Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications

Christian Kirschner[◇†], **Judith Eckle-Kohler**[◇], **Iryna Gurevych**[◇†]

◇ UKP Lab, Technische Universität Darmstadt

† German Institute for Educational Research

http://www.ukp.tu-darmstadt.de

Abstract

This paper presents the results of an annotation study focused on the fine-grained analysis of argumentation structures in scientific publications. Our new annotation scheme specifies four types of binary argumentative relations between sentences, resulting in the representation of arguments as small graph structures. We developed an annotation tool that supports the annotation of such graphs and carried out an annotation study with four annotators on 24 scientific articles from the domain of educational research. For calculating the inter-annotator agreement, we adapted existing measures and developed a novel graph-based agreement measure which reflects the semantic similarity of different annotation graphs.

1 Introduction

Argumentation mining aims at automatically identifying arguments and argumentative relations in argumentative discourse, e.g., in newspaper articles (Feng and Hirst, 2011; Florou et al., 2013), legal documents (Mochales-Palau and Moens, 2011), or scientific publications. Many applications, such as text summarization, information retrieval, or faceted search could benefit from a fine-grained analysis of the argumentation structure, making the reasoning process directly visible. Such an enhanced information access would be particularly important for scientific publications, where the rapidly increasing amount of documents available in digital form makes it more and more difficult for users to find specific information nuggets without investing a lot of time in reading (parts of) documents which are not relevant.

According to well-established argumentation theories in Philosophy and Logic (e.g. Toulmin (1958), Freeman (2011), Walton et al. (2008)), an *argument* consists of several *argument components* which often are of a specific type, such as premise or claim. *Argumentative relations* are usually directed relations between two argument components. Different relation types are distinguished, like *support* or *attack* (Peldszus and Stede, 2013) which indicate that the source argument component is a reason or a refutation for the target component. Argument components and argumentative relations together form the *argumentation structure*. Figure 1 shows the argumentation structure of one argument consisting of 6 argument components and 6 relations between them. Previous work has developed approaches to classify

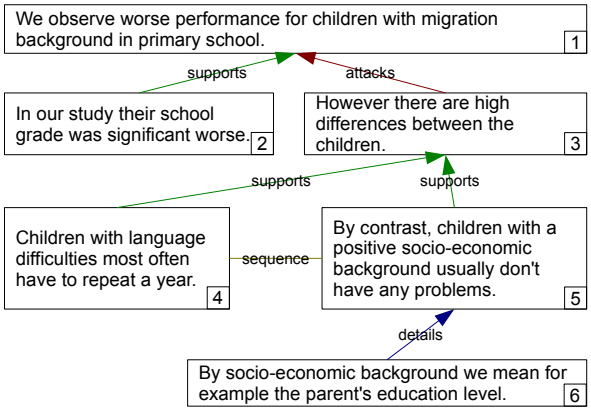

Figure 1: Illustration of one argument consisting of 6 argument components.

sentences in scientific papers according to their argumentative role (Teufel, 1999; Liakata et al., 2012), distinguishing up to seven types of argumentative roles (e.g., Background, Other, Own). However, this results in a coarse-grained analysis of the argumentation structure present in a scientific paper, which merely reflects the more or less standardized way scientific papers are written in many domains (e.g., Natural Sciences or Computer Science). Such a coarse-grained analysis does not reveal how an author connects his thoughts in order to create a convincing line of argumentation. To the best of our knowledge, there exists no prior work which tries to identify argumentative relations between argument components on such a fine-grained level in scientific full-texts yet. This is a challenging task since scientific publications are long and complex documents, and even for researchers in a specific field it can be hard to fully understand the underlying argumentation structures.

We address this gap and aim at developing methods for the automatic identification of argumentation structures in scientific publications. We chose scientific journal articles from the educational research as a prototypical domain, because it is of particular interest not only for educational researchers, but also for other groups in the society, such as policy makers, teachers or parents.

This paper presents the results of our annotation of 24 articles from educational research (written in German) – a crucial step towards developing and testing automatic methods. Our contributions can be summarized as follows: (i) We introduce an annotation scheme and an annotation tool for the fine-grained analysis of argumentation structures in scientific publications, which represents arguments as small graph structures. (ii) We developed a novel graph-based inter-annotator agreement measure, which is able to reflect the semantic similarity of different annotation graphs. (iii) Finally, we present the results of a detailed quantitative and qualitative analysis of the annotated dataset where we characterize the argumentation structures in scientific publications and identify major challenges for future work.

The rest of the paper is organized as follows: First we discuss related work (section 2). In section 3, we describe our annotation scheme and the annotation study, and in section 4 the inter-annotator agreement measures are introduced. The results of the quantitative and qualitative analysis are discussed in section 5. Section 6 concludes.

2 Related Work

This section discusses related work regarding the annotation of argumentation structure on the one hand, and annotating scientific articles on the other hand. We give an overview of (i) annotation schemes for annotating argumentation and discourse structure, (ii) inter-annotator agreement (IAA) metrics suitable for this annotation task, (iii) previous annotation studies.

Annotation Schemes Previously, annotation schemes and approaches for identifying arguments in different domains have been developed. For instance, Mochales-Palau and Moens (2011) identify arguments in legal documents, Feng and Hirst (2011) focus on the identification of argumentation schemes (Walton, 1996) in newspapers and court cases, Florou et al. (2013) apply argumentation mining in policy modeling, and Stab and Gurevych (2014) present an approach to model arguments in persuasive essays. Most of the approaches focus on the identification and classification of argument components. There are only few works which aim at identifying argumentative relations and consequently argumentation structures. Furthermore it is important to note that the texts from those domains differ considerably from scientific publications regarding their length, complexity, purpose and language use.

Regarding argumentation mining in scientific publications, one of the first approaches is the work called *Argumentative Zoning* by Teufel (1999) which was extended by Teufel et al. (2009). According to the extended annotation scheme, each sentence in a scientific publication is annotated with exactly one of 15 categories (e.g. Background or Aim), reflecting the argumentative role the sentence has in the text. Mapping this scheme to our terminology (see section 1), a sentence corresponds to an argument component. The aim of this annotation scheme is to improve information access and to support applications like automatic text summarization (Teufel and Moens, 2002; Ruch et al., 2007;

Contractor et al., 2012). While the authors themselves do not consider argumentative relations, Angrosh et al. (2012) transfer the argumentation inherent in the categories of the *Argumentative Zoning* to the Toulmin model (Toulmin, 1958) and therefore describe how argument components of several types relate to each other. For example, research findings are used to support "statements referring to the problems solved by an article" and "statements referring to current work shortcomings" support "statements referring to future work". However, the paper focuses on citation contexts and considers relations only on a coarse-grained level.

Several similar annotation schemes for scientific publications exist. For instance, Liakata et al. (2012) proposed *CoreSC* ("Core Scientific Concepts"), an annotation scheme consisting of 11 categories[1]. Mizuta and Collier (2004) provide a scheme consisting of 7 categories (plus 5 subcategories) for the biology domain. In addition Yepes et al. (2013) provide a scheme to categorize sentences in abstracts of articles from biomedicine with 5 categories.

Furthermore, Blake (2010) describes approaches to identify scientific claims or comparative claim sentences in scientific articles (Park and Blake, 2012). Again these works do not consider argumentative relations on a fine-grained level, but focus on the classification of argument components. While all of these works use data from the natural sciences, there are only few works in the domain of social sciences (e.g. Ahmed et al. (2013)), and to the best of our knowledge no previous work has addressed scientific publications in the educational domain.

A field that is closely related to the annotation of argumentation structures is the annotation of discourse structure which aims at identifying discourse relations that hold between adjacent text units, e.g. sentences, clauses or nominalizations (Webber et al., 2012). Often, the text units considered in discourse analysis correspond to argument components, and discourse relations are closely related to argumentative relations. Most previous work in automated discourse analysis is based on corpora annotated with discourse relations, most notably the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) and the Rhetorical Structure Theory (RST) Discourse Treebank (Carlson et al., 2001). However, the data consists of newspaper articles (no scientific articles), and only relations between adjacent text units are identified. In addition, it is still an open question how the proposed discourse relations relate to argumentative relations (the difference of the relations is best illustrated by the work of Biran and Rambow (2011)). Nevertheless, annotated corpora like this can be valuable resources for training automatic classifiers later.

IAA Metrics Current state-of-the-art annotation studies use chance corrected measures to compute IAA, i.e., random agreement is included in the calculation. The values can be in the range of -1 to 1, a value of 0 indicates random agreement and a value of 1 perfect agreement (negative values indicate a negative correlation). One of the most popular chance corrected measures for two raters is Cohen's κ (Cohen, 1960). While Cohen's κ assumes different probability distributions for each rater, there exist other approaches which assume a single distribution for all raters (Scott, 1955). In addition, extensions to multiple raters exist. Multi-π is the extension of Scott's π by Fleiss (1971). Multi-κ is the extension of Cohen's κ by Hubert (1977).

All of these measures are well suited for tasks where we have a fixed set of independent and uniformly distributed entities to annotate. However, as soon as the annotation of one entity depends on the annotation of another entity, or some entities have a higher overall probability for a specific annotation than others, the measures may yield misleadingly high or low values (see section 4). Apart from that, chance-corrected measures are criticized because they "are often misleading when applied to unbalanced data sets" (Rehbein et al., 2012) and can be "problematic in categorization tasks that do not have a fixed number of items and categories" (van der Plas et al., 2010). Therefore, many researchers still report raw percentage agreement without chance correction.

Annotation Studies Table 1 gives an overview of previous annotation studies performed for scientific publications. In all of these studies, the annotators have to label argument components (typically, each sentence represents exactly one argument component) with one out of 3 - 15 categories. In most of

[1]For a comparison between Argumentative Zoning and CoreSC, see Liakata et al. (2010).

Author	Data	Annotators	#Cat	Guidelines	IAA
Teufel (1999)	22 papers (CL)	3 (semi-experts)	3	6 pages	0.78
	26 papers (CL)	3 (semi-experts)	7	17 pages	0.71
	3x1 paper (CL)	3x6 (untrained)	7	1 page	0.35-0.72
Teufel et al. (2009)	30 papers (Chemistry)	3 (different)	15	111 pages	0.71
	9 papers (CL)	3 (experts)	15	111 pages	0.65
Liakata et al. (2012)	41 papers (Biochemistry)	3 (experts)	11	45 pages	0.55
Blake (2010)	29 papers (Biomedicine)	2 (students)	5	discussion	0.57-0.88

Table 1: Comparison of annotation studies on scientific full-texts (CL = computational linguistics, #Cat = number of categories which can be annotated, IAA = chance-corrected inter-annotator agreement).

the studies, the annotators are at least semi-experts in the particular domain and get detailed annotation guidelines. Regarding the IAA, Teufel et al. (2009) report that untrained annotators performed worse than trained expert annotators. All of the agreement measures in table 1 are chance corrected and therefore comparable.

There are also annotation studies outside the domain of scientific articles which deal with argumentative relations. Mochales-Palau and Moens (2011) report an IAA of Cohen's $\kappa = 0.75$ (legal documents) but only for the identification of argument components (here claims and premises) and not for argumentative relations. Stab and Gurevych (2014) report an IAA of Fleiss' $\pi = 0.8$ for argumentative support and attack relations in persuasive essays. However, these relations are annotated between pre-annotated premises and claims, which simplifies the task considerably: annotators already know that premises have outgoing support and attack relations and claims incoming ones, i.e., they only have to annotate the target or source components of the relations as well as their type. Furthermore, compared to scientific articles, persuasive essays are much shorter and less complex regarding language use.

3 Annotation Study

This section describes our annotation study: we introduce the dataset, the annotation scheme and describe the annotation tool we developed.

Dataset For the annotation study, we selected 24 publications from 5 controversial educational topics (teaching profession, learning motivation, attention deficit hyperactivity disorder (ADHD), bullying, performance rating) from different journals in the domain of educational psychology and developmental psychology.[2] All of the articles are in German, about 10 pages of A4, describe empirical studies, and are composed of similar sections (introduction, methods, results, discussion). In our annotation study, we annotated the introduction and discussion sections and left out the methods and results sections, because these sections usually just describe the experimental setup without any assessment or reasoning.

The dataset contains the following annotatable[3] text units: 529 paragraphs (22 per document), 2743 sentences (114 per document), 79680 tokens (3320 per document). On average, we have a comparably high number of 29 tokens per sentence, which indicates the high complexity of the texts (Best, 2002).

At least three annotators with different backgrounds annotated the journal articles, some documents were annotated by a forth annotator. Two of the annotators were students (psychology and sociology), one was a PhD student (computer science) and the forth annotator had a PhD degree (computational linguistics). We developed annotation guidelines of about 10 pages of A4[4] and trained the annotators on these guidelines. In a pre-study, the annotators annotated five documents about language learning (not included in the dataset described above). During this pre-study, the annotations were discussed several times and the annotation guidelines were adapted. All in all, the annotation study extended over several months part time work. The annotation of one single document took about two hours.

Annotation Scheme Our annotation scheme specifies argument components and binary relations

[2] published by Hogrefe & Huber Verlagsgruppe, http://psycontent.metapress.com

[3] without headings, abstract, method/results section.

[4] We plan to make the guidelines publicly available.

between argument components. Every sentence correponds to an argument component. Our observations show that most of the arguments can be found on the sentence level. This simplification helps to keep the identification of argumentative relations manageable: Scientific publications are highly complex texts containing argumentation structures that are often hard to understand even for researchers in the respective field.

