were explicitly met, but implicit argumentation and opinions, resulting in annotators having to interpret what's missing in the text, caused disagreements.

1 Introduction

In recent years, argumentation mining has grown into a central research topic within the field of computational linguistics. With the aim of automatically identifying and analyzing argumentation in text, its envisioned applications are many, from more effective document retrieval to learning aids (Lawrence and Reed, 2020). There are many different approaches to how argumentation can be modeled and annotated and there are now many data sets of different size and level of annotation, with domains ranging from legal documents to social media. However, in many of the existing data sets, inter-annotator agreement is not very high and it is because annotating argumentation turns out to be a quite challenging task. There is still a need of more annotated data, as well as investigating how to reliably annotate data. It is also important to investigate other languages than English. Because of this, we have conducted a small annotation study on Swedish social-media data where the focus has been on identifying instances of argumentation but not analyzing them further.[1] This is both to select documents for further analysis of the identified argumentation instances but also in order to investigate how reliably annotators can agree on what is argumentation or not.

2 Related work

Annotating with the aim to distinguish what is argumentative from what is not argumentative has not been the most common goal in argumentation mining, although it is necessarily part of studies that annotate components of argumentation, either implicitly or explicitly as a first step in an argumentation mining pipeline. When it comes to documents from the web, the annotation of argumentation is usually done with respect to a topic. For example Habernal et al. (2014), annotated comments and blog posts as argumentative with respect to a topic in order to select documents for further annotation. On this they reach a 0.51 Fleiss κ and 0.59 Cohen's κ. Similarly, Habernal and Gurevych (2017) annotated documents from web discourse as 'non-persuasive' and 'on topic persuasive' before moving on to annotate micro-structure. They reached Fleiss κ of 0.59 on this task. In some studies presence of argumentation has been annotated together with the stance or the type of the argumentation. For example, Stab et al.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

[1] In the literature it seems that the assumption is made that argumentation is universally present in all languages and that its form is comparable across languages. This is obviously subject to empirical verification, but we have not seen any literature addressing this question. Impressionistically, descriptions of the kinds and structure of argumentation made for English seem to apply also to Swedish, but more thorough studies of this would be needed.

Proceedings of the 7th Workshop on Argument Mining, pages 100–105
Barcelona, Spain (Online), December 13, 2020.

(2018) annotated sentences from the web for supporting or opposing argument, or not an argument with respect to a topic. They reached Cohen's κ 0.72 and an observed agreement between 0.86–0.84. More recently, Trautmann et al. (2020) annotated sentences from the web with both expert and crowd-sourced annotators. The sentences were annotated with argument spans, and the spans were marked with stance with respect to a topic. The reached 0.6 Krippendorff's α_u and the crowd-sourced annotators reached 0.71 α_u.

3 Data

The data in this study is from two of Sweden's largest online discussion forums, Familjeliv ("Family life" FM), and Flashback (FB). Familjeliv is generally considered to be more about relationships and family life and Flashback more about politics, although both forums cover a broad range of topics. Both forums have a simple thread structure, where a thread is started with a post by a user and then other users reply with subsequent posts, shown in chronological order. There is a possibility for the users to cite each other, but there is no visually explicit tree structure as for example on Reddit. For this study, nine threads were randomly chosen among the threads which had a length of about 30 posts. These threads are shown in table 1. Threads 1–5 are from Familjeliv and threads 6–9 are from Flashback.

Thread	no. posts	no. users	no. tokens	no. cite tokens	tot no. tokens	Thread title
1	25	7	1426	562	1988	Corona at my kid's preschool
2	51	17	5795	442	6237	Thinking about cheating on my partner?
3	28	17	2627	45	2672	The stepchildren don't want to stay with us
4	20	8	1549	89	1638	To you who made the Trip, mainly Slovakia
5	33	25	1425	461	1886	Abolish home economics
6	32	28	1407	658	2065	The government wants to establish a new department ”for psychological defence”
7	22	12	2032	725	2757	Tehran vs. Pyongyang.
8	25	19	1442	822	2264	Who will name their son Anders in the future?
9	30	17	3589	3369	6958	It was right to keep the schools open
Tot	266	150	21292	7173	28465	

Table 1: Thread statistics

4 Annotation

4.1 Annotation guidelines and setup

We employed 8 annotators in this study: one expert (the author) and 7 with linguistic background. For the annotation, the annotation tool WebAnno (Eckart de Castilho et al., 2016) was used. The annotators were asked to annotate spans of argumentation, the spans could not overlap but otherwise there was no restriction on span length. Argumentation was only to be annotated within posts. The annotation guidelines[2] provide the annotators with a definition of argumentation, inspired by a simplified version of the definition given in Van Eemeren et al. (2013). The definition also includes persuasiveness, as this is a fundamental part of argumentation, as discussed in Habernal and Gurevych (2017) among others. The definition is seen below, and says that argumentation should include:

[2] Please note that the guidelines were written in Swedish, which means some of the nuances of the following descriptions might be lost in translation.

