Developing an argument annotation scheme based on a semantic classification of arguments

Lea Kawaletz, Heidrun Dorgeloh, Stefan Conrad and **Zeljko Bekcic**
Heinrich-Heine-University Düsseldorf, Germany
{lea.kawaletz, dorgeloh, stefan.conrad,zeljko.bekcic}@hhu.de

Abstract

Corpora of argumentative discourse are commonly analyzed in terms of argumentative units, consisting of claims and premises. Both argument detection and classification are complex discourse processing tasks. Our paper introduces a semantic classification of arguments that can help to facilitate argument detection. We report on our experiences with corpus annotations using a function-based classification of arguments and a procedure for operationalizing the scheme by using semantic templates.

1 Introduction

The corpus-based analysis of argumentative texts is a widely used discourse processing task needed both for an in-depth understanding of this basic discourse type, and in the field of argument mining. We here present an annotation scheme that has been developed as part of a project for gaining detailed insight into the linguistic features of arguments. These features can be used for machine learning as well as for the task of argument detection in the study of discourse and discourse processing.

In contrast to other approaches in the field, our method aims at the identification and classification of arguments, and not at the analysis of an overall argumentation structure (cf., for example, Peldszus et al. 2016). We argue that the annotation scheme will facilitate the annotation process in many applications of argument detection, enabling both researchers and annotators to zoom into linguistic characteristics that pertain to a specific class of arguments rather than to the notion of 'argument' as a whole. The approach therefore reduces some of the vagueness of the category of 'argument' and adds to the transparency of annotators' decisions.

Arguments are used for different purposes, aiming to persuade an addressee to believe, evaluate, or do something (see e.g. Eggs 2008; Stede and Schneider 2019). We use this functional versatility of arguments as a starting point for our annotation scheme. More precisely, we propose a systematic testing procedure during which annotators use a set of linguistic templates on a given text passage to determine whether it is an argument or not, and, if so, which argumentative function it has. We are currently developing and evaluating this approach with a corpus of COVID-19-related news opinion texts from *The New York Times*.

This paper is structured as follows. In section 2, we introduce the general idea of a function-based argument classification and briefly describe our corpus. In section 3, we present and evaluate our initial, rather ad hoc annotation efforts. In section 4, we introduce our function-based annotation scheme and report on our progress in terms of workflow and inter-annotator agreements. In section 5, we summarize our insights and provide an outlook.

2 Background

2.1 Arguments and argument categories

Theories of discourse generally claim that arguments do not have a particular linguistic form, but appear in all sorts of linguistic structures (e.g., Smith 2003; Virtanen 2010; Dorgeloh and Wanner 2010). Accordingly, the annotation of arguments in corpora is still a challenge because "a substantial amount of knowledge needed for the correct recognition of the argumentation, its composing elements and their relationships is not explicitly present in the text" (Moens 2018, 1; see also Lawrence and Reed 2020). Resulting from this difficulty, argument detection schemes so far often avoid cross-topic transfer (e.g., Nguyen and Litman 2015; Liebeck et al. 2016), but schemes for more heterogeneous corpora also exist (e.g., Stab et al. 2018; Cabrio and Villata 2018; Ein-Dor et al. 2020). Such work from argument mining typically relies on recurrent patterns identified by the NLP model used, but does not imply a systematic, truly topic-independent classification of arguments.

Argumentative discourse is characterized by presenting a central, disputed issue, the *major claim*, which the author argues for or against (Stab and Gurevych 2017). That is, they aim to persuade an addressee to believe and/or evaluate and/or do something, and they provide a number of arguments to this end (van Eemeren and Grootendorst 2004; Stede and Schneider 2019). This variability of what an argument is ultimately intended to do is often commented on in existing approaches, for example as an argument being either the expression of (positive or negative) stance, or of a policy or action to be taken (e.g., Hidey et al. 2017; Ein-Dor et al. 2020).