There are four types of relations: the directed relations *support*, *attack*, *detail*, and the undirected *sequence* relation. The *support* and *attack* relations are argumentative relations, which are known from related work (Peldszus and Stede, 2013), whereas the latter two correspond to discourse relations used in Rhetorical Structure Theory (RST) (William and Thompson, 1988). The *sequence* relation corresponds to "Sequence" in RST, the *detail* relation roughly corresponds to "Background" and "Elaboration". We added the *detail* relation, because we observed many cases in scientific publications, where some background information (for example the definition of a term) is given, which is important for understanding the overall argumentation.

A *support* relation between an argument component A and another argument component B indicates that A supports (reasons, proves) B. Similarly, an *attack* relation between A and B is annotated if A attacks (restricts, contradicts) B. The *detail* relation is used, if A is a detail of B and gives more information or defines something stated in B without argumentative reasoning. Finally, we link two argument components with the *sequence* relation, if two (or more) argument components belong together and only make sense in combination, i.e., they form a multi-sentence argument component.[5]

Annotation Tool We developed our own web-based annotation tool DiGAT which we think is better suited for annotating relations in long texts than existing tools like WebAnno (Yimam et al., 2013), brat (Stenetorp et al., 2012) or GraPAT (Sonntag and Stede, 2014). Although all of them allow to annotate relations between sentences, the view quickly becomes confusing when annotating relations. In WebAnno and brat, the relations are drawn with arrows directly in the text. Only GraPAT visualizes the annotations in a graph. However, the text is included in the nodes directly in the graph, which again becomes confusing for texts with multiple long sentences.

DiGAT has several advantages over existing tools. First, the full-text with its layout structure (e.g., headings, paragraphs) is displayed without any relation annotations on the left-hand side of the screen. The argumentation structure which emerges by adding relations is visualized as a graph on the right-hand side of the screen. Second, the tool automatically marks each sentence in an argumentative paragraph by a different color for better readability. In addition, discourse markers in the text are highlighted to support the annotation of relations.[6]

4 IAA Measures for Relations

This section introduces the measures we used for calculating IAA. We will describe the adaption of measures discussed in section 2 to relation annotations. We also motivate and introduce a novel graph-based measure.

Adaptation of the Dataset to use Chance-corrected IAA Measures In this work, we focus on binary argumentative relations between two argument components. In order to use the chance-corrected measures introduced in section 2, we have to consider each possible pair of argument components in a document as either being connected via a relation (of different types) or not. Then we calculate the IAA with the multi-κ measure (Hubert, 1977) because it is suitable for multiple raters and assumes different probability distributions for each rater.

One drawback of this approach is that the probability of a relation between two argument components decreases with the distance between the components in the text. It is much more likely that two consecutive argument components are related than two components which are in different paragraphs (we observe about 70% of all relations to be between adjacent argument components). Consequently, we get a very high number of non-relations and a very unbalanced dataset because for a document with n=100 argument components, we get

[5] This is necessary because we fixed the size of one argument component to exactly one sentence.

[6] A previous annotation study showed that often discourse markers are signals of argumentative relations (Kluge, 2014).

$\frac{(n-1)*n}{2} = 4950$ pairs, only 1-2% of which are usually related.

Therefore, we limited our evaluation to pairs with a distance of d<6 argument components, since we observed only very few relations with a higher distance. We define the *distance* between two argument components as the number of argument components between them in the text flow. Thus, two adjacent argument components have the distance 0. For a document with n=100 argument components, this reduces the number of pairs to $(n-d)*(d) = 564$. Since we still have a higher probability for a relation with a small distance compared to a relation with a larger distance, we additionally calculated the agreement individually considering only relations with a particular distance (d=0, d=1, d=2, d>2) and averaged over the results weighting them according to the average probability for the distances (69.5% d=0, 18.5% d=1, 7% d=2, 5% d>2). We call this value *Weighted Average* (WA) in the next sections.

Adapted Percentage Agreement / F_1-Score As pointed out in section 2, many researches still report raw percentage agreement. Usually percentage agreement is calculated by dividing the number of annotation items where all annotators agreed by the total number of all annotation items. The high number of non-relations would result in a high agreement that would be meaningless. Therefore, we divide the number of annotation items where the annotators agreed by the number of annotation items where at least one annotator found a relation. We call this approach *adapted percentage agreement* (APA), also called "positive agreement" (Cicchetti and Feinstein, 1990).

We have to keep two things in mind: First, this APA approach leads to worse agreement results than usual percentage agreement because the agreement for non-relations is not considered at all. Second, the APA decreases with an increasing number of annotators because the number of pairs where all annotators agree decreases, and simultaneously the number of pairs where at least one anotator found a relation increases. Therefore, we average over the pairwise APA. This approach is quite similar to the F_1-Score = $\frac{2TP}{2TP+FP+FN}$ (TP = true positives = both annotators found a relation, FP/FN = false positives/false negatives = the annotators disagree). The only difference is the factor 2 for the true positives both in numerator and denominator which gives more weight to the agreements. For the two annotation graphs in figure 2, we get an APA of $\frac{1}{3} = 0.33$ or a F_1-Score of $\frac{2*1}{2*1+2} = 0.5$ (ignoring the direction of the relations).