1. A standpoint/stance.
2. This standpoint is expressed with claims, backed by reasons.
3. There is a real or imagined difference of opinion concerning this standpoint which leads to:
4. the intent to persuade a real or imagined other part about the standpoint.

What is considered as argumentation or an argument in argumentation mining tasks varies and is often adjusted to fit the task or the domain, see for example Bosc et al. (2016) who annotated tweets containing opinions as arguments due to the implicit argumentation on Twitter. In some studies a definition of argumentation is not given, but rather definitions of what is being annotated, for example argumentative components such as premises or claims. The definition described here is not meant to cover all phenomena which could be considered argumentative, the intent is to describe something which hopefully annotators can apply successfully and agree on. From this definition above these three questions were derived:

- Does the poster's text signal that he or she is taking a stance / has a standpoint?
- Does the poster motivate why?
- Do you perceive the poster as trying to persuade someone?

If the annotator considered the answer to be affirmative for all the questions for some span of text, they were instructed to mark it as argumentation. In addition to these questions two tests were supplied in order to aid in answering the questions. The first test asked the annotator to reformulate the argumentation as "A, because B", in order to answer the first two questions. The second test asked the annotator to insert "I agree/I don't agree" into the text. If doing so would not change the meaning of the text, this might indicate that the poster is arguing, and was intending to persuade. These two tests were not meant to give a definite answer but rather to guide the annotators. The guidelines also included examples of argumentation from the forums, as well as examples on how to apply the tests. Four of the annotators were also asked to write down the reformulation of the "A because of B" test in the annotation tool. We've chosen to treat the results from the all the annotators equally in this study as we've yet to analyze the reformulations.

4.2 Annotation statistics

The annotators took between 4.5 and 12 hours each to annotate all the threads. The annotators which had to write down a reformulation took the longer time. Table 2 shows annotation statistics for each annotator. Annotator A is the expert annotator and seems not to diverge from the others. The annotators annotated mostly one argument per post, in some cases two arguments per post (compare number of arguments and number of posts in table 2). The annotators differ in how many argument spans they have annotated. The annotators also differ in how many sentences on average they have included in the argumentation spans, which is reflected in how many of the total tokens they have annotated. The annotators usually marked spans respecting sentence boundaries, but sometimes annotated half a sentence. When a post was annotated with a span, all but one annotator annotated at least half the post on average.

4.3 Inter-annotator agreement

When calculating inter-annotator agreement (IAA) sentences were considered as being argumentative if at least half of the tokens in it were labeled as argumentation, posts were considered as being argumentative if they contained at least one argument span. Observed agreement for tokens are 25%, for sentences 40% and 39% for posts. If we include posts where all but one annotator agree, observed agreement is 60% and if we include posts were all but two agree it's 86%. 70% of all the posts are labeled with an argument span by at least one of the annotators; 47% of those posts are annotated with a span by at least 6 of the annotators. Cohen's κ was measured pair-wise for all annotators and, as used in Toledo et al. (2019), averages from Cohen's κ were calculated and are shown in table 3. Annotator F has the highest

Annotator	no. arg spans	no. arg tokens	no. arg sents	no. of arg posts	% of tokens annotated	avg no. sent / arg span
A	135	9346	601	124	46%	4.45
B	174	11721	765	149	57%	4.40
C	81	6049	414	79	30%	5.11
D	109	6755	451	97	33%	4.14
E	75	2094	140	70	10%	1.87
F	141	5704	367	114	28%	2.60
G	167	12578	821	153	61%	4.92
H	134	7118	495	121	35%	3.39

Table 2: Annotation statistics for each annotator.

average κ, 0.55, and annotator E has the lowest average. Values between 0.21 and 0.40 are considered fair agreement, values between 0.41 and 0.61 are considered moderate agreement (Landis and Koch, 1977). Table 4 shows Krippendorff's α, for each thread and in total. α varies between threads. IAA is the highest for posts.

Annotator	A	B	C	D	E	F	G	H	Task-average
Average Cohen's κ - sents	0.44	0.42	0.30	0.33	0.30	0.38	0.38	0.35	0.35
Average Cohen's κ - posts	0.52	0.53	0.42.	0.46	0.39	0.55	0.48	0.52	0.48

Table 3: Average Cohen's κ for each annotator.

