A Universal Model for Discourse-Level Argumentation Analysis

HENNING WACHSMUTH, Bauhaus-Universität Weimar
BENNO STEIN, Bauhaus-Universität Weimar

The argumentative structure of texts is increasingly exploited for analysis tasks, e.g., for stance classification or the assessment of argumentation quality. Most existing approaches, however, model only the local structure of single arguments. This article considers the question of how to capture the global discourse-level structure of a text for argumentation-related analyses. In particular, we propose to model the global structure as a flow of "task-related rhetorical moves", such as discourse functions or aspect-based sentiment. By comparing the flow of a text to a set of common flow patterns, we map the text into the feature space of global structures, thus capturing its discourse-level argumentation. We show how to identify different types of flow patterns and we provide evidence that they generalize well across different domains of texts. In our evaluation for two analysis tasks, the classification of review sentiment and the scoring of essay organization, the features derived from flow patterns prove both effective and more robust than strong baselines. We conclude with a discussion of the universality of modeling flow for discourse-level argumentation analysis.

ACM Reference Format:
Henning Wachsmuth and Benno Stein, 2016. A Universal Model for Discourse-Level Argumentation Analysis. *ACM Trans. Internet Technol.* V, N, Article XXXX (January 2016), 22 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Several types of argumentative texts exist in social media and the web as a whole, ranging from reviews on e-commerce platforms over opinionated blog posts and news editorials to essays and discussions in writing and debate communities. Generally, an argumentative text can be seen as a written form of argumentation. In the article at hand, we focus on monological argumentation, where an author composes arguments to justify a thesis or an opinion on a given topic. Such argumentation puts particular emphasis on argumentative structure [Besnard and Hunter 2008]. Different models have been proposed to capture this structure: some are based on argumentation theory [Toulmin 1958; Walton et al. 2008], considering different types of argument units and relations; others adapt models like the rhetorical structure theory [Mann and Thompson 1988], or they rely on proprietary representations of arguments.

In the last five years, researchers started to model argumentative structure to tackle argumentation-related analysis tasks that aim at predicting a class or a numerical value. Examples are the stance classification of essays and comments [Faulkner 2014; Sobhani et al. 2015] or the quality assessment of essays and their arguments [Ong et al. 2014; Persing and Ng 2015]. So far, however, no argumentation model robustly applies to all text types [Al-Khatib et al. 2016]. Moreover, nearly all approaches restrict their view to the local structure of single arguments, ignoring the impact of the global *discourse-level* argumentation of complete texts. E.g., the sentiment of reviews depends on whether positive aspects precede negatives or vice versa [Wachsmuth et al. 2014b].

Author's addresses: H. Wachsmuth and B. Stein, Faculty of Media, Bauhaus-Universität Weimar, Germany.
 1533-5399/2016/01-ARTXXXX $15.00
DOI: http://dx.doi.org/10.1145/0000000.0000000

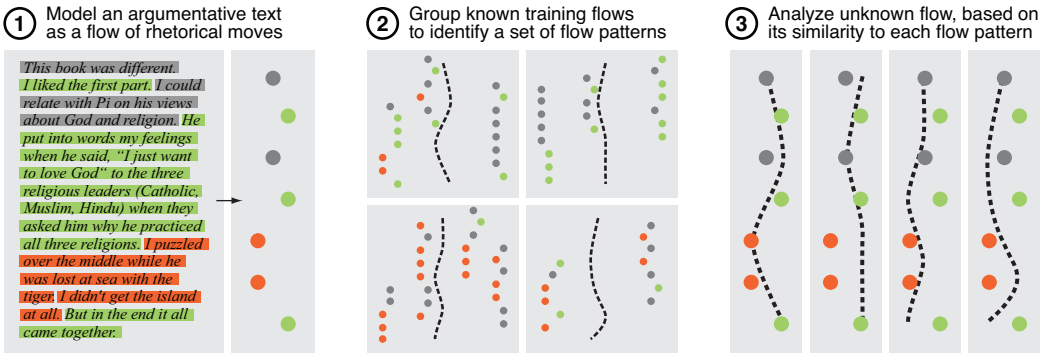

Fig. 1. How to model and analyze the discourse-level argumentation of a text (overall view of our approach).

News editorials have a strategy of when to state pro arguments and when to attack con arguments to convince readers [Kiesel et al. 2015]. An essay should introduce its thesis before it opposes pros and cons, and concludes [Persing et al. 2010]. Similarly, scientific articles follow structural conventions to achieve clarity, e.g., to first present methods and then the results [Teufel et al. 2009]. To account for such global dialectical structure, argumentation must be modeled and analyzed at the discourse level.