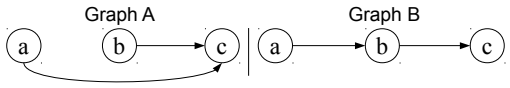

Figure 2: Two simple annotation graphs (each node represents an argument component).

New Graph-based IAA Measure The measures described above consider each pair of argument components independently in isolation. However, we do not annotate pairs of argument components in isolation, but we consider the complete text-flow and represent the argumentation structure in an annotation graph consisting of the argument components as nodes and the relation annotations as edges between them. This means that the annotation of one entity can influence the annotation of a second entity. So the measures do not consider the overall annotation graph structure. For example, in figure 2 both annotators think that the nodes a and b directly or indirectly support/attack node c which we cannot capture if we only consider pairs of argument components in isolation.

Consequently, we also need a method to calculate the IAA for annotation graphs, considering the graph structure. To the best of our knowledge, such a graph-based IAA metric has not been developed so far. There are approaches in graph theory which aim at calculating the similarity of graphs. However, most of these approaches are very complex because they target larger graphs and a matching of the nodes is required (which is not necessary in our case). Hence, we propose a novel graph-based agreement measure, which can identify different annotations with similar meaning. For example, it considers that in figure 2 both annotators directly or indirectly found a relation from node a to node c. Hence, the new measure results in a higher agreement than the standard measures.

The measure determines to what extent graph A is included in graph B and vice versa (note that relation types are ignored in this approach). To calculate to what extent graph A is included in graph B, we av-

erage over the sum of the inverse of the shortest path distance between two nodes which are connected by a relation of graph A in graph B:

$$\frac{1}{|E_A|} \sum_{(x,y) \in E_A} \frac{1}{SP_B(x,y)}$$

E_A is the set of edges in graph A with the elements (x,y) whereas x is the source and y the target node. $SP_B(x, y)$ is the shortest path between the nodes x and y in graph B.

We illustrate the process with an example (see figure 2). Starting from graph A (1), we find the two edges $a - c$ (distance 2 in graph B) and $b - c$ (distance 1 in graph B). Starting from graph B (2), we find the two edges $a - b$ (distance ∞ in graph A) and $a - c$ (distance 1 in graph A). So the graph-based agreement is:

(1) $\frac{1}{2} * (\frac{1}{2} + \frac{1}{1}) = 0.75$
(2) $\frac{1}{2} * (\frac{1}{1} + \frac{1}{\infty}) = 0.5$

On average, the graph-based agreement for the graphs A and B is $\frac{(0.75+0.5)}{2} = 0.625$. Considering (1) and (2) as precision and recall, we can also calculate F_1-Score = $\frac{2*precision*recall}{precision+recall}$. This measure has the advantage that it becomes higher for similar precision and recall values (also called "harmonic mean"). So in the example from figure 2 the F_1-Score is $\frac{2*0.5*0.75}{0.5+0.75} = 0.6$.

5 Results and Discussion

In this section, we will perform both a quantitative and a qualitative analysis of the annotated argumentative relations and the argumentation structure.

5.1 Quantitative Analysis

We analyze the IAA for the relations identified by the annotators. Table 2 gives an overview of the class distributions for each annotator (A1 - A4). While the distribution of relation distances is quite homogeneous (about 70% of identified relations are between adjacent argument components), there are large differences regarding the number of identified relations and the distribution of relation types. Especially A4 found more relations than the other annotators. In part, this is due to the fact that A4 annotated only 5 of the 24 documents which had above-average length. Nevertheless, we observe that A3 found only few relations compared to the other annotators, especially of the types sequence and detail (also in absolute numbers) and annotated most of the relations as support which is still less than the other annotators in absolute numbers.

Table 3 presents the IAA for the relations. We get multi-κ values up to 0.63 considering all distances d<6, which is a fair agreement considering the difficulty of the task. We already observed in the individual statistics that A3 identified much fewer relations than the other annotators. This is reflected in the agreement values which are lower for all pairs of annotators where A3 is involved. However, an analysis of the relations annotated by A3 using the graph-based measure reveals that most of these relations where also identified by the other annotators: the graphs by A3 are to a large extent contained in the graphs of the other annotators (0.63 - 0.68). The other way round, the graphs by A3 only marginally contain the graphs of the other annotators (0.29 - 0.41). This indicates that A3 only annotated very explicit argumentative relations (see section 5.2).