Krippendorff's α Thread	1	2	3	4	5	6	7	8	9	All
Tokens	0.311	0.187	0.419	0.118	0.355	0.358	0.31	0.166	0.166	0.296
Sents	0.365	0.22	0.434	0.112	0.486	0.462	0.398	0.299	0.327	0.356
Posts	0.525	0.363	0.676	0.425	0.437	0.412	0.573	0.309	0.369	0.482

Table 4: Krippendorff's α.

In order to compare the annotators observed agreement and α were calculated holding out each annotator. Holding out annotator E had the largest effect, changing observed agreement on post level from 0.39 to 0.45 and post level α from 0.48 to 0.52.

4.4 Analysis of the annotation results

A manual inspection of the annotation of posts was done on the two threads with highest α, thread 6 and 4, and the two threads with lowest α, threads 5 and 7. These four threads cover different topics, but the ones with lower α have fewer tokens and shorter posts. High agreement was deemed to be when 6 or more annotators agreed, otherwise the agreement was considered low. High agreement seemed to occur when the poster is very explicit with his or her opinion and writes it in terms of ”I” and not ”one”. Explicitly addressing a previous user, using confrontational language and contradicting also seems to occur within high agreement posts. Below is an example of a post were all annotators agreed it contained argumentation. The poster is clearly taking a stance, and is also signaling that they think the person they are addressing doesn't know what they are talking about.

> “So? And how do you think the children are feeling right now? That it's so hard to live with their dad that they'd rather refrain from doing it altogether? It doesn't matter that you thought it was boring to not live with your boyfriend. I agree with the others in this thread that you should stop living together. For the sake of the children. You can't just think of yourself.”

Disagreements between the annotators seemed to occur when a poster is not explicit with his or her stance or opinion, as well when the poster is using irony. Implicit argumentation (if there is any) such as that will force the annotator to interpret what's not being said in the text and this probably caused disagreement. General statements that are not tied explicitly to the opinion of the poster also seem to cause disagreements. The post below has a similar message as the previous example, but this poster is more sarcastic, and the argumentation is more implicit, if there is any. Here the annotators disagreed.

> "A three-year old should be grateful because you split up his parents? Oh my god! Are you for real?"

Another example of disagreement is seen in the post below where the user could be interpreted as speculating, rather than arguing.

> "The popularity of first names is varying over generations. Names that were popular in the 1900's first half such as Albin, Arvid etc ., have returned a bit. Names which were common a few decades ago, Johan, Andreas, Magnus, and Anders seem to have completely disappeared now. I think Anders is or was at least a few years ago the most common name for persons in high positions in the business world."

The guidelines asked for a stance or standpoint, which might be why posts where the author is clearly taking a stance have high agreement. The third condition, the intent to persuade, might be the reason posts with confrontational (and sometimes condescending) language have high agreement —if someone strongly disagrees with someone they might also intent to persuade them that they are wrong.

5 Conclusions & future directions

IAA values such as the ones reported here are not uncommon in argumentation mining tasks. Still, both the Cohen's κ of 0.48 and the Krippendorff's α 0.48 are lower than the previously reported studies, (for example 0.59 Cohen's κ in Habernal et al. (2014) or 0.71 Krippendorff's α_u in Trautmann et al. (2020)). However, as opposed to those studies, the annotators were not asked to annotate with respect to a topic, so the results are not fully comparable. Annotating only 9 threads might have affected the IAA, especially since the IAA varied between the threads. When manually inspecting the annotations, it seemed as when the conditions asked for in the guidelines were very explicitly met, annotators agreed. When the argumentation (or not argumentation) was more implicit the annotators disagreed. This is something which has to be considered when further developing the guidelines. Another thing to consider when annotating complex phenomena such as argumentation is that even though the annotators disagree, it might not be the case that one is right and the other is wrong. As shown in for example Lindahl et al. (2019) there are cases where two different annotations could both be considered correct. If one allows for several annotations to be correct, this would need to be reflected in both the guidelines and evaluation.

In the future we plan to test the guidelines in a domain where one can assume that people are more explicit with their argumentation, such as newspapers. We also plan to extend the guidelines to annotate components of argumentation to see how this affects the annotation.

Acknowledgements

The work presented here has been partly supported by an infrastructure grant to Språkbanken Text, University of Gothenburg, for contributing to building and operating a national e-infrastructure funded jointly by the participating institutions and the Swedish Research Council (under contract no. 2017-00626). We would also like to thank the anonymous reviewers for their constructive comments and feedback.

References

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: a dataset of arguments and their relations on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. 43(1):125–179.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *ArgNLP*.

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.

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, Italy, August. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *AAAI*, pages 9048–9056.

Frans H Van Eemeren, Rob Grootendorst, Ralph H Johnson, Christian Plantin, and Charles A Willard. 2013. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge.