Annotating Argument Components and Relations in Persuasive Essays

Christian Stab[‡] **and Iryna Gurevych**[†‡]
[‡]Ubiquitous Knowledge Processing Lab (UKP-TUDA),
Department of Computer Science, Technische Universität Darmstadt
[†]Ubiquitous Knowledge Processing Lab (UKP-DIPF),
German Institute for Educational Research
http://www.ukp.tu-darmstadt.de

Abstract

In this paper, we present a novel approach to model arguments, their components and relations in persuasive essays in English. We propose an annotation scheme that includes the annotation of claims and premises as well as support and attack relations for capturing the structure of argumentative discourse. We further conduct a manual annotation study with three annotators on 90 persuasive essays. The obtained inter-rater agreement of $\alpha_U = 0.72$ for argument components and $\alpha = 0.81$ for argumentative relations indicates that the proposed annotation scheme successfully guides annotators to substantial agreement. The final corpus and the annotation guidelines are freely available to encourage future research in argument recognition.

1 Introduction

The ability of formulating persuasive arguments is a crucial aspect in writing skills acquisition. On the one hand, well-defined arguments are the foundation for convincing an audience of novel ideas. On the other hand, good argumentation skills are essential for analyzing different stances in general decision making. By automatically recognizing arguments in text documents, students will be able to inspect their texts for plausibility as well as revise the discourse structure for improving argumentation quality. This assumption is supported by recent findings in psychology, which confirm that even general tutorials effectively improve the quality of written arguments (Butler and Britt, 2011). In addition, *argumentative writing support systems* will enable tailored feedback by incorporating argument recognition. Therefore, it could be expected that they provide appropriate guidance for improving argumentation quality as well as the student's writing skills.

An argument consists of several components (i.e. claims and premises) and exhibits a certain structure constituted by argumentative relations between components (Peldszus and Stede, 2013). Hence, recognizing arguments in textual documents includes several subtasks: (1) separating argumentative from non-argumentative text units, (2) identifying claims and premises, and (3) identifying relations between argument components.

There exist a great demand for reliably annotated corpora including argument components as well as argumentative relations (Reed et al., 2008; Feng and Hirst, 2011) since they are required for supervised machine learning approaches for extracting arguments. Previous argument annotated corpora are limited to specific domains including legal documents (Mochales-Palau and Moens, 2008), newspapers and court cases (Reed et al., 2008), product reviews (Villalba and Saint-Dizier, 2012) and online debates (Cabrio and Villata, 2012). To the best of our knowledge, no work has been carried out to annotate argument components and argumentative relations in persuasive essays (section 2). In addition, the reliability of the corpora is unknown, since only few of these works provide holistic inter-rater agreement scores and none a detailed analysis and discussion of inter-rater agreement.

In this work, we introduce a new argument annotation scheme and a corpus of persuasive essays annotated with argument components and argumentative relations. Our primary motivation is to create

This work is licensed under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: http://creativecommons.org/licenses/by/4.0/

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1501–1510, Dublin, Ireland, August 23-29 2014.

a corpus for argumentative writing support and to achieve a better understanding of how arguments are represented in texts. In particular, the contributions of this paper are the following: First, we introduce a novel annotation scheme for modeling arguments in texts. Second, we present the findings of a pre-study and show how the findings influenced the definition of the annotation guidelines. Third, we show that the proposed annotation scheme and guidelines lead to substantial agreement in an annotation study with three annotators. Fourth, we provide the annotated corpus as freely available resource to encourage future research.[1]

2 Related Work

2.1 Previous Argument Annotated Corpora

Currently, there exist only a few corpora that include argument annotations. The work most similar to ours with respect to the annotation scheme is Araucaria (Reed et al., 2008) since it also includes structural information of arguments. It is based on the *Argumentation Markup Language* (AML) that models argument components in a XML-based tree structure. Thus, it is possible to derive argumentative relations between components though they are not explicitly included. In contrast to our work, the corpus consists of several text genres including newspaper editorials, parliamentary records, judicial summaries and discussion boards. In addition, the reliability of the annotations is unknown. Nevertheless, researchers use the corpus for different computational tasks, e.g. separating argumentative from non-argumentative sentences (Mochales-Palau and Moens, 2011), identifying argument components (Rooney et al., 2012) and classifying argumentation schemes (Feng and Hirst, 2011).

