The Impact of Modeling Overall Argumentation with Tree Kernels

Henning Wachsmuth
Bauhaus-Universität Weimar
Faculty of Media, Webis Group
henning.wachsmuth@uni-weimar.de

Giovanni Da San Martino
Qatar Computing Research Institute
Arabic Language Technologies
gmartino@hbku.edu.qa

Dora Kiesel
Bauhaus-Universität Weimar
Faculty of Media, VR Group
dora.kiesel@uni-weimar.de

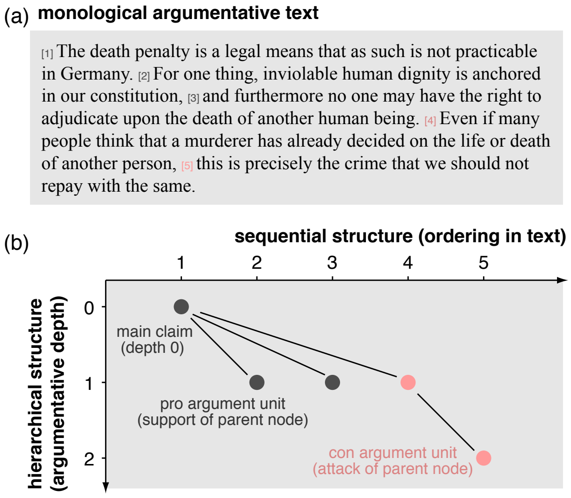
Benno Stein
Bauhaus-Universität Weimar
Faculty of Media, Webis Group
benno.stein@uni-weimar.de

Abstract

Several approaches have been proposed to model either the explicit sequential structure of an argumentative text or its implicit hierarchical structure. So far, the adequacy of these models of overall argumentation remains unclear. This paper asks what type of structure is actually important to tackle downstream tasks in computational argumentation. We analyze patterns in the overall argumentation of texts from three corpora. Then, we adapt the idea of positional tree kernels in order to capture sequential and hierarchical argumentative structure together for the first time. In systematic experiments for three text classification tasks, we find strong evidence for the impact of both types of structure. Our results suggest that either of them is necessary while their combination may be beneficial.

1 Introduction

Argumentation theory has established a number of major argument models focusing on different aspects, such as the roles of an argument's units (Toulmin, 1958), the inference scheme of an argument (Walton et al., 2008), or the support and attack relations between arguments (Freeman, 2011). The common ground of these models is that they conceptualize an argument as a conclusion (in terms of a claim) inferred from a set of pro and con premises (reasons), which in turn may be the conclusions of other arguments. For the overall argumentation of a monological argumentative text such as the one in Figure 1(a), this results in an implicit hierarchical structure with the text's main claim at the lowest depth. In addition, the text has an explicit linguistic structure that can be seen as a regulated sequence of speech acts (van Eemeren and Grootendorst, 2004).



[1] The death penalty is a legal means that as such is not practicable in Germany. [2] For one thing, inviolable human dignity is anchored in our constitution, [3] and furthermore no one may have the right to adjudicate upon the death of another human being. [4] Even if many people think that a murderer has already decided on the life or death of another person, [5] this is precisely the crime that we should not repay with the same.

(b)
sequential structure (ordering in text)
1
2
3
4
5
hierarchical structure (argumentative depth)
0
1
2
main claim (depth 0)
pro argument unit (support of parent node)
con argument unit (attack of parent node)

Figure 1: (a) Example text with five argument units, taken from the *Arg-Microtexts* corpus introduced in Section 3. (b) Graph visualization of the sequential and hierarchical overall argumentation of the text.

Figure 1(b) illustrates the interplay of the two types of overall structure in form of a tree-like graph.

Natural language processing research has largely adopted the outlined hierarchical models for mining arguments from text (Stab and Gurevych, 2014; Habernal and Gurevych, 2015; Peldszus and Stede, 2016). However, the adequacy of the resulting overall structure for downstream analysis tasks of computational argumentation has rarely been evaluated (see Section 2 for details). In fact, a computational approach that can capture patterns in hierarchical overall argumentation is missing so far. Even more, our previous work indicates that a sequential model of overall structure is preferable for analysis tasks such as stance classification or quality assessment (Wachsmuth and Stein, 2017).

In this paper, we ask and investigate what model of (monological) overall argumentation is important to tackle argumentation-related analysis tasks. To this end, we consider three corpora with fully annotated argument structure (Section 3). Each corpus allows studying one text classification task, two of which we hypothesize to benefit from modeling argumentation (myside bias, stance), the third not (genre). An empirical analysis of the corpora reveals class-specific patterns of how people argue (Section 4). In order to combine the explicit sequential and the implicit hierarchical structure of an argumentative text for the first time, we then adapt the approach of *route kernels* (Aiolli et al., 2009), modeling overall argumentation in form of a positional tree (Section 5).

On this basis, we design an experiment to evaluate the impact of the different types of argumentative structure (Section 6). In particular, we decompose our approach into four complementary modeling steps, both for a general model of overall argumentation and for the specific models of the given corpora. Using the structure annotated in the corpora, we systematically compare the effectiveness of all eight resulting models and two standard baselines in the three classification tasks.