We suggest that this functional complexity of argumentation is exactly what is needed for the aim of developing a topic-independent classification scheme that can be applied to arguments as a whole. In the annotation scheme we developed, we distinguish between *epistemic*, *ethical* and *deontic* arguments, as first proposed by Eggs (2008; see Stede and Schneider 2019 for a summary in English). The three types are illustrated in Table 1.

Table 1: Argument categories

polarity	epistemic	ethical	deontic
positive	x is true	x is good	do x
negative	x is false	x is bad	don't do x

In addition to an argument being understood by its function, the most common definition is that it has two components, the *claim* and the *premise*. The claim is typically described as a controversial statement which provides the topic of the argument, and its premise is then a statement which provides evidence or expresses reasoning that either supports or attacks the claim (Stab et al. 2018). The link between a claim and its premise can thus be conceptualized as a directed argumentative relation, with a premise as the source and a (major) claim as its target (Stab and Gurevych 2014b). Each argument classified by our annotation scheme needs to have these two components expressed in the text.

2.2 Corpus compilation

Our corpus is currently being developed at Heinrich-Heine-University Düsseldorf ('HHU') as part of a collaborative project of both linguists and computer scientists working on argumentative discourse. So far, it consists of 25 COVID-19-related news opinion texts from *The New York Times* (29,466 words), and it will be consecutively expanded as the annotations progress. The corpus is designed to provide us with an inventory of arguments, divided into components and categorized by function and polarity. This inventory will first be used for linguistic analysis and, at a later stage, for an experiment with human subjects on argument-specific discourse relations, as well as for machine learning experiments.

3 The initial annotation process

Our first set of annotations ('set 1') was created before the introduction of our annotation scheme. Four annotators were instructed to apply a basic, simplified notion of 'argument,' consisting of a claim that is either supported or attacked. Practical issues of claim detection and annotation (e.g. size of the discourse unit, treatment of quotes within the texts) were discussed at regular meetings, leading the group from an initial, very thorough exemplary discussion of three texts (subset 1-1, 3,653 words) to an annotation of another ten texts in one hit (subset 1-2, 11,646 words). Annotations were created in the INCEpTION tool (Klie et al. 2018), hosted on a HHU server.

As the annotation task is not only a coding but also a unitizing task, we measure the inter-annotator agreement using Krippendorff's unitizing alpha (Krippendorff et al., 2016). This measure works with an arbitrary number of annotators (where not all have to annotate all texts) and determines the degree of observed disagreement in relation to the expected disagreement (assuming random annotations). Values range from -1 to 1, with values around zero representing random annotations, positive values representing more agreement among the annotators, and negative values representing more disagreement than expected by chance. The results for both subsets are displayed in Table 2.[1] We counted whether the annotators identified a given text passage as a *premise,* as a *claim,* or not as an argument component at all.

While subset 1-1 showed promising inter-annotator agreement scores, subset 1-2 comes with disappointing scores. The good values for subset 1-1 are likely the result of the initial, intensive discussion between and with the annotators, producing biased annotations. Comparing this to the weaker values for subset 1-2, it seems obvious that the an-

[1]The ID numbering starts at 10 because the very first annotations did not turn out to be suitable for our purposes, which is why the first nine texts were excluded from the corpus.

Table 2: Inter-annotator agreement ('iaa') of set 1 for annotating *premise* vs. *claim* vs. nothing, by text (Krippendorff's unitizing alpha, Krippendorff et al. 2016)

subset	id	iaa	# of annotators
1-1	10	0.2713	3
1-1	11	0.4078	3
1-1	12	0.2646	3
1-2	13	0.1932	3
1-2	14	-0.0268	3
1-2	15	0.3851	3
1-2	16	0.3002	3
1-2	17	0.0123	3
1-2	18	0.1705	3
1-2	19	0.0941	3
1-2	20	0.3853	3
1-2	21	0.1891	3
1-2	22	0.0681	3

notators need more precise guidelines than what was provided in this second annotation round. This is supported by the fact that introducing a systematic annotation scheme has also been shown to improve inter-annotator agreement in previous projects involving argument annotation (see Stab and Gurevych 2014a). Therefore, our logical next step was to introduce such a scheme, as described in the next section.