In this article, we investigate the question of how to generally capture the structure of an argumentative text for analysis tasks related to the overall argumentation of the text. Figure 1 illustrates the three high-level steps of the approach that we provide. In particular, we propose (1) a flexible argumentation model that represents a text as a flow of task-related rhetorical moves, such as the discourse functions of paragraphs or, as illustrated, the aspect-based sentiment within sentences. The shallow nature of the model supports a reliable pattern identification. (2) Given a collection of training texts, we group the flows of all texts in order to determine a set of common flow patterns. (3) For each unknown text, we then compute the similarity of its flow to all patterns, and we use each similarity as an individual feature for learning to address the analysis task at hand. This way, we map the text into the feature space of global structures, thereby capturing the discourse-level argumentation of the complete text.

In Section 2, we present flow as a model of argumentation. We discuss the kinds of rhetorical moves to represent in a flow and how to represent them for analysis purposes. Section 3 deals with the identification of flow patterns, contrasting "normalization" and "abstraction" as two means to make flows comparable. Patterns are derived from clusters of training flows or from frequency counts then. The last step of our approach is to use the patterns as similarity features for supervised learning (Section 4).

In empirical studies, Section 5 reveals common flow patterns for different text types (reviews and essays), rhetorical moves (discourse functions and sentiment, among others), and granularity levels (sentences and paragraphs). Thereby, we obtain insights into how people argue in practice. Section 6 evaluates major aspects of our approach for two diverse analysis tasks. We find that analyzing flow outperforms strong baselines in the cross-domain sentiment analysis of reviews, and it improves the state of the art in scoring essay organization. These results suggest to discuss flows as a universal model of discourse-level argumentation, which is done in Section 8, after a discussion of related work in Section 7. In conclusion, the contributions of this article are:

(1) We provide a comprehensive and general view of how to model and analyze flows to assess the argumentation of a text at the discourse level.
(2) We report on several statistical patterns of how people argue in reviews and essays.
(3) We offer evidence that capturing the global structure of argumentative texts benefits effectiveness and robustness in two argumentation-related analysis tasks.

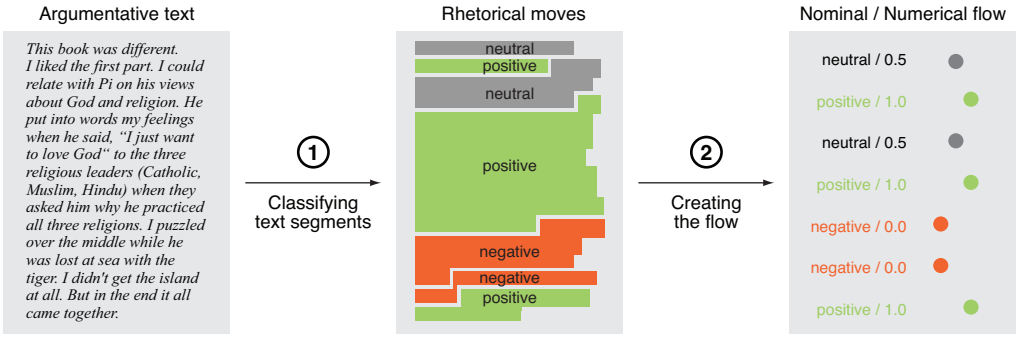

Fig. 2. The main steps of modeling the discourse-level argumentation of a text as a flow of rhetorical moves.

2. MODELING DISCOURSE-LEVEL ARGUMENTATION AS A FLOW

This section presents how to model the flow of argumentative texts for discourse-level argumentation analysis. We discuss the kinds of rhetorical moves to consider in the flow—depending on the text type and analysis task at hand—as well as how to represent these moves. As detailed below, Figure 2 sketches these two steps for a book review, which is mapped to a flow of sentence-level sentiment values.

2.1. Classifying Text Segments as Task-related Rhetorical Moves

The argumentative structure of a text corresponds to a graph whose nodes represent argument units and whose edges model relations between the units. However, many argumentative texts, especially user-generated texts, are written ad-hoc. Reviews such as the one in Figure 2, e.g., often remain with a sequential structure. Also, they miss explicit relations between the reasons they give to justify their ratings. Following Walton and Godden [2006], we see the discourse-level argumentation of texts as a regulated sequence of speech acts. In particular, we propose to identify the *rhetorical moves* of speech acts [Teufel et al. 2009] for modeling the argumentation.