Our results provide strong evidence that both sequential and hierarchical structure are important. As indicated by related work, sequential structure nearly competes with hierarchical structure, at least based on the specific argument models. Even more impressively, modeling hierarchical structure practically solves the task of identifying argumentation with myside bias, achieving an outstanding accuracy of 97.1%. For stance classification, the combination captured by positional trees works best. In contrast, all types of structure fail in distinguishing genres, suggesting that they indeed capture properties of argumentation. We conclude that the impact of modeling overall structure on downstream analysis tasks is high, while the required type may vary.

Contributions To summarize, the main contributions of this paper are the following:

1. Empirical insights into how people structure argumentative texts in overall terms.
2. The first approach to model and analyze the sequential and hierarchical overall structure of argumentative texts in combination.
3. Evidence that modeling overall structure impacts argumentation-related analysis tasks.

2 Related Work

The study of overall argumentation traces back to Aristotle (2007) who outlined the impact of the sequential arrangement of the different parts of a speech. Conceptually, theory agrees that a monological argumentative text has an implicit tree-like hierarchical structure: Toulmin (1958) defines an argument as a claim supported by data that is reasoned by a warrant, which in turn may have a backing. In addition, a rebuttal may be given showing exceptions to the claim. The role of support and attack relations is investigated by Freeman (2011) who models dialectical arguments that discuss both a proponent's and an opponent's view on the main claim argued for. Walton et al. (2008) put the focus on the inference scheme that describes how an argument's conclusion follows from its premises, which may themselves be conclusions of arguments.

In natural language processing, argumentation research deals with the mining of argument units and their relations from text (Mochales and Moens, 2011). Several corpora with annotated argument structure have been published in the last years. Many of the corpora adapt the hierarchical models from theory (Reed and Rowe, 2004; Habernal and Gurevych, 2015; Peldszus and Stede, 2016) or propose comparable models (Stab and Gurevych, 2014). Since we target monological overall argumentation, we use those that capture the complete structure of texts, as detailed in Section 3. Corpora focusing on dialogical argumentation (Walker et al., 2012), topic-related arguments (Rinott et al., 2015), or sequential structure (Wachsmuth et al., 2014b; Al Khatib et al., 2016) are out of scope.

We do not mine the structure of argumentative texts, but we exploit the previously mined structure to tackle downstream tasks of computational argumentation, namely, to classify the myside bias and stance of texts. For myside bias, Stab and Gurevych (2016) use features derived from discourse structure, whereas Faulkner (2014) and Sobhani et al. (2015) model arguments to classify stance. Ong et al. (2014) and we ourselves (Wachsmuth et al., 2016) do similar to assess the quality of persuasive essays, and Beigman Klebanov et al. (2016) examine how an essay's content and structure influence quality. Other works predict the outcome of legal cases based on the applied types of reasoning (Brüninghaus and Ashley, 2003) or analyze inference schemes for given arguments (Feng and Hirst, 2011). In contrast to the local structure of single arguments employed by all these approaches, we study the impact of the global overall structure of complete monological argumentative texts.

In (Wachsmuth et al., 2017), we point out that the argumentation quality of a text is affected by interactions of its content at different levels of granularity, from single argument units over arguments to overall argumentation. Stede (2016) explores how different depths of overall argumentation can be identified, observing differences across genres. Unlike in our experiments, however, the genres considered there reflect diverging types of argumentation. Related to argumentation, Feng et al. (2014) build upon rhetorical structure theory (Mann and Thompson, 1988) to assess the coherence of texts, while Persing et al. (2010) score the organization of persuasive essays based on sequences of sentence and paragraph functions.

We introduced the first explicit computational model of overall argumentation in (Wachsmuth et al., 2014a). There, we compared the flow of local sentiment in a review to a set of learned flow patterns in order to classify global sentiment. Recently, we generalized the model in order to make flows applicable to any type of information relevant for argumentation-related analysis tasks (Wachsmuth and Stein, 2017). However, flows capture only *sequential* structure, whereas here we also model the *hierarchical* structure of overall argumentation. To this end, we make use of kernel methods.

Kernel methods are a popular approach for learning on structured data, with several applications in natural language processing (Moschitti, 2006b) including argument mining (Rooney et al., 2012). They employ a similarity function defined between any two input objects that are represented in a task-specific implicit feature space. The evaluation of such a kernel function relies on the common features of the input objects (Cristianini and Shawe-Taylor, 2000). The kernel function encodes knowledge of the task in the form of these features.

Several kernel functions have been defined for structured data. To assess the impact of sequential argumentation, we refer to the function of Mooney and Bunescu (2006), which computes common subsequences of two input sequences. For trees, most existing approaches count common subtrees of a certain type (Collins and Duffy, 2001; Moschitti, 2006a; Kimura and Kashima, 2012), but they do not take the ordering of the nodes in the subtrees into account. In contrast, Aiolli et al. (2009) developed a kernel that combines the content of substructures with the relative positions inside trees, called the *route kernel*. Similarly, the tree kernel of Daumé III and Marcu (2004) includes positional information for document compression. For overall argumentation, we decided to use the route kernel in Section 5, as it makes the modeling of the sequential positions of an argument unit in a text straightforward. This allows us to capture both the sequential and the hierarchical structure at the same time. To our knowledge, no work has done this before.[1]

Neural networks denote an alternative for learning on structured data. They become particularly effective when few prior knowledge about what is important to address a task at hand is available, because they can learn any feature representation in principle (Goodfellow et al., 2016). Due to this flexibility, however, large amounts of data are required for training an effective model, making neural networks inadequate for the small datasets that allow studying overall argumentation.