4 Introducing an annotation scheme

Our updated annotation process is divided into three major steps (see the similar approaches in e.g. Stab and Gurevych 2014a; Peldszus et al. 2016):

1. Identify the major claim: The annotator reads the full text in order to understand the overall argumentation, and annotates or formulates the major claim.
2. Identify claims and premises: The annotator identifies claims and premises according to a set of criteria, and labels them by semantic category.
3. Review and submit: The annotator goes through the whole text again to finalize their annotation, and submits their annotated text.

We here focus on step two, the identification of claims and premises. Specifically, we describe the approach we apply to identify arguments by systematically categorizing them semantically. For further information on steps one and three see our annotation guidelines (Kawaletz et al. in prep).

In order for a pair of text passages to be included in our database as an argument, it must meet the following criteria:

1. x is a controversial statement (the claim)
2. x is supported or attacked by y (the premise)
3. x supports, attacks or repeats the major claim
4. x is an epistemic, ethical or deontic claim

The first two criteria represent the standard definition of claim and premise (see above), while the third one guarantees that our resulting database has a homogeneous subject matter (in order to facilitate future experiments involving cross-topic transfer). The final criterion, which distinguishes our approach from other, existing ones, is the obligatory assignment of the claim to one of three semantic categories.

In order to test a pair of text passages for these criteria, annotators insert them into linguistic templates (see Kawaletz et al. in prep for details). For the final, semantic criterion, these templates take the form 'x, [___] y' as presented in Table 3. These templates make use of the connectors *and* (for support relations) and *but* (for attack relations), of sentential negation (e.g. *not true* negating *true*), lexical negation (e.g. *false* negating *true*), lexical cues (e.g. *approve/disapprove* for ethical claims), and indication of stress by means of italics to increase grammatical acceptability. All templates may be adapted by the annotator to fit a given syntactic context.

The application of these templates is exemplified in (1). There, we see a claim (bold print) and a premise (underlined) from our corpus, inserted in the template which tests for a supported, positive, deontic claim (represented in Table 3 by *and do this because*). By inserting the two text passages into this template, both the argumentative function and the relation between claim and premise are made explicit.

(1) a. **[M]asking should be mandated and enforced.**

b. *And this should be done because* [i]t's not just about your individual risk tolerance, but about keeping everyone safe.

By systematically applying such templates, our annotation process is now based on principled linguistic judgments rather than on ad hoc decisions. At the point of writing this paper, we have applied our annotation scheme to 12 texts (14,167 words), with promising results: Annotators have reported that applying the provided patterns and being obliged to think about a given text passage in functional terms facilitates argument identification from the start. Thus, by specifying the in-

Table 3: Templates testing for claim categories

claim category	positive claim	negative claim
epistemic		
support	and this is true because and this is the case because	and this is false because and this is not the case because
attack	but this is not true because but this is not the case because	but this is not false because but this *is* the case because
ethical		
support	and this is good because and I find this good because and I approve because and what is good about this is	and this is bad because and I find this bad because and I disapprove because and what is bad about this is
attack	but this is bad because but what is bad about this is	but this is good because but what is good about this is
deontic		
support	and do this because	and don't do this because
attack	but don't do this because	but *do* do this because

ternal, semantic structure of the category *claim* thoroughly, its separation from premises as well as non-argument units becomes clearer. Furthermore, discussions about the status of text passages as argumentative discourse units go more smoothly.

These impressions are backed up by a clear trend toward increasing inter-annotator agreements, as illustrated in Figure 1. In set 2, annotators reached an agreement of up to a rounded 0.6 (as compared to 0.4 for set 1), with no negative values. However, this difference does not come out as significant, as is shown by an unpaired t-test comparing set 1 (M = 0.208831, SD = 0.143289) and set 2 (M = 0.315300, SD = 0.190852); t(23) = 1.5857, p = 0.1265. The fact that we have not been able to support our intuition statistically is likely due to the small sample size and is currently being tested on more texts as the project progresses.