Mochales-Palau and Moens (2008) conduct an argument annotation study in legal cases of the *European Court of Human Rights* (ECHR). They experiment with a small corpus of 10 documents and obtain an inter-rater agreemnt of $\kappa = 0.58$. In a subsequent study, they elaborated their guidelines and obtain an inter-rater agreement of $\kappa = 0.75$ on a corpus of 47 documents (Mochales-Palau and Moens, 2011). Unfortunately, the annotation scheme is not described in detail, but it can be seen from the examples that it includes annotations for claims and supporting or refuting premises. Unlike our work, the annotation scheme does not include argumentative relations.

Cabrio and Villata (2012) annotate argumentative relations in debates gathered from *Debatepedia*. Instead of identifying argument components, they are interested in relations between arguments to identify which are the ones accepted by the community. They apply textual entailment for identifying support and attack relations between arguments and utilize the resulting structure for identifying accepted arguments. Therefore, they annotate a pair of arguments as either entailment or not. In contrast to our work, the approach models relationships between pairs of arguments and does not consider components of individual arguments. In addition, the work does not include an evaluation of the annotation's reliability.

Villalba and Saint-Dizier (2012) study argumentation annotation in a corpus of French and English product reviews. Their goal is to identify arguments related to opinion expressions for recognizing reasons of customer opinions. Their annotation scheme is limited to eight types of support (e.g. justification, elaboration, contrast). Compared to our annotation scheme, the work distinguishes between different premise types. However, the approach is tailored to product reviews, and the work does not provide an inter-rater agreement study.

In contrast to previous work, our annotation scheme includes argument components and argumentative relations. Both are crucial for argument recognition (Sergeant, 2013) and argumentative writing support. First, argumentative relations are essential for evaluating the quality of claims, since it is not possible to examine how well a claim is justified without knowing which premises belong to a claim (Sampson and Clark, 2006). Second, methods that recognize if a statement supports or attacks a claim enable the collection of additional evidence from other resources to recommend argument improvement. In addition, we provide a detailed analysis of the inter-rater agreement and an analysis of disagreements.

[1] http://www.ukp.tu-darmstadt.de/data/argumentation-mining

2.2 Persuasive Essays

Persuasive essays are extensively studied in the context of *automated essay grading* (Shermis and Burstein, 2013), which aims at automatically assigning a grade to a student's essay by means of several features. Since the argument structure is crucial for evaluating essay quality, Burstein et al. (1998) propose an approach for identifying the argumentative discourse structure by means of discourse marking. They utilize a surface cue word and phrase lexicon to identify the boundaries of arguments at the sentence level in order to evaluate the content of individual arguments and to enrich their feature set for determining precise grades. Although the identification of argument boundaries is important for argument recognition, our work allows a more fine-grained analysis of arguments since it also includes argument components and argumentative relations.

Madnani et al. (2012) studied persuasive essays for separating organizational elements from content. They argue that the detection of organizational elements is a step towards argument recognition and inferring the structure of persuasive discourse. Further, they refer to organizational elements as claim and premise indicating word sequences which they call *shell expressions*. They annotate 200 essays and estimate an inter-rater agreement of $\kappa = 0.699$ and $F_1 = 0.726$ on a subset of 50 essays annotated by two annotators. However, their annotation scheme is limited to shell expressions and compared to our work it does not include argument components or argumentative relations.

Additional annotation studies on persuasive essays focus on identifying style criteria (Burstein and Wolska, 2003), factual information (Beigman Klebanov and Higgins, 2012), holistic scores for argumentation quality (Attali et al., 2013) or metaphors (Beigman Klebanov and Flor, 2013). We are not aware of an annotation study including argument components and argumentative relations in persuasive essays.

3 Annotation Scheme

The goal of our proposed annotation scheme is to model argument components as well as argumentative relations that constitute the argumentative discourse structure in persuasive essays. We propose an annotation scheme including three argument components and two argumentative relations (figure 1).