3 Tasks and Datasets

We seek to study the impact of modeling overall argumentation on downstream tasks without the noise from argument mining errors. To this end, we rely on three ground-truth argument corpora. Each corpus is suitable for evaluating one text classification task and comes with a specific model of overall argumentation, as detailed in the following.

Myside Bias on AAE-v2 The *Argument Annotated Essays* corpus was originally been presented by Stab and Gurevych (2014). We use version 2 of the corpus (available on the website of the authors), which consists of 402 persuasive student essays. In each essay, all *main claims*, *claims*, and *premises* are annotated as such. Each claim has a *pro* or *con* stance towards each instance of the main claim, whereas each premise *supports* or *attacks* a claim or another premise. Thereby, argumentation is modeled as one tree structure for each major claim.

Stab and Gurevych (2016) added a *myside bias* class to each essay, reflecting whether its argumentation is one-sided considering only arguments for the own stance (251 cases) or not (151 cases).

Stance on Arg-Microtexts The *Arg-Microtexts* corpus of Peldszus and Stede (2016) contains 112 short argumentative texts. They cover 18 different controversial topics and are annotated according to Freeman (2011): Each argument unit takes the role of the *proponent* or *opponent* of a main claim. What

[1] While extensions of the route kernel idea have been published later on (Aiolli et al., 2011, 2015), we resort to the original version in this paper for simplicity.

	AAE-v2	Arg-Microtexts	Web Discourse
Argument units	6089	576	1149
Avg. units/text	15.1	5.1	3.4
Min. units/text	7	3	0
Max. units/text	28	10	16
Arguments	5687	443	560
Avg. depth	2.8	2.0	0.6
Min. depth	2	1	0
Max. depth	5	4	1
Texts	402	112	340

Table 1: Statistics of the argument units and arguments in the three corpora analyzed in this paper.

the main claim is follows from a tree-like overall structure emerging from four types of relations: *normal* or *example* support from one unit to another, a *rebuttal* of units by other units, and *undercutters* where a relation is attacked by another unit.

For 88 texts, the *stance* towards a specified topic is labeled as *pro* (46) or *con* (42). We use these labels for classification, but we do not access the topic. This way, stance needs to be identified only based on a text itself — a very challenging task.[2]

Genre on Web Discourse Finally, we consider the *Argument Annotated User-Generated Web Discourse* corpus of Habernal and Gurevych (2015). There, 340 texts are annotated according to a modified version of the specific model of Toulmin (1958) where *claims* are supported by *premises* or attacked by *rebuttals*. *Rebuttals* in turn may be attacked by *refutations*. Besides, emotional units not participating in the actual arguments are marked as *pathos*. The support and attack relations build up the overall argumentation of a text.

The corpus composes argumentative texts of four *genres*, namely, 5 *articles*, 216 *comments* to articles, 46 *blog posts*, and 73 *forum posts*. The genre is specified in form of a label for each text. Due to the low number, we ignore the articles below.

To give an idea of the sequential and hierarchical overall structure in each corpus, Table 1 presents statistics of the argument units, the arguments (in terms of relations between two or more units), and the depth of the resulting argumentation.

While the size of the given corpora and the variety of tasks are limited, the only other available corpus with fully annotated argument structure that we are aware of is *AraucariaDB* (Reed and Rowe, 2004). No downstream task can be tackled on AraucariaDB besides inference scheme classification (Feng and Hirst, 2011). As all schemes compose a conclusion and a set of premises (without more specific roles), analyzing overall structure hardly makes sense, which is why we omit the corpus.

4 Insights into Overall Argumentation

Before we approach overall argumentation computationally, this section analyzes the three given corpora empirically to provide insights into how people argue in overall terms. For this, we unify the specific corpus models of overall argumentation outlined above in one general model.

4.1 A Unified View of Overall Argumentation

The texts in all corpora are segmented into argument units, partly with non-argumentative spans in between that we ignore here for lack of relevance. To capture the sequential ordering of the segmentation, we assign a global index to each unit.

As described in Section 3, the specific models of all three corpora in the end consider an argument as a composition of one unit serving as the conclusion with one or more units that support or attack the conclusion (the premises). This composition is defined through multiple relations from one premise to one conclusion each. There is one exception, namely, the undercutter relations in the Arg-Microtexts corpus have a relation as their target. To obtain a unified form in the general model, we modify the undercutters such that they target the premise of the undercutted relation.

In all corpora, a premise may be the conclusion of another argument, while no argument unit serves as a premise in multiple arguments. This leads to a tree structure for each main claim of the associated text. A main claim corresponds to a unit that is not a premise. In AAE-v2 and in Web Discourse, more than one such unit may exist per text.