Introducing the argument categories has not only had beneficial effects, however. The annotators have also reported that actually deciding on one functional label is often difficult, due to ambiguities in the text. Interestingly, this sentiment is not reflected in the inter-annotator agreements for set 2: As shown in Table 4, for any given text the difference between the more basic decision (*premise* vs. *claim* vs. nothing) and the more complex decision on a specific claim label (*premise* vs. *epistemic claim* vs. *ethical claim* vs. *deontic claim* vs. nothing) is negligible. A paired t-test reveals that there is indeed no significant difference between the two (p/c/Ø: M = 0.315300, SD = 0.190852; ep/et/d/Ø: M = 0.316033, SD = 0.188318; t(11) = 0.2020, p = 0.8436).

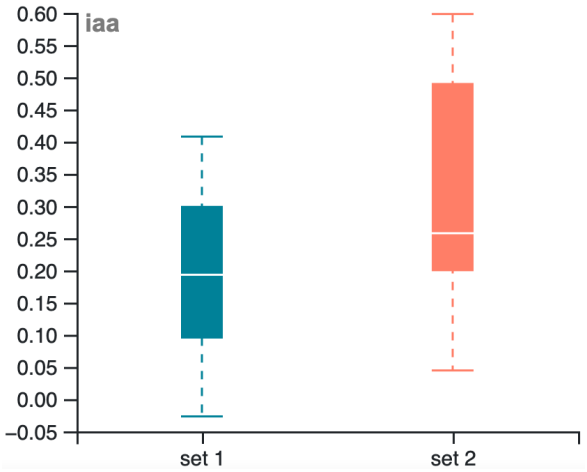

Figure 1: Inter-annotator agreement ('iaa') of sets 1 and 2 for annotating *premise* vs. *claim* vs. nothing (Krippendorff's unitizing alpha, Krippendorff et al. 2016)

Table 4: Inter-annotator agreement ('iaa') of set 2 by text (Krippendorff's unitizing alpha, Krippendorff et al. 2016), comparing *premise* vs. *claim* vs. nothing ('p/c/Ø') and *premise* vs. *epistemic claim* vs. *ethical claim* vs. *deontic claim* vs. nothing ('p/ep/et/d/Ø').

subset	id	iaa (p/c/Ø)	iaa (p/ep/et/d/Ø)	# of annotators
2-1	23	0.1811	0.181	4
2-1	24	0.2809	0.2657	4
2-2	25	0.0951	0.113	3
2-2	26	0.2516	0.2798	3
2-2	27	0.5906	0.5953	3
2-2	28	0.2496	0.2391	3
2-2	29	0.0446	0.0428	3
2-2	30	0.2638	0.2623	3
2-2	31	0.5531	0.5525	3
2-2	32	0.2046	0.2027	3
2-2	33	0.5982	0.5829	3
2-2	34	0.4704	0.4753	3

Apart from further improvements in inter-annotator agreement, we presume that applying a semantic classification of arguments is likely to reduce false positives in manual annotation as well. The text passage in (2), for example, was wrongly classified as an argument by one of three annotators during initial annotation. Applying the templates from Table 3, however, shows that the passage does not fit in either of the twelve categories, as exemplified in (2′) with the pattern *and this is true because* for the category *positive, epistemic, supported claim*.

(2) a. The U.S. Supreme Court threatens to get into the action, too.

b. In May, four conservative justices [...] dissented from an order in South Bay United Pentecostal Church v. Newsom allowing California's COVID-19-related restrictions to remain in place for gatherings at places of worship.

(2′) a. The U.S. Supreme Court threatens to get into the action, too.

b. # *And this is true because*, [i]n May, four conservative justices [...] dissented from an order in South Bay United Pentecostal Church v. Newsom allowing California's COVID-19-related restrictions to remain in place for gatherings at places of worship.

In this example, (2a) is a controversial statement and thus a valid candidate for a claim, but (2b) does not support (nor attack) it. Rather, it specifies more exactly what happened, as can be shown by applying another one of our templates, namely 'X. *What happened is that* y.':

(2″) a. The U.S. Supreme Court threatens to get into the action, too.

b. *What happened is that*, [i]n May, four conservative justices [...] dissented from an order in South Bay United Pentecostal Church v. Newsom allowing California's COVID-19-related restrictions to remain in place for gatherings at places of worship.