There are only small differences between the graph-based measure which considers the argumentation structure, and multi-κ which considers each pair of argument components in isolation. This can be attributed to the fact that about 50% of all argument components are in connected graph components[7] with only two nodes, i.e., there is no argumentation structure to consider for the graph-based measure.

In contrast, if we only consider graph components with at least 3 nodes for any pair of annotators, the graph-based IAA improves by about 0.15 (while the other measures do not change). This clearly demonstrates the advantages of our new graph-based approach for detecting different annotations with similar meaning.

Table 5 shows IAA when considering the relation types. This annotation task requires to decide between 5 different classes (support, attack, detail, sequence, none). The chance-corrected multi-κ values downgrade by about 0.1. If we consider the individual distances (WA measure), we get significantly lower results compared to considering all distances together (d<6).

Table 4 shows the multi-κ values for the different

[7] In a connected graph component, there exists a path between all nodes (assuming undirected edges).

Annotator	Relations									Distance between relations							
	#Sup	%	#Att	%	#Seq	%	#Det	%	#ALL	#d=0	%	#d=1	%	#d=2	%	#d>2	%
A1	45.0	58.8	8.9	11.6	12.0	15.7	10.7	13.9	76.7	51.5	67.2	14.1	18.4	5.9	7.7	5.1	6.7
A2	40.0	43.7	11.5	12.5	26.9	29.3	13.3	14.5	91.7	61.1	66.6	19.3	21.1	6.9	7.5	4.4	4.8
A3	36.5	73.6	3.6	7.2	5.5	11.0	4.1	8.2	49.7	37.7	75.8	8.0	16.1	2.3	4.6	1.7	3.4
A4	54.8	45.7	15.2	12.7	28.6	23.9	21.2	17.7	119.8	82.0	68.4	21.8	18.2	10.0	8.3	6.0	5.0
ALL	44.1	55.4	9.8	11.0	18.2	20.0	12.3	13.6	84.5	58.1	69.5	15.8	18.5	6.3	7.0	4.3	5.0

Table 2: Individual statistics for identified relations (#Sup/Att/Seq/Det = average number of support/attack/sequence/detail relations per document; #ALL = average number of relations per document; #d = average number of relations with distance d per document).

Annotators	Weighted Average (WA)			d<6			Graph-based			
	APA	F_1	multi-κ	APA	F_1	multi-κ	1-2	2-1	Avg.	F_1
A1-A2	0.5030	0.6380	0.4668	0.4681	0.6327	0.5822	0.5102	0.6460	0.5781	0.5607
A1-A4	0.5040	0.6467	0.4421	0.4859	0.6492	0.5988	0.5083	0.7343	**0.6213**	**0.5959**
A4-A2	**0.5553**	**0.6855**	**0.4744**	**0.5265**	**0.6873**	**0.6335**	0.5730	0.6069	0.5900	0.5881
A3-A1	0.3776	0.5261	0.3613	0.3693	0.5345	0.4903	0.6285	0.4059	0.5172	0.4795
A3-A2	0.3813	0.5189	0.3388	0.3629	0.5257	0.4767	**0.6815**	0.3380	0.5097	0.4424
A3-A4	0.3251	0.4690	0.2459	0.3152	0.4782	0.4229	0.6770	0.2868	0.4819	0.3992
ALL	0.4270	0.5559	0.3912	0.4044	0.5683	0.5257	-	-	0.5387	0.4984

Table 3: IAA for relation annotation, relation type is ignored (APA = adapted percentage agreement, weighted average = averaged results for relations with distance 0, 1, 2 and >2, weighted according to their probability; d<6 = agreement for all relations with a distance d<6; 1-2 or 2-1 (graph-based) = measures how much the annotation of annotator 1 is included in the annotation of annotator 2 or vice versa).

Annotators	multi-κ					
	d=0	d=1	d=2	d>2	WA	d<6
A1-A2	0.5426	0.3346	0.2625	0.1865	0.4668	0.5822
A1-A4	0.4756	0.3868	0.3729	0.2768	0.4421	0.5988
A4-A2	0.5388	0.3349	0.3878	0.2151	0.4744	0.6335
A3-A1	0.4079	0.2859	0.2562	0.1949	0.3613	0.4903
A3-A2	0.4002	0.2234	0.1779	0.1369	0.3388	0.4767
A3-A4	0.2889	0.1397	0.1353	0.1950	0.2459	0.4229
ALL	0.4488	0.2856	0.2488	0.1801	0.3912	0.5257

Table 4: IAA for relation annotation with multi-κ measure for different distances (relation type is ignored).