Depending on the corpus, the distinction of support and attack is encoded through a specified relation type, a unit's stance, or both. We unify these alternatives by modeling the stance of each unit towards its parent in the associated tree. This stance can be derived in all corpora.[3] All other unit and relation types from the specific models are ignored, since there is no clear mapping between them.

[2] We do not include the topic, in order not to conflate the impact of modeling argumentation with the influence of the topic. The corpus is too small to analyze topic differences.

[3] Alternatively, the stance towards the main claim could be modeled. We decided against this alternative to avoid possibly wrong reinterpretations, e.g., it is unclear whether a unit that attacks its parent always supports a unit attacked by the parent.

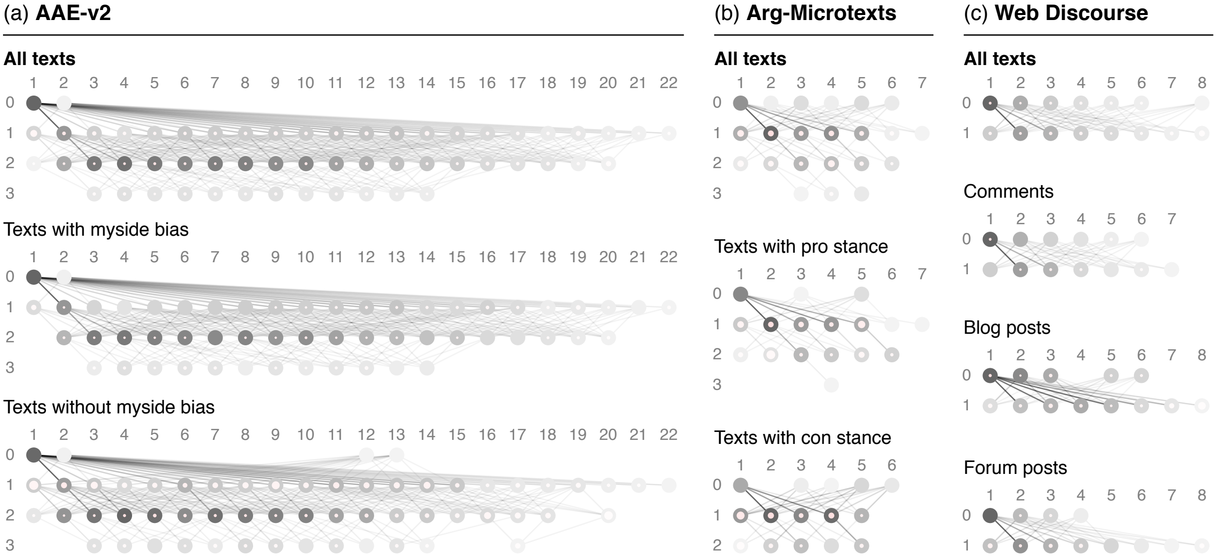


Figure 2: Visualization of the overall argumentation in the three considered corpora based on the introduced general model, averaged over all texts as well over all those texts that belong to a particular class. Left to right: Position in the text. Top to bottom: Depth in graph. Brightness: Inverse relative frequency of each position/depth combination in the corpus. Light red to gray: Proportion of argument units with con stance.

General Model As a result, we model the overall argumentation of an argumentative text as a forest of trees. Each node in a tree corresponds to an argument unit. It has an assigned stance (pro or con) as well as a global index that defines its position in the text. Each edge defines a relation from a premise (the child node) to a conclusion (the parent node). Each main claim defines the root of a tree.

Figure 1(b) has already illustrated an instance of the general model. The general model is slightly less expressive than the specific models. We evaluate in Section 6 to what extent this reduces its use for tackling argumentation-related analysis tasks. The advantage of the general model is that it allows a comparison of patterns of overall argumentation across corpora, as we do in the following.[4]

4.2 Visualization of Argumentation Patterns

Based on the general model, we empirically analyze class-specific patterns of overall argumentation on the three corpora. To this end, we compute one "average graph" for all texts in each complete corpus and one such graph for all texts with a particular class (e.g., for all "no myside bias" texts in case of AAE-v2). In an average graph, each node is labeled with the relative frequency of the associated combination of position and depth in all texts (edges accordingly). We align positions of different texts based on their start node, due to our observation that the first argument unit overproportionally often represents the main claim.[5] In addition the relative frequency, we determine the proportion of con to pro stance for each node.

As we aim to provide intuitive insights into how people argue in overall terms, we discuss the graphs in an informal visual way instead of listing exact numbers.[6] In the visualizations in Figure 2, brightness captures (inverse) frequency, so darker nodes represent more frequent argument units. The diameter of the inner light-red part of each node reflects its proportion of con stance. Nodes with a relative frequency below 0.3% and/or an absolute frequency below 3 are pruned, along with all their associated edges.

AAE-v2 Figure 2(a) stresses that most students state the main claim (depth 0, position 1) in a persuasive essay first. When the first argument unit is a premise of the main claim instead, it often attacks the main claim, as the large light-red proportion of the node at depth 1 and position 1 conveys. While, on average, texts with myside bias do not differ in length from those without, the latter show more con stance, especially at depth 1. Also, argumenta-

[4] Besides, although not in the focus here, we also assume stance to be easier to detect in practice than fine-grained roles.