As these examples, contrasting with (1) above, illustrate, the point of the semantic classification and of the corresponding paraphrases is to enable annotators in the early stage of argument detection to make informed, well-founded decisions. Previous work with semantic types left the initial argument detection to experts and applied a semantic classification separately (see Hidey et al. 2017), while our approach aims at an improved identification of arguments, which then become available for thorough linguistic investigation.

5 Conclusion and outlook

In this paper, we have sketched an annotation scheme which builds on a function-based classification of arguments. By systematically applying an array of linguistic templates to pairs of text passages, the annotation process is streamlined and facilitated. A first trend for improved inter-annotator agreements, however, has yet to be statistically confirmed. In the long run, we expect significant improvements in annotator recall as well as a less labor-intensive creation of a gold standard (i.e., the curation of the annotated texts by an expert linguist annotator).

In order to further improve our results in terms of inter-annotator agreement and annotator recall, we are currently refining our work flow: For the third set of annotations, we have restricted our corpus to editorials, a more homogeneous subgenre of newspaper opinion pieces, and we are limiting text length to between 40 and 70 sentences in order to avoid too much variation in how the texts deal with argumentation in general. In addition, all annotators are actively involved in the text selection process, pre-assessing and potentially rejecting each text according to a growing catalogue of criteria (e.g. too anecdotal, too many direct quotes).

In the future, apart from the methodological benefits of applying a semantically-grounded annotation scheme, ultimately we will also be able to investigate the semantic types per se. Possible research questions are, for example, which linguistic features annotators and/or machines use to categorize arguments, and how our classification scheme relates to others (e.g. Hidey et al. 2017 on interpretation, evaluation, and agreement/disagreement).

References

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. IJCAI Organization.

Heidrun Dorgeloh and Anja Wanner, editors. 2010. *Syntactic Variation and Genre*. Number 70 in Topics in

English Linguistics. De Gruyter Mouton, Berlin/New York.

Ekkehard Eggs. 2008. 39. Vertextungsmuster Argumentation: Logische Grundlagen. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann, and Sven F. Sager, editors, *Text- und Gesprächslinguistik 1. Halbband*, volume 16/1 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 397–414. De Gruyter Mouton, Berlin/Boston.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining: A working solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7683–7691.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Lea Kawaletz, Heidrun Dorgeloh, Stefan Conrad, and Zeljko Bekcic. in prep. Annotation guidelines for the project 'Probing patterns of argumentative discourse'. Unpublished Manuscript.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity: International Journal of Methodology*, 50(6):2347–2364.

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

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the German online participation project tempelhofer feld. In *Proceedings of the 3rd Workshop on Argument Mining (54th Annual Meeting of the ACL), ArgMining@ACL, Berlin*, Berlin, Germany. Association for Computational Linguistics.

Marie-Francine Moens. 2018. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument & Computation*, 9(1):1–14.

Huy Nguyen and Diane Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO. Association for Computational Linguistics.

Andreas Peldszus, Saskia Warzecha, and Manfred Stede. 2016. Annotation guidelines for argumentation structure. English translation of chapter "Argumentationsstruktur" in Manfred Stede (ed.): Handbuch Textannotation – Potsdamer Kommentarkorpus 2.0. Universitätsverlag Potsdam, 2016.

Carlota S. Smith. 2003. *Modes of Discourse: The Local Structure of Texts*. Cambridge University Press, Cambridge.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 1501–1510, Dublin. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

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

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2019. *Argumentation Mining*. Number 40 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A systematic theory of argumentation: the pragma-dialectical approach*. Cambridge University Press, New York.

Tuija Virtanen. 2010. Variation across texts and discourses: Theoretical and methodological perspectives on text type and genre. In Heidrun Dorgeloh and Anja Wanner, editors, *Syntactic Variation and Genre*, volume 70, pages 53–84. De Gruyter Mouton, Berlin/New York.