2.1.1. Rhetorical Moves (related to an analysis task). Rhetorical moves represent the communicative functions of segments of an argumentative text, which are linked to the general communicative objective of the type of text [Swales 1990]. Thereby, they support the author's strategy of persuasion or justification. Unlike Swales, we will not consider a fixed set of rhetorical moves for a given text type, but we propose to choose a set in compliance with the analysis task to be addressed: For the organization of essays, Persing et al. [2010] model different discourse functions. For stance classification, it may be more beneficial to explicitly model argument units, e.g., pro claims and con claims along with their premises. Many stance classifiers approximate pros and cons with local sentiment [Faulkner 2014]. This makes particular sense in case of reviews, like the one in Figure 2. There, the ordering of local sentiment adds to the justification of the global sentiment. Also, discourse relations provide insights, such as the review's final contrast. We experiment with various rhetorical moves in Sections 5 and 6.

2.1.2. Text Segments (specific to a text type). Each speech act is represented by a specific span of text. The identification of rhetorical moves therefore requires to first segment the text appropriately. While the right segment size may vary in and between texts, we observe that many text types have a particular level on which the discourse is usually advanced. E.g., many reviews consecutively state opinions on different aspects rather than providing arguments. Their local sentiment is classified on the clause level [Wachsmuth et al. 2014b] or sentence level [Täckström and McDonald 2011]. In contrast, arguments in essays ideally match with paragraphs, although discourse functions are known for both sentences and paragraphs [Persing et al. 2010].

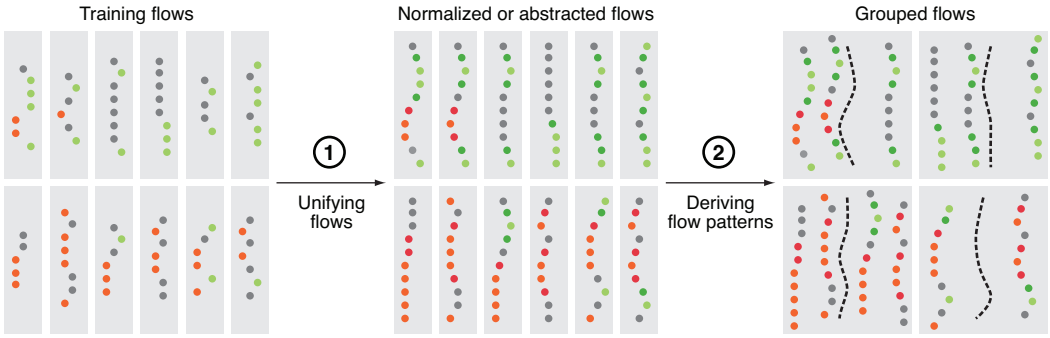

Fig. 3. The main steps when identifying common flow patterns based on a set of known training flows.

We explore all these levels in Sections 5 and 6. Irrespective of the level, the rhetorical move of each segment must be classified. As with most natural language text analyses, such a classification will not be free of noise, which in turn affects the correctness of the resulting model instance. The effect of classification errors is analyzed in Section 6.

2.2. Creating a Flow based on the Rhetorical Moves

Given the rhetorical moves of an argumentative text, we propose a straightforward model of the global discourse-level argumentation of a text for analysis purposes:

2.2.1. Flows. We model global argumentation solely by the sequence of rhetorical moves in a text, which we call the *flow* of the text. In particular, we fully abstract from content. The shallow nature of our model is motivated by the goal of predicting output classes or values (cf. Section 1). Our hypothesis is that similar flows refer to similar outputs, when being based on task-related rhetorical moves. Moreover, we hypothesize that comparable flows are used in the same way across different text domains (in terms of topic or similar). For comparison purposes, we unify flows below.

2.2.2. Nominal versus Numerical Flows. One model parameter deserves special attention, namely, whether rhetorical moves are represented as nominal classes or as numerical values. For instance, sentiments are frequently mapped onto values between 0 and 1 or to some predefined scores [Pang and Lee 2005]. We also do this in Sections 5 and 6. The right side of Figure 2 illustrates a respective flow for local sentiment. The two flow versions shown there can also be denoted in vector form:

Nominal flow. *(Neutral, Positive, Neutral, Positive, Negative, Negative, Positive)*
Numerical flow. *(0.5, 1.0, 0.5, 1.0, 0.0, 0.0, 1.0)*

For other rhetorical moves, a reasonable mapping may not exist, since it reduces their expressiveness to a ratio scale. For example, we model discourse relations as moves in Section 5. We evaluate the impact of using numerical instead of nominal flows in Section 6 where possible. Numerical flows offer advantages for pattern identification.