[5] We also considered using the main claim as the fix point, but the resulting graphs would be much wider than the longest argumentation, which may be misleading.

[6] We provide files with the exact frequencies of all nodes and edges at: http://www.arguana.com/software.html

tion without myside bias shows more variance, as indicated, for instance, by the nodes at depth 0 and position 12 and 13 respectively. In contrast, clear patterns in the sequential ordering of pro and con stance are not recognizable in AAE-v2.

Arg-Microtexts According to the graphs in Figure 2(b), the position of the main claim varies in the microtexts. While the proportion of con stance seems rather similar between pro and con texts, our visualization reveals that their overall structure is "mirror-inverted" to a limited extent: Most pro texts start with the main claim (depth 0, position 1), discuss con stance later (red proportions increase to the right), and deepen the argumentation in a top-down fashion (most edges from top left to bottom right). Vice versa, con texts more often present the main claim later, attack it earlier, and seem to argue more bottom-up. This suggests that both sequential and hierarchical structure play a role here.

Web Discourse The web discourse texts, finally, comprise rather shallow argumentation across all genres. Slight structural differences can be seen, especially, the comments appear a little shorter and richer of pro stance on average. Besides, the blog posts have more con stance later. Still, the darker and thus more frequent nodes are at similar positions in all graphs. So, if at all, differences may be reflected in a sequential model of argumentation, which implicitly covers length. In terms of the hierarchical structure of the frequent nodes, the graphs of all genres are rather indistinguishable.

Altogether, the visualizations give first support for the impact of modeling overall argumentation. In particular, we hypothesize that hierarchical overall structure is decisive for myside bias, whereas a combination of sequential and hierarchical structure helps to distinguish pro-stance from con-stance texts. In contrast, we expect that the impact on classifying genres in the Web Discourse corpus is low.

5 Modeling Overall Argumentation

This section presents our kernel-based approaches for argumentation-related analysis tasks. They rely on a tree representation of overall argumentation.

5.1 Representation of Overall Argumentation

We model the overall structure of an argumentative text in form of a positional tree $T = (V, E)$ that, in principle, equals those exemplified in Figure 1 and analyzed above. Each node $v \in V$ represents an argument unit and each edge $e = (v_1, v_2) \in E$ a relation between two units. Technically, we therefor map the forest of trees representing a text (see Section 4) to a single tree by adding a "virtual" root node v_0 to V that is the parent of all tree roots.

In analysis tasks, we seek to compare sequential and hierarchical structures irrespective of the actual texts and the size of the associated trees. To this end, we represent labels and positions as follows:

Labels The tree kernel approaches in natural language processing discussed in Section 2 include text (usually words) in the leaf nodes. In contrast, we label each node $v \in V$ with the type of the associated argument unit only. Thereby, we almost fully abstract from the content of texts, which benefits the identification of common structures. In case of the general model, the only two labels are *pro* and *con*. In case of the specific models, we combine the role of a unit with the type of the relation the unit is the source of (if any). On Arg-Microtexts, for instance, this creates labels such as *opponent-support* or *opponent-undercutter*.

Positions As we adapt the route kernels of Aiolli et al. (2009) below, we follow their representation of sequential structure with one exception. In particular, the authors assigned an index to each edge that numbers the child nodes of each node ascending from 1. Thereby, they encoded the relative positions of sibling nodes *to each other*. To capture the ordering of argument units in a text from left to right, we also model positions as indices of the edges in E. Unlike Aiolli et al. (2009), however, we use indices decreasing from -1 in the left direction of the parent node and ascending from 1 to the right (derived from the nodes' global indices). While such a simple relabeling allows us to reuse their algorithm for computing kernels, it makes a decisive difference, namely, it encodes the relative positions of child nodes *to their parent*. This in turn implies the sequential structure of the whole tree.

Figure 3(a) exemplifies the tree representation for the argument unit types of the general model, omitting the virtual root v_0 for simplicity. Analogously, the types of the specific models of the three considered corpora could be used.

5.2 Kernel-based Modeling Approaches

Based on the tree representation, we now introduce four approaches for modeling overall argumentation. Figure 3(b) illustrates the kernel representations of each approach. As discussed in Section 2,

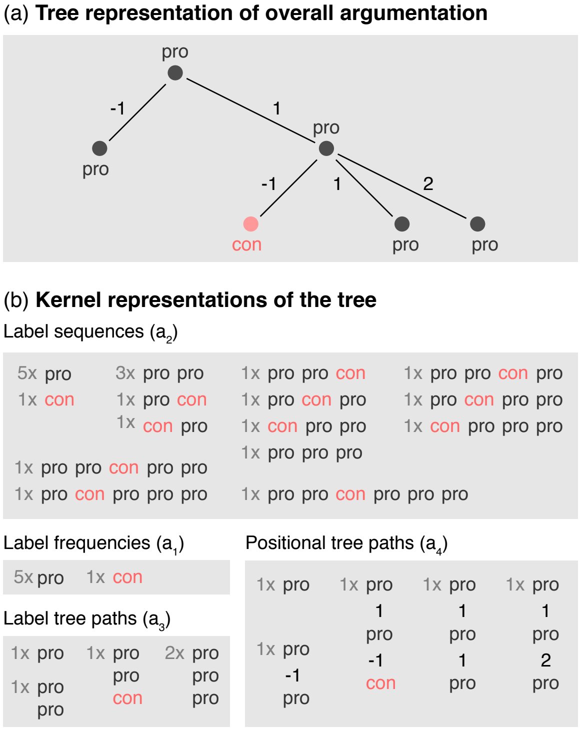


Figure 3: (a) Tree representation exemplified for the general model; node labels are unit types, edge indices relative positions of child nodes. (b) Kernel representations of the tree for all four approaches.

the associated kernel function compares the representations of the trees T, T' of any two texts.