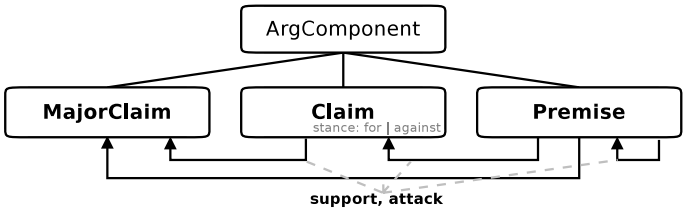

Figure 1: Argument annotation scheme including argument components and argumentative relations indicated by arrows below the components.

3.1 Argument Components

Persuasive essays exhibit a common structure. Usually, the introduction includes a *major claim* that expresses the author's stance with respect to the topic. The major claim is supported or attacked by arguments covering certain aspects in subsequent paragraphs. Sentences (1–3) illustrate three examples of major claims (the major claim is in bold face).[2]

(1) "*I believe that* ***we should attach more importance to cooperation during education***."
(2) "*From my viewpoint,* ***people should perceive the value of museums in enhancing their own knowledge***."
(3) "*Whatever the definition is,* ***camping is an experience that should be tried by everyone***."

In the first example, the author explicitly states her stance towards cooperation during education. The major claims in the second and third example are taken from essays about museums and camping

[2]We use examples from our corpus (5.1) without correcting grammatical or spelling errors.

respectively. In (1) and (2) a *stance indicating expression* (*"I believe"* and *"From my viewpoint"*) denotes the presence of the major claim. Although, these indicators are frequent in persuasive essays, not every essay contains an expression that denotes the major claim. In those cases, the annotators are asked to select the expression that is most representative with respect to the topic and author's stance (cf. (3)).

The paragraphs between introduction and conclusion of persuasive essays contain the actual arguments which either support or attack the major claim.[3] Since argumentation has been a subject in philosophy and logic for a long time, there is a vast amount of argumentation theories which provide detailed definitions of argument components (Toulmin, 1958; Walton et al., 2008; Freemen, 2011).[4] All these theories generally agree that an *argument* consists of several components and that it includes a *claim* that is supported or attacked by at least one *premise*. Examples (4) and (5) illustrate two arguments containing a claim (in bold face) and a premise (underlined).

(4) "***It is more convenient to learn about historical or art items online***. *With Internet, people do not need to travel long distance to have a real look at a painting or a sculpture, which probably takes a lot of time and travel fee.*"
(5) "***Locker checks should be made mandatory and done frequently*** *because they assure security in schools, makes students healthy, and will make students obey school policies.*"

The claim is the central component of an argument. It is a controversial statement that is either true or false and should not be accepted by readers without additional support. The premise underpins the validity of the claim. It is a reason given by an author for persuading readers of the claim. For instance, in (4) the author underpins his claim that Internet usage is convenient for exploring cultural items because of time and travel fee savings. In this example, both components cover a complete sentence. However, a sentence can also contain several argument components like in example (5). Therefore, we do not predefine the boundaries of the expression to be annotated (markable) in advance and annotate each argument as a *statement*, which is a sequence of words that constitutes a grammatically correct sentence.

To indicate if an argument supports or attacks a major claim, we add a *stance attribute* to the claim that denotes the polarity of an argument with respect to the author's stance. This attribute can take the values *for* or *against*. For example, the argument given in (4) refutes the major claim in example (2). Thus, the stance attribute of the claim in (4) is set to *against* in this example.

3.2 Argumentative Relations

Argumentative relations model the discourse structure of arguments in persuasive essays. They indicate which premises belong to a claim and constitute the structure of arguments. We follow the approach proposed by Peldszus and Stede (2013) and define two directed relations between argument components: *support* and *attack*.[5] Both relations can hold between a premise and another premise, a premise and a (major-) claim, or a claim and a major claim (figure 1). For instance, in example (4) the premise in the second sentence is a reason or justification for the claim in the first sentence and the claim in (4) attacks the major claim of example (2). Thus, an argumentative relation between two components indicates that the source component is a reason or a refutation for the target component. The following example illustrates a more complex argument including one claim and three premises.