3. IDENTIFYING COMMON FLOW PATTERNS

As outlined in Section 1, we aim to capture the global structure of an argumentative text by comparing its flow to common flow patterns. This section describes two alternatives of how to unify flows for pattern identification. Then, we present two ways of deriving patterns from groups of training texts. The variants are evaluated in Section 5. Figure 3 sketches one unification and derivation variant for sentiment flows.

3.1. Unifying a Set of Flows

The similarity of two numerical flows can be computed from the numerical difference of their values. For nominal flows, only two cases exist: a pair of values is equal or not.

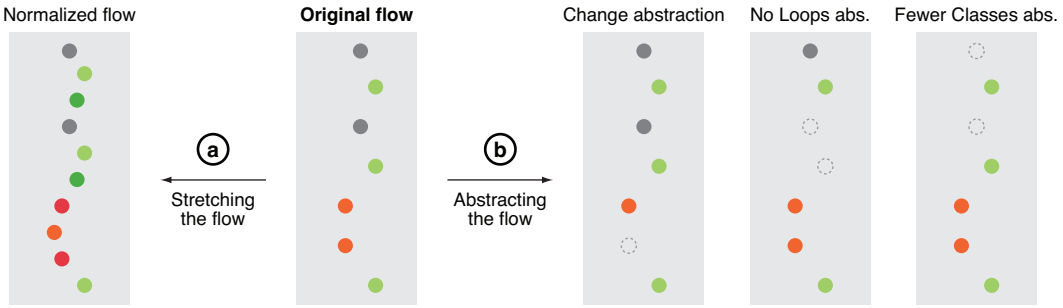

Fig. 4. Comparison of the two proposed unification approaches for a sample flow: (a) Normalization of the flow's dimensionality by stretching it to 10 positions, (b) Application of one or more of three flow abstractions.

Many similarity functions require vectors with uniform dimensions as input, such as the Manhattan or Cosine distance [Cha 2007]. To this end, the flows must be normalized to the same vector space, as e.g. done by Mao and Lebanon [2007]. An adequate normalization can preserve all information, but flows are similar only if they are similar in the same dimensions then. An alternative is to abstract from variances, thus mapping similar flows to the same flow. We introduce both approaches in the following.

3.1.1. Normalized Flows (in terms of dimensionality). With dimensionality, we refer to the number of positions distinguished in a flow. Each position originally represents a particular rhetorical move. For normalization purposes, positions may have to be stretched or squeezed, which brings up two questions: (1) What number of dimensions to use in the normalized form and (2) how to interpolate during stretching and squeezing?

The first question relates to the bias-variance tradeoff in machine learning: Few positions will oversimplify long flows, because they may omit potentially relevant rhetorical moves. Many positions will require too much data to distinguish patterns from noise. A reasonable number should therefore be chosen depending on the expected number of text segments to be represented. The second question is particularly relevant for numerical flows. In case of the sentiment flows from above, a weighted (possibly linear) interpolation seems beneficial, e.g., for comparing flows like *(1.0, 0.5, 0.0)* and *(1.0, 0.0)*. In contrast, nominal classes hardly provide an alternative to simple removal and duplication. Figure 4 depicts a weighted interpolation of the flow from Figure 2, contrasting normalization and abstraction.

3.1.2. Abstracted Flows (in terms of irrelevant variations). While we hypothesize that similar flows occur across domains of argumentative texts (cf. Section 2), we cannot expect the original flows from our model to generalize well: Different domains show different flows due to variations in the text length or the explicitness of rhetorical moves. E.g., a movie review from Rotten Tomatoes is typically over twice as long as a TripAdvisor hotel review or an Amazon product review. At the same time, it contains less and more subtle sentiment. These variations are exemplified in Section 5. Now, we deal with the problem of reducing them. We propose to apply up to three abstractions:

Change. The deletion of repeating rhetorical moves in a flow. The rationale is to reduce differences in explicitness by considering move changes only.
No Loops. The deletion of repeating sequences of two or more rhetorical moves. The rationale is to reduce length differences by merging identical sub-flows.
Fewer Classes. The deletion of specific rhetorical moves of minor relevance. The rationale is to reduce explicitness and length differences caused by these types.[1]

[1] *Fewer Classes* generalizes *2Class* from [Wachsmuth et al. 2015], allowing any number of remaining classes.