Label Frequencies (a_1) Our simplest model of overall argumentation does not encode structure at all. Instead, it compares only the frequencies of each node label in T and T'. We represent the model with a linear kernel, which in the end corresponds to a standard feature representation.

Label Sequences (a_2) To encode sequential overall structure, we refer to the kernel of Mooney and Bunescu (2006), representing the sequential ordering of node labels in a tree by all contiguous subsequences. The similarity of two trees T and T' follows from the proportion of common subsequences, but longer subsequences are penalized by a decay factor. This approach can be seen as an imitation of our flow model (Wachsmuth and Stein, 2017).[7]

Label Tree Paths (a_3) We capture hierarchical overall structure adapting the non-positional part of the route kernel of Aiolli et al. (2009), *label paths*. A label path $\xi(v_i, v_j)$ denotes the sequence of labels of the nodes in the shortest path between v_i, v_j in a tree (including v_i, v_j). Following Aiolli et al. (2009), we consider only label paths starting at the root $v_i = v_0$, abbreviated here as $\xi(v_j)$. Implicitly, other paths may still be considered through the use of polynomial kernels with degree $d > 1$. As the authors, we compare any two paths with a function δ whose values is 1 when the paths are identical and 0 otherwise. Given two trees $T = (V, E)$ and $T' = (V', E')$, we then define a normalized polynomial kernel $K_\xi(T, T')$ over all label paths as:

$$\left(\sum_{v \in V} \sum_{v' \in V'} \frac{\delta(\xi(v), \xi(v'))}{|V| \cdot |V'|} \right)^d$$

Positional Tree Paths (a_4) In addition to label paths, Aiolli et al. (2009) define a *route* $\pi(v_i, v_j)$ as the sequence of edge indices on the shortest path between any two nodes v_i, v_j in a tree, i.e., the sequence of local positions. As above, they restrict their view to routes starting at the root, which we denote as $\pi(v_j)$, and compare them using δ. To combine positional information with label information, the authors build the product of a kernel based on the label paths and a kernel based on routes. As a result, sequential and hierarchical overall structure are compared at the same time. For overall argumentation, we define the resulting normalized polynomial product kernel $K_{\xi\pi}(T, T')$ as:

$$\left(\sum_{v \in V} \sum_{v' \in V'} \frac{\delta(\xi(v), \xi(v')) \cdot \delta(\pi(v), \pi(v'))}{(|V| \cdot |V'|)^2} \right)^d$$

Each approach, a_1–a_4, can be seen as representing one particular step of modeling overall argumentation; a_4 combines the complementary steps of a_2 and a_3, both of which implicitly include a_1.

6 Evaluation

Finally, we evaluate all four approaches to model overall argumentation from Section 5 on the three tasks associated to the corpora from Section 3.[8]

6.1 Experimental Set-up

Our goal is to assess the theoretical impact of each introduced step of modeling overall argumentation as far as possible. To this end, we conduct a systematic experiment where we use the ground-truth argument structure in each corpus for the associated downstream task based on the following set-up:

[7] We use a sequence kernel instead of flows in order to obtain a uniform setting. In Wachsmuth and Stein (2017), we also analyze flow abstractions (e.g., collapsing sequences of the same label). Here, we resort only to the original sequence.

[8] The Java source code for reproducing the experiment results is available at: http://www.arguana.com/software.html

		Myside Bias on AAE-v2		**Stance on Arg-Microtexts**		**Genre on Web Discourse**	
#	**Approach**	*General model*	*Specific model*	*General model*	*Specific model*	*General model*	*Specific model*
b_1	POS n-grams	63.3	63.3	58.8	58.8	74.0 (99.9% >a_2)	74.0 (99.9% >a_2)
b_2	Token n-grams	*70.5* (95% >b_1)	*70.5* (95% >b_1)	*65.2* (99% >a_2)	*65.2*	*75.6* (99.9% >a_2)	*75.6* (99.9% >a_2)
a_1	Label frequencies	83.4 (99.9% >b_2)	85.7 (99.9% >b_2)	49.7	54.4	62.6 (95% >a_4)	61.4
a_2	Label sequences	87.9 (99.9% >b_2)	94.7 (99.9% >a_1)	52.2	62.3	*64.5* (95% >a_3)	*64.5* (99.9% >a_3)
a_3	Label tree paths	***97.1*** (99.9% >a_2)	***97.1*** (95% >a_2)	59.8 (95% >a_1)	61.9	58.1	55.5
a_4	Positional tree paths	95.8 (99.9% >a_2)	95.6 (99.9% >a_1)	*66.7* (99% >a_2)	*67.8* (95% >a_1)	53.4	55.2
ba	**Best b_i + Best a_j**	**97.1** (99.9% >a_2)	**97.1** (95% >a_2)	69.8 (99.9% >a_2)	**71.0** (95% >a_1)	75.7 (99.9% >a_2)	**75.9** (99.9% >a_2)
	Majority baseline	62.4	62.4	52.3	52.3	64.5	64.5