(6) "***Living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet***. *One who is living overseas will of course struggle with loneliness, living away from family and friends*$_1$ *but those difficulties will turn into valuable experiences in the following steps of life*$_2$. *Moreover, the one will learn living without depending on anyone else*$_3$."

Figure 2 illustrates the structure of this argument. The claim is attacked by premise$_1$, whereas premise$_2$ is a refutation of premise$_1$. The third premise is another reason that underpins the claim in this paragraph.

[3] In some cases, the introduction or conclusion contains arguments as well, those are also annotated in the annotation study.
[4] A review of argumentation theory is beyond the scope of this paper but a survey can be found in (Bentahar et al., 2010)
[5] Peldszus and Stede also define a *counter-attacking relation* that is omitted in our scheme, since it can also be represented as a chain of attacking premises.

This shows that it is not necessary to explicitly distinguish between supporting and attacking premises, since the relational structure and the type of argumentative relations implicitly denote the role of argument components. Additionally, argumentative relations enable the modeling of relationships between pairs of arguments on the macro level, e.g., by linking claims to the major claim.

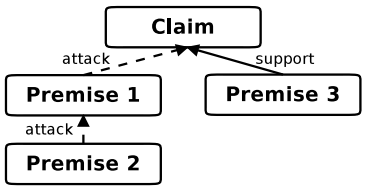

Figure 2: Argumentation structure of example (6)

4 Pre-Study

We conduct a preliminary study to define the annotation guidelines on a corpus of 14 short text snippets (1–2 sentences) that are either gathered from example essays or written by one of the authors. We ask five non-trained annotators to classify each text as argumentative or non-argumentative. If a text is classified as argumentative, the annotators are asked to identify the claim and the premise. In the first task, we obtain an inter-rater agreement of 58.6% and multi-$\pi = 0.171$ (Fleiss, 1971)[6]. We identified the markables for measuring the inter-rater agreement of the second task by manually determining the statements in each of the 14 text snippets. In total, we determined 32 statements and obtained an inter-rater agreement of 55.9% and multi-$\pi = 0.291$. These results indicate a low reliability of the annotations. In addition, they emphasize the demand for a precisely defined argument annotation strategy. In subsequent discussions, we discovered that the primary source of uncertainty is due to the missing context. Since the text snippets are provided without any information about the topic, the annotators found it difficult to decide if a snippet includes an argument or not. In addition, the annotators report that the author's stance might facilitate the separation of argumentative from non-argumentative text and to determine the components of arguments.

According to these findings, we define a new *top-down annotation process* starting with the major claim and drill-down to the claims and premises. Therefore, the annotators are aware of the author's stance after identifying the major claim. In addition, we ask the annotators to read the entire essay in order to identify the topic before starting with the actual annotation task. Although, this approach is more time-consuming than a direct identification of argument components, we show in our annotation study (section 5) that it yields reliably annotated data. In particular, the annotation guidelines consist of the following steps:

1. *Topic and stance identification*: Before starting with the annotation process, annotators identify the topic and the author's stance by reading the entire essay.

2. *Annotation of argument components*: In this step, the major claim is identified either in the introduction or in the conclusion of an essay. Subsequently, annotators identify the claims and premises in each paragraph. We instruct the annotators to annotate each argument component as a statement covering an entire sentence or less. We consolidate the annotations of all annotators before continuing with the next step (section 5.4).

3. *Annotation of argumentative relations*: Finally, the claims and premises are linked within each paragraph, and the claims are linked to the major claim either with a support or attack relation.

[6]Although the coefficient was introduced by Fleiss as a generalization of Cohen's κ (Cohen, 1960), it is actually a generalization of Scott's π (Scott, 1955), since it assumes a cumulative distribution of annotations by all annotators (Artstein and Poesio, 2008). We follow the naming proposed by Artstein and Poesio and refer to the measure as multi-π.