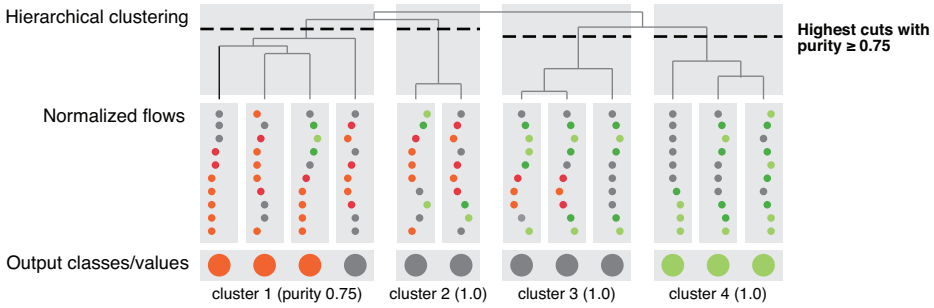

Fig. 5. Cuts in a hierarchical clustering of 12 normalized training flows with known output class/value.

The third abstraction requires expert knowledge about the importance of the move types in the analysis task at hand. Typical candidates are "filler classes" (such as neutral local sentiment) or very frequent and, thus, potentially less discriminative classes. Besides, the three abstractions are not commutative. In Section 5, we analyze combinations of the abstractions. Higher abstraction will benefit generality, but it may decrease the impact of flows in analysis tasks. While the best combinations are not known in advance for the task at hand, they can be learned from a training set.

The general goal of abstraction is to reduce the number of different flows, thus favoring common flows that can be directly used as patterns. Unlike with normalization, similar flows are found even if they are not scaled versions of each other; the avoidance of a vector space enables more powerful similarity functions, as detailed in Section 4. Moreover, abstraction lowers the risk of misclassifying rhetorical moves. E.g., if one negative sentiment in the original flow from Figure 4 is classified as positive, *Change* eliminates the effect. If classified as neutral, *Change* and *Fewer Classes* do so together.

3.2. Deriving Flow Patterns from a Set of Unified Flows

Having unified the set of all training flows through normalization or abstraction, common flow patterns can be derived. Again, we present two alternative ways to approach this step: (1) Simple frequency counts and (2) a clustering of the flows.

3.2.1. Abstracted Flow Patterns (derived from frequency counts). In general, the number of different flows that remains after unification is not predictable. However, flow abstraction as defined above often yields several frequent flows. In this case, a way to obtain flow patterns is to simply take the most frequent flows, i.e., those that represent a certain fraction of all training flows.

3.2.2. Normalized Flow Patterns (derived from flow clusters). Even with significant abstraction, many training sets will contain only very few flows that occur multiple times in exactly the same form, because of the diversity of natural language. Still, many flows will be similar in terms of some adequate similarity function (for normalized flows, we use the Manhattan distance below). In such situations, we propose to determine a set of flow patterns such that each pattern represents a distinct set of similar training flows. Ideally, flows that refer to the same flow pattern should be as similar as possible and others as dissimilar as possible. This suggests to partition the training flows with clustering and to then derive flow patterns from the resulting clusters.

Our hypothesis is that similar flows entail the same output class or value within the task to be addressed (cf. Section 2). Since we know the correct output for each training flow, we propose a supervised variant of clustering that exploits this knowledge to ensure that all clusters are of high *purity*. Here, purity denotes the proportion of flows that have the majority output within a cluster [Manning et al. 2008]. At the same time, we aim for few clusters in order to prefer generality over specificity of the flow

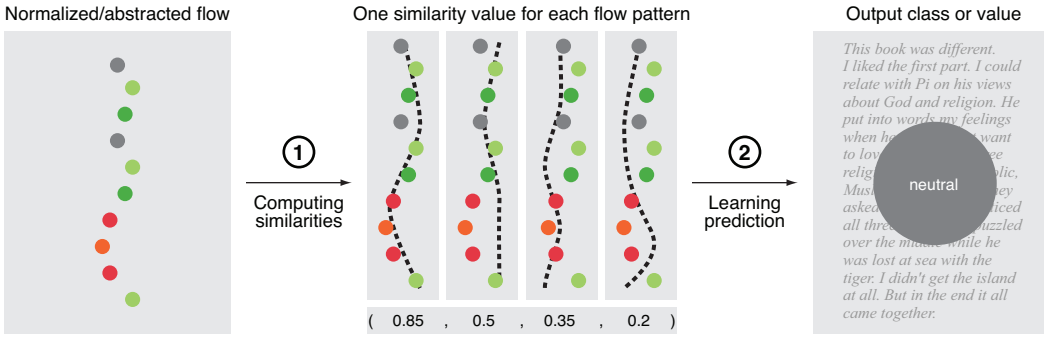

Fig. 6. The main steps of analyzing an argumentative text based on its similarity to a set of flow patterns.