Table 2: Accuracy in 10-fold cross-validation (10 repetitions, fairness in training) of all evaluated approaches on each of the three task/corpus combinations, both based on a *general model* of arguments and based on the *specific model* of the respective corpus. The highest value on each corpus is marked in bold; the best b_i and a_j in each column are italicized. In parenthesis: The confidence level in percent at which the respective approach is significantly better than the specified approach and all worse approaches.

Approaches The modeling steps are reflected by the approaches a_1–a_4 from Section 5. For each task, we measure the accuracy of all four approaches. We do this once for our *general model* of overall argumentation from Section 4 and once for the *specific model* annotated in the respective corpus, in order to assess the loss of resorting to our always applicable general model.

Baselines As a basic task-intrinsic measure, we compare a_1–a_4 to the *majority baseline* that always predicts the majority class in the given corpus. In addition, we employ two standard feature types and combine them with a_1–a_4, in order to roughly assess the need for modeling argumentation:

- b_1 *POS n-grams.* The frequency of each part-of-speech 1- to 3-gram found in $\geq 5\%$ of all texts. This style feature has been effective in argumentation-related analysis tasks (Persing and Ng, 2015; Wachsmuth et al., 2016).
- b_2 *Token n-grams.* The frequency of each token 1- to 3-gram found in $\geq 5\%$ of all texts. This content feature is strong in many text analysis tasks (Joachims, 1998; Pang et al., 2002).

From the tackled tasks, only myside bias has been approached on the given datasets in previous work. While we mention the respective results for completeness below, a comparison is in fact unfair due to our resort to ground-truth argument structure.

Experiments The evaluation of all approaches and baselines was done using the kernel-based machine learning platform *KeLP* (Filice et al., 2015), performing classification with the available implementation of *LibSVM* (Chang and Lin, 2011). As we target the theoretically possible impact of modeling overall argumentation, we tested a number of hyperparameter configurations.[9] We performed 10-fold cross-validation on the complete corpora and repeated each experiment 10 times, with instance shuffling in between. Then, we averaged the accuracy of each configuration over all folds and repetitions. To prevent the classifiers from using knowledge about the class distributions, we used fairness during training, i.e., each class was given an equal weight (Filice et al., 2014). Thus, the majority baseline is not a trivial competitor.

6.2 Results

Table 2 presents the best obtained results of each evaluated approach for each task/corpus combination. To clarify the reliability of the differences between the results, the table includes the confidence level (starting at 95%) at which each approach is significantly better than all weaker approaches according to a two-tailed paired student's t-test.[10]

Myside Bias on AAE-v2 The highest accuracy reported for classifying myside bias is 77.0 (Stab and Gurevych, 2016). While the comparability is limited (see above), we see that label frequencies (a_1) already achieve 83.4 and 85.7 for the general and specific model respectively, outperforming all baselines with 99.9% confidence. Matching the insights from Section 4, the sole proportion of attacks thus seems a good predictor of myside bias.

[9] SVM C parameter: 0.01, 0.1, 1, 10, 100; sequence kernel decay factor: 0, 0.5, 1; polynomial tree kernel degree: 1, 2, 3.

[10] While selecting the best result a posteriori gives an upper bound on the true effectiveness, we do this to assess to what extent each approach captures task-relevant information.

Label sequences (a_2) further improve over a_1, which underlines that also the sequential position of con stance and attack relations has an impact. a_2 is particular strong under the specific model (94.7). Unlike the general model, this model reflects some hierarchical information via the roles of argument units, such as *premise*. a_2 performs only slightly worse than the label tree paths (a_3), indicating that an adequate sequential model can compete with a hierarchical model, as we hypothesized in previous work (Wachsmuth and Stein, 2017).

Nevertheless, a_3 turns out best on AAE-v2, most likely due to its capability to capture the depth at which con stance occurs. Considering that no corpus annotation is perfect, the outstanding accuracy of 97.1 conveys an important finding: Modeling the tree structure of an argumentation basically *solves* the myside bias task without requiring other features. Neither the positional tree paths (a_4) nor the combination with token n-grams (**ba**) can add to that. Also, there is no difference between the general and the specific model, underlining that the unit roles in AAE-v2 are implicitly covered by the hierarchical structure in the general model.

Stance on Arg-Microtexts The accuracy results for the given challenging variant of stance classification (see Section 3) are much lower. Under the general model, the label frequencies (49.7) do not even compete with the majority baseline (52.3). Notable gains are achieved by the label sequences under the specific model (62.3), slightly beating the label tree paths (61.9). Putting them together in the positional tree paths (a_4) yields an accuracy of 66.7 and 67.8 respectively; more than the token n-grams (b_2, 65.2). Combining a_4 and b_2 in **ba** in turn results in the best observed accuracy value (71.0 on the specific model).