Annotators	Weighted Average (WA)			d<6		
	APA	F_1	multi-κ	APA	F_1	multi-κ
A1-A2	0.3144	0.4588	0.3784	0.2980	0.4516	0.4742
A1-A4	**0.3624**	**0.5124**	**0.4105**	**0.3479**	**0.5111**	**0.5153**
A4-A2	0.3126	0.4611	0.3546	0.3024	0.4594	0.4911
A3-A1	0.2838	0.4275	0.3341	0.2756	0.4278	0.4299
A3-A2	0.1986	0.3167	0.2535	0.1933	0.3187	0.3615
A3-A4	0.1884	0.3048	0.2065	0.1835	0.3078	0.3335
ALL	0.2699	0.4002	0.3246	0.2582	0.4023	0.4285

Table 5: IAA for relation annotation (relation type is considered).

distances in detail. As we can see, the agreement degrades significantly with increasing distance and even for distance d=0 the values are lower than for d<6. The reason for this is the high number of non-relations compared to relations, especially for distances with d>2.

5.2 Qualitative Analysis

In order to get a better understanding of the reasons for the partially low IAA, we performed a qualitative analysis. We focused on support and attack relations and compared instances annotated with high agreement[8] with instances where annotators disagreed.

Relations annotated with high agreement: Support or attack relations annotated with high agreement can be considered as explicit argumentative relations. We identified different types of argument components with explicit incoming support relations, which are typically marked by surface indicators (e.g. sentence mood, discourse markers, stylistic devices): a claim (expressed e.g. as a rhetorical question), an opinion statement marked by words expressing sentiment (e.g. *überraschend (surprisingly)*), a hypothesis marked by a conjunctive sentence mood and modal verbs (e.g. *könnte (could)*), a conclusion or summarizing statement marked by discourse markers (e.g. *daher (therefore)*), or a generalizing statement marked by adverbial expressions (e.g. *gemeinsam sein (have in common)*).

Another explicit support relation was annotated for argument components supporting an observation that is based on a single piece of evidence; here the supporting argument component contained lexical indicators such as *konform gehen (be in line with)*. Explicit attack relations, on the other hand, appeared

[8] High agreement means that 3 or 4 annotators agreed.

to be marked by a combination of discourse markers expressing concession (e.g., *jedoch, allerdings (however), aber (but)*) and negation or downtoning markers (e.g. *kaum (hardly)*). We found negation to be expressed in many variants, including not only explicit negation, such as *nicht (not), kein (no)*, but also (implicit) lexicalized negation, e.g. verbs such as *ausstehen (is pending)*.

Relations where annotators disagreed: Our analysis of support relations that were annotated only by one of 3 (4) annotators revealed that there are many cases, where the disagreement was due to an alternation of support and detail or support and sequence relation. These cases can be regarded as weakly argumentative, i.e. the argument component with the incoming relation is not considered by all annotators as a statement that requires argumentative support.

We performed the same qualitative analysis for attack relations and found that in most cases either a concession marker in the attacking argument component is present, or some form of negation, but not both as in the explicit case of the attack relation.

Ambiguity as the main reason for disagreement: One of the main challenges in identifying argumentative relations on a fine-grained level in scientific publications is ambiguity (Stab et al., 2014). All the measures used to calculate IAA assume that there is one single correct solution to the annotation problem. Actually, we believe that in many cases several correct solutions exist depending on how the annotators interpret the text. In our qualitative analysis, we found that this is especially true for argument components that are lacking discourse markers or other surface indicators. For example, the following text snippet can be interpreted in two ways:

*"School grades have severe consequences for the academic career of students.**(a)** Students with good grades can choose among numerous career options.**(b)** According to Helmke (2009), judgments of teachers must therefore be accurate, when qualification certificates are granted.**(c)**"*[9]

According to one interpretation, there is a relation chain between a, b, and c (a supports b and b supports c), while the other interpretation considers a and b as a sequence which together supports c (a supports c and b supports c).

Another source of ambiguity is the ambiguity of discourse markers, which sometimes seems to trigger annotation decisions that are based on the presence of a discourse marker, rather than on the semantics of the relation between the two argument components. A prototypical example are discourse markers expressing concession, e.g. *jedoch, allerdings (however)*. They are often used to indicate attacking argument components, but they can also be used in a different function, namely to introduce counter-arguments. In this function, which has also been described by (Grote et al., 1997), they appear in an argument component with incoming support relations.