5 Annotation Study

Three annotators participate in the study and annotate the essays independently using our described annotation scheme. We conduct several training sessions after each annotator has read the annotation guidelines. In these sessions, annotators collaboratively annotate 8 example essays for resolving disagreements and obtaining a common understanding of the annotation guidelines. For the actual annotation task, we used the *brat annotation tool* that is freely available.[7] It allows the annotation of text units with arbitrary boundaries as well as the linking of annotations for modeling argumentative discourse structures.

5.1 Data

Our corpus consists of 90 persuasive essays in English, which we selected from *essayforum*[8]. This forum is an active community that provides writing feedback for different kinds of texts. For instance, students post their essays for retrieving feedback about their writing skills while preparing themselves for standardized tests. We randomly selected the essays from the *writing feedback* section of the forum and manually reviewed each essay. Due to the non-argumentative writing style and significant language flaws, we replaced 4 of them during a manual revision of the corpus. The final corpus includes 1,673 sentences with 34,917 tokens. On average, each essay has 19 sentences and 388 tokens.

5.2 Inter-rater Agreement

We evaluate the reliability of the argument component annotations using two strategies. Since there are no predefined markables in our study, annotators have to identify the boundaries of argument components. We evaluate the annotations using Krippendorff's α_U (Krippendorff, 2004). It considers the differences in the markable boundaries of several annotators and thus allows for assessing the reliability of our annotated corpus. In addition, we evaluate if a sentence contains an argument component of a particular category using percentage agreement and two chance-corrected measures: multi-π (Fleiss, 1971) and Krippendorff's α (Krippendorff, 1980). Since only 5.6% of the sentences contain several annotations of different argument components, evaluating the reliability at the sentence-level provides a good approximation of the inter-rater agreement. In addition, it enables comparability with future argument annotation studies that are conducted at the sentence-level. The annotations yield the following class distribution at the token-level: 3.5% major claim, 18.2% claim, 48.1% premise and 30.2% are not annotated. At the sentence-level 5.4% contain a major claim, 26.4% a claim, 61.1% a premise and 19.3% none annotation. Thus, 12.2% of the sentences contain several annotations.

	%	π	α	α_U
MajorClaim	.9827	.8334	.8365	.7726
Claim	.8690	.6590	.6655	.6033
Premise	.8618	.7075	.7131	.7594

Table 1: Inter-rater agreement of argument component annotations

We obtain the highest inter-rater agreement for the annotations of the major claim (table 1). The inter-rater agreement of 98% and multi-$\pi = 0.833$ indicates that the major claim can be reliably annotated in persuasive essays. In addition, there are few differences regarding the boundaries of major claims ($\alpha_U = 0.773$). Thus, annotators identify the sentence containing the major claim as well as the boundaries reliably. We obtain an inter-rater agreement of multi-$\pi = 0.708$ for premise annotations and multi-$\pi = 0.66$ for claims. This is only slightly below the "*tentative conclusion boundary*" proposed by Carletta (1996) and Krippendorff (1980). The unitized α of the major claim and the claim are lower than the sentence-level agreements (table 1). Only the unitized α of the premise annotations is higher compared to the sentence-level agreement. Thus, the boundaries of premises are more precisely identified. The joint unitized measure for all categories is $\alpha_U = 0.724$. Hence, we tentatively conclude that the annotation of argument components in persuasive essays is reliably possible.

[7] http://brat.nlplab.org

[8] http://www.essayforum.com

The agreement of the stance attribute is computed for each sentence. We follow the same methodology as for the computation of the argument component agreement, but treat each sentence containing a claim as either for or against according to the stance attribute (sentences not containing a claim are treated as not annotated, but are included in the markables). Thus, the upper boundary for the stance agreement constitutes the agreement of the claim annotations. The agreement of the stance attribute is only slightly below the agreement of the claim (86%; multi-$\pi = 0.643$; $\alpha = 0.65$). Hence, the identification of either attacking or rebutting claims is feasible with high agreement.