We conclude that both sequential and hierarchical overall structure are important for the distinction of pro from con argumentation, supporting our hypothesis from Section 4. They complement content-oriented approaches, such as b_2. Moreover, the fine-grained unit and relation types of the specific model annotated in Arg-Microtexts seem useful, consistently obtaining higher accuracy than the general model. Notice, though, that due to the small size of the corpus, only few reported gains are statistically significant, as shown in Table 2.

Genre on Web Discourse Although Section 4 has made minor structural differences in Web Discourse visible, Table 2 shows that a_1–a_4 all fail in genre classification: None of them beats the majority baseline (64.5), suggesting that no decisive discriminative patterns are learned. Both POS and token n-grams (b_1–b_2) significantly outperform a_1–a_4 at 99.9% confidence. While combining b_2 with a_2 (**ba**) minimally increases accuracy from 75.6 to 75.9, the results reveal that overall argumentation hardly impacts genre — as hypothesized.

7 Conclusion

This paper provides answers to the question of how the overall structure of a monological argumentative text should be modeled in order to tackle downstream tasks of computational argumentation. We have adopted the idea of including positional information in tree kernels in order to capture the explicit sequential and the implicit hierarchical overall structure of the text at the same time. In systematic experiments, we have demonstrated the strong impact of modeling overall argumentation. Most impressively, we have found that hierarchical structure decides about myside bias alone, while the combination of sequential and hierarchical structure has turned out beneficial for classifying stance. The missing impact on genre supports that the presented approaches actually capture argumentation-related properties of a text.

So far, however, we have restricted our view to ground-truth argument structure, leaving the integration of computational argument mining approaches to future work. While the noise from mining errors might qualify some of our findings, we also expect that larger corpora will allow us to discover more reliable and discriminative patterns. After all, our results underline the general importance of modeling overall argumentation.

References

Fabio Aiolli, Giovanni Da San Martino, and Alessandro Sperduti. 2009. Route kernels for trees. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 17–24.

Fabio Aiolli, Giovanni Da San Martino, and Alessandro Sperduti. 2011. Extending tree kernels with topological information. In *Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I*, pages 142–149.

Fabio Aiolli, Giovanni Da San Martino, and Alessandro Sperduti. 2015. An efficient topological distance-based tree kernel. *IEEE Transactions on Neural Networks and Learning Systems*, 26(5):1115–1120.

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443. The COLING 2016 Organizing Committee.

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, translator). Clarendon Aristotle series. Oxford University Press.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75. Association for Computational Linguistics.

Stefanie Brüninghaus and Kevin D. Ashley. 2003. Predicting outcomes of case based legal arguments. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, pages 233–242.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

Michael Collins and Nigel Duffy. 2001. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14*, pages 625–632. MIT Press.

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, 1st edition. Cambridge University Press, New York, NY, USA.

Hal Daumé III and Daniel Marcu. 2004. A tree-position kernel for document compression. In *Proceedings of the Fourth Document Understanding Conference*.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge, UK.

Adam Robert Faulkner. 2014. *Automated Classification of Argument Stance in Student Essays: A Linguistically Motivated Approach with an Application for Supporting Argument Summarization*. Dissertation, City University of New York.

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. Association for Computational Linguistics.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949. Dublin City University and Association for Computational Linguistics.

Simone Filice, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2014. Effective kernelized online learning in language processing tasks. In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416*, pages 347–358.

Simone Filice, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2015. KeLP: A kernel-based learning platform for natural language processing. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 19–24. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Springer.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137. Association for Computational Linguistics.

Thorsten Joachims. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142.

Daisuke Kimura and Hisashi Kashima. 2012. Fast computation of subpath kernel for trees. In *Proceedings of the 29th International Conference on Machine Learning*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

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

Raymond J. Mooney and Razvan C. Bunescu. 2006. Subsequence kernels for relation extraction. In *Advances in Neural Information Processing Systems 18*, pages 171–178. MIT Press.

Alessandro Moschitti. 2006a. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of the 17th European Conference on Machine Learning*, pages 318–329.

Alessandro Moschitti. 2006b. Making tree kernels practical for natural language learning. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: 1st European Conference on Argumentation*. College Publications.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552. Association for Computational Linguistics.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, 14:961–980.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence — An automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450. Association for Computational Linguistics.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *Proceedings of the 25th International FLAIRS Conference*, pages 272–275.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77. Association for Computational Linguistics.

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

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118. Association for Computational Linguistics.

Manfred Stede. 2016. Towards assessing depth of argumentation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3308–3317. The COLING 2016 Organizing Committee.

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

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Alberdingk Tim Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Henning Wachsmuth and Benno Stein. 2017. A universal model for discourse-level argumentation analysis. *Special Section of the ACM Transactions on Internet Technology: Argumentation in Social Media*, 17(3):28:1–28:24.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014a. Modeling review argumentation for robust sentiment analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 553–564. Dublin City University and Association for Computational Linguistics.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014b. A review corpus for argumentation analysis. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics, Part II*, pages 115–127.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 812–817. European Language Resources Association (ELRA).

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