Apart from ambiguity, we found that another difficulty are different granularities of some argument components. Sentences might relate to coarse-grained multi-sentence units and this is not representable with our fine-grained annotation scheme. This is illustrated by the following example where *against this background* relates to a long paragraph describing the current situation: *Against this background, the accuracy of performative assessment received growing attention recently.*

6 Conclusion

We presented the results of an annotation study to identify argumentation structures on a fine-grained level in scientific journal articles from the educational domain. The annotation scheme we developed results in a representation of arguments as small graph structures. We evaluated the annotated dataset quantitatively using multiple IAA measures. For this, we proposed adaptions to existing IAA measures and introduced a new graph-based measure which reflects the semantic similarity of different annotation graphs. Based on a qualitative analysis where we discussed characteristics of argument components with high and low agreement, we identified the often inherent ambiguity of argumentation structures as a major challenge for future work on the development of automatic methods.

[9] Südkamp and Möller (2009), shortened and translated.

Acknowledgements

This work has been supported by the German Institute for Educational Research (DIPF) as part of the graduate program "Knowledge Discovery in Scientific Literature" (KDSL). We thank Stephanie Bäcker and Greta Koerner for their valuable contributions.

References

Shameem Ahmed, Catherine Blake, Kate Williams, Noah Lenstra, and Qiyuan Liu. 2013. Identifying claims in social science literature. In *Proceedings of the iConference*, pages 942–946, Fort Worth, USA. iSchools.

M.A. Angrosh, Stephen Cranefield, and Nigel Stanger. 2012. A Citation Centric Annotation Scheme for Scientific Articles. In *Australasian Language Technology Association Workshop*, pages 5–14, Dunedin, New Zealand.

Karl-Heinz Best. 2002. Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel. In *Göttinger Beiträge zur Sprachwissenschaft 7*, pages 7–31.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 05(04):363–381.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of biomedical informatics*, 43(2):173–189.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Aalborg, Denmark.

Domenic V. Cicchetti and Alvan R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551 – 558.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Danish Contractor, Yufan Guo, and Anna Korhonen. 2012. Using Argumentative Zones for Extractive Summarization of Scientific Articles. In *Proceedings of the 23th International Conference on Computational Linguistics (COLING 2012)*, pages 663–678, Mumbai, India.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying Arguments by Scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, USA.

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

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, Sofia, Bulgaria.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*, volume 18 of *Argumentation Library*. Springer.

Brigitte Grote, Nils Lenke, and Manfred Stede. 1997. Ma(r)king concessions in English and German. *Discourse Processes*, 24(1):87–118.

Lawrence Hubert. 1977. Kappa revisited. *Psychological Bulletin*, 84(2):289.

Roland Kluge. 2014. Automatic Analysis of Arguments about Controversial Educational Topics in Web Documents, Master Thesis, Ubiquitious Knowledge Processing Lab, TU Darmstadt.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin R Batchelor. 2010. Corpora for the Conceptualisation and Zoning of Scientific Papers. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC)*, pages 2054–2061, Valletta, Malta.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Yoko Mizuta and Nigel Collier. 2004. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 29–35, Geneva, Switzerland.

Raquel Mochales-Palau and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Dae Hoon Park and Catherine Blake. 2012. Identifying comparative claim sentences in full-text scientific articles. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 1–9, Jeju, Republic of Korea.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 28–30, Marrakech, Morocco.

Ines Rehbein, Joseph Ruppenhofer, Caroline Sporleder, and Manfred Pinkal. 2012. Adding nominal spice to SALSA-frame-semantic annotation of German nouns and verbs. In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS12)*, pages 89–97, Vienna, Austria.

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbühler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, et al. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International journal of medical informatics*, 76(2):195–200.

William A Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3):321–325.

Jonathan Sonntag and Manfred Stede. 2014. GraPAT: a Tool for Graph Annotations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of the the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510, Dublin, Ireland.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective. In *Frontiers and Connections between Argumentation Theory and Natural Language Processing*, Bertinoro, Italy.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 102–107, Avignon, France.

Anna Südkamp and Jens Möller. 2009. Referenzgruppeneffekte im Simulierten Klassenraum. *Zeitschrift für Pädagogische Psychologie*, 23(3):161–174.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1493–1502, Singapore.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

Stephen E. Toulmin. 1958. *The uses of Argument*. Cambridge University Press.

Lonneke van der Plas, Tanja Samardžić, and Paola Merlo. 2010. Cross-lingual validity of PropBank in the manual annotation of French. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 113–117, Uppsala, Sweden.

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

Douglas N Walton. 1996. *Argumentation schemes for presumptive reasoning*. Routledge.

Bonnie Webber, Mark Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18:437–490.

Mann William and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Antonio Jimeno Yepes, James G. Mork, and Alan R. Aronson. 2013. Using the argumentative structure of scientific literature to improve information access. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 102–110, Sofia, Bulgaria.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 1–6, Sofia, Bulgaria.