We determine the markables for evaluating the reliability of argumentative relations as the set of all pairs between argument components according to our annotation scheme. So, the markables correspond to all relations that were possible during the annotation task. In total, the markables include $5{,}137$ pairs of which 25.5% are annotated as support relation and 3.1% as attack relations. We obtain an inter-rater agreement above 0.8 for both support and attack relations (table 2) that is considered by Krippendorff (1980) as good reliability. Therefore, we conclude that argumentative relations can be reliably annotated in persuasive essays.

	$\%$	π	α
support	.9267	.8105	.8120
attack	.9883	.8052	.8066

Table 2: Inter-rater agreement of argumentative relation annotations

5.3 Error Analysis

To study the disagreements encountered during the annotation study, we created *confusion probability matrices* (CPM) (Cinková et al., 2012) for argument components and argumentative relations. A CPM contains the conditional probabilities that an annotator assigns a certain category (column) given that another annotator has chosen the category in the row for a specific item. In contrast to traditional confusion matrices, a CPM also enables the evaluation of confusions if more than two annotators are involved in an annotation study.

	Major Claim	Claim	Premise	None
Major Claim	.675	.132	.148	.045
Claim	.025	.552	.338	.086
Premise	.014	.163	.754	.069
None	.012	.123	.204	.660

Table 3: Confusion probability matrix for argument component annotations (Category 'None' indicates argument components that are not identified by an annotator.)

The major disagreement is between claims and premises (table 3). This could be expected since a claim can also serve as premise for another claim, and it is difficult to distinguish these two concepts in the presence of reasoning chains. For instance, examples (7–9) constitute a reasoning chain in which (7) is supported by (8) and (8) is supported by (9):

(7) "*Random locker checks should be made obligatory*."
(8) "*Locker checks help students stay both physically and mentally healthy*."
(9) "*It discourages students from bringing firearms and especially drugs*."

Considering this structure, (7) can be classified as claim. However, if (7) is omitted, (8) becomes a claim that is supported by (9). Thus, the distinction between claims and premises depends not only on the context and the intention of the author but also on the structure of a specific argument. Interestingly, the distinction between major claims and claims is less critical. Apparently, the identification of the major claim is easier since it is directly related to the author's stance in contrast to more general claims that cover a certain aspect with respect to the overall topic of the essay.

The CPM for relations (table 4) reveals that the highest confusion is between support/attack relations and none classified relations. This could be due to the fact that it is difficult to identify the correct target of a relation, especially in the presence of multiple claims or reasoning chains in a paragraph. For instance,

	support	attack	none
support	.750	.013	.238
attack	.104	.691	.205
none	.092	.001	.898

Table 4: Confusion probability matrix for argumentative relation annotations

in the previous example an annotator could also link (9) directly to (7) or even to (7) and (8). In both cases, the argument would be still meaningful. The distinction between support and attack relations does not reveal high disagreements. To sum up, the error analysis reveals that the annotation of argumentative relations yields more reliable results than that of argument components. This could be due to the fact that in our studies, argument components are known before annotating the relations and thus the task is easier. Nevertheless, it could be interesting to annotate relations before classifying the types of argument components and to investigate if it positively influences the reliability of annotations.

5.4 Creation of the Final Corpus

The creation of the final corpus consists of two independent tasks. First, we consolidate the argument components before the annotation of argumentative relations. So each annotator works on the same argumentative components when annotating the relations. Second, we consolidate the argumentative relations to obtain the final corpus. We follow a majority voting in both steps. Thus, an annotation is adopted in the final corpus if at least two annotators agree on the category as well as on the boundaries. In applying this strategy, we observed seven cases for argument components and ten cases for argumentative relations that could not be solved by majority voting. Those cases were discussed in the group of all annotators to reach an agreement. Table 5 shows an overview of the final corpus. It includes 90 major

	ALL	avg. per essay	standard deviation
Sentence	1,673	19	7
Tokens	34,917	388	124
MajorClaim	90	1	0
Claim	429	5	2
Claim (for)	365	4	2
Claim (against)	64	1	1
Premises	1,033	11	6
support	1,312	15	7
attack	161	2	2

Table 5: Statistics of the final corpus

claims (each essay contains exactly one), 429 claims and 1,033 premises. This proportion between claims and premises is common in argumentation and confirms the findings of Mochales-Palau and Moens (2011, p. 10) that claims are usually supported by several premises for "*ensuring a complete and stable standpoint*".

6 Conclusion & Future Work

We presented an annotation study of argument components and argumentative relations in persuasive essays. Previous argument annotation studies suffer from several limitations: Either they do not follow a systematic methodology and do not provide detailed inter-rater agreement studies or they do not include annotations of argumentative relations. Our annotation study is the first step towards computational argument analysis in educational applications that provides both annotations of argumentative relations and a comprehensive evaluation of the inter-rater agreement. The results of our study indicate that the annotation guidelines yield substantial agreement. The resulting corpus and the annotation guidelines are freely available to encourage future research in argument recognition.

In future work, we plan to utilize the created corpus as training data for supervised machine learning methods in order to automatically identify argument components as well as argumentative relations. In addition, there is a demand to scale the proposed annotation scheme to other genres e.g. scientific articles or newspapers and to create larger corpora.

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Piyush Paliwal and Krish Perumal for their valuable contributions and we thank the anonymous reviewers for their helpful comments.

References

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

Yigal Attali, Will Lewis, and Michael Steier. 2013. Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, 30(1):125–141.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-Relevant Metaphors in Test-Taker Essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, GA, USA.

Beata Beigman Klebanov and Derrick Higgins. 2012. Measuring the use of factual information in test-taker essays. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 63–72, Montreal, Quebec, Canada.

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

Jill Burstein and Magdalena Wolska. 2003. Toward evaluation of writing style: finding overly repetitive word use in student essays. In *Proceedings of the tenth conference of European chapter of the Association for Computational Linguistics*, EACL '03, pages 35–42, Budapest, Hungary.

Jill Burstein, Karen Kukich, Susanne Wolff, Ji Lu, and Martin Chodorow. 1998. Enriching Automated Essay Scoring Using Discourse Marking. In *Proceedings of the Workshop on Discourse Relations and Discourse Markers*, pages 15–21, Montreal, Quebec, Canada.

Jodie A. Butler and M. Anne Britt. 2011. Investigating Instruction for Improving Revision of Argumentative Essays. *Written Communication*, 28(1):70–96.

Elena Cabrio and Serena Villata. 2012. Natural language arguments: A combined approach. In *Proceedings of the 20th European Conference on Artificial Intelligence*, ECAI '12, pages 205–210, Montpellier, France.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Silvie Cinková, Martin Holub, and Vincent Kríž. 2012. Managing uncertainty in semantic tagging. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 840–850, Avignon, France.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

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 - Volume 1*, HLT '11, pages 987–996, Portland, OR, USA.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

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

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage.

Klaus Krippendorff. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Quality & Quantity*, 38(6):787–800.

Nitin Madnani, Michael Heilman, Joel Tetrault, and Martin Chodorow. 2012. Identifying High-Level Organizational Elements in Argumentative Discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 20–28, Montreal, Quebec, Canada.

Raquel Mochales-Palau and Marie-Francine Moens. 2008. Study on the Structure of Argumentation in Case Law. In *JURIX the twenty-first annual conference on legal knowledge and information systems*, pages 11–20, Florence, Italy.

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

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.

Chris Reed, Raquel Mochales-Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC '08, pages 2613–2618, Marrakech, Morocco.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, FLAIRS '12, pages 272–275, Marco Island, FL, USA.

Victor D. Sampson and Douglas B. Clark. 2006. Assessment of argument in science education: A critical review of the literature. In *Proceedings of the 7th International Conference on Learning Sciences*, ICLS '06, pages 655–661, Bloomington, IN, USA.

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

Alan Sergeant. 2013. Automatic argumentation extraction. In *Proceedings of the 10th European Semantic Web Conference*, ESWC '13, pages 656–660, Montpellier, France.

Mark D. Shermis and Jill Burstein. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge Chapman & Hall.

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

Maria Paz Garcia Villalba and Patrick Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. In *Proceeding of the 2012 conference on Computational Models of Argument*, COMMA '12, pages 23–34, Vienna, Austria.

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