patterns. We account for both aims by using hierarchical clustering. This way, we can sensibly control the number of distinct clusters through cuts in the clustering dendogram (the binary tree of the associated hierarchy). To minimize the number, we search all cuts closest to the root of the tree that create clusters whose purity lies above some threshold. Figure 5 exemplifies this approach for a sample threshold of 0.75.

The centroid of each cluster becomes a flow pattern.[2] For numerical flows, it can be computed as the average of the values of all flows in the cluster, as demonstrated in Section 5. For nominal flows, we use the majority class at each position of the flows.

4. ANALYZING DISCOURSE-LEVEL ARGUMENTATION BASED ON FLOW PATTERNS

To analyze an argumentative text at the discourse level, we compare its flow to the identified flow patterns. In this section, we discuss how to compute similarities for this purpose and how to predict output classes or values from the similarities within an analysis task. Figure 6 illustrates these steps for classifying global sentiment.

4.1. Computing Similarities between Flows and Flow Patterns

In case the flow of a text does not already occur in the set of training flows, its output output class or value is unknown. The flow is hence transformed into the unified form of the given flow patterns (cf. Section 3) and its similarity to each flow pattern is computed. We consider two alternative approaches in this regard.

4.1.1. Vector Space Distance (for normalized flows). If the training flows were mapped into a vector space through normalization, a similarity function (e.g., the Manhattan distance) has already been chosen for the subsequent pattern derivation. The same function should also be used to compute the similarity flows and flow patterns in order not to mix up different similarity concepts. The center part of Figure 6 shows exemplary similarity values that result in this case.

4.1.2. Edit Distance (for abstracted flows). If unification was achieved via abstraction, an adequate similarity function is still to be chosen. As argued in Section 3, a function is preferred that detects similar flows even if they are not similar at the same positions. Thus, we compare flows and flow patterns in terms of their normalized minimum edit distance. This requires the specification of costs for all edit operations. We consider the substitution, insertion, and deletion of a single rhetorical move. Given the moves r and r' at some position of a pair of flow and pattern, we define the costs as follows:

$$d(r,r') = \begin{cases} \Delta(r,r') & \text{if } r' \text{ substitutes } r. \\ \alpha + (1-\alpha) \cdot \Delta(r,r') & \text{if } r' \text{ is inserted or deleted after } r. \end{cases}$$

[2]For noise reduction, flow patterns can be derived only from clusters of some minimum size. Also, rare flows can already be discarded before clustering (e.g., flows that occur only once in the training set).

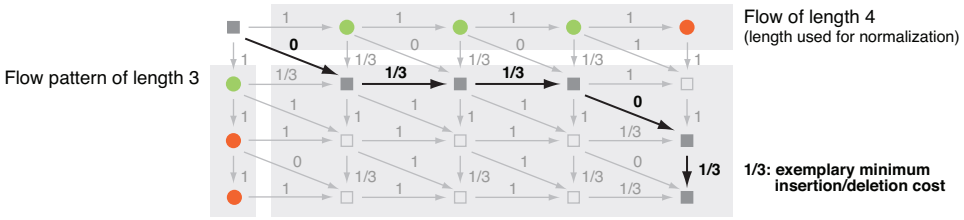

Fig. 7. Illustration of the normalized edit distance of a flow and a flow pattern: $(2 \cdot 0 + 3 \cdot 1/3) \,/\, 4 = 1/4$.

Here, $\Delta(r, r')$ is the numerical difference in case r and r' are represented as numerical values. For nominal classes, $\Delta(r, r') = 0$ if $r = r'$, and $\Delta(r, r') = 1$ otherwise. The idea is to increase the cost the more r and r' differ, but to induce a minimum cost value $\alpha \in [0, 1]$ for insertions and deletions—to capture differences that remain after abstraction. We set α to 1/3 in Section 6, which worked best among 1, 1/2, 1/3, and 1/4 in experiments of Wachsmuth et al. [2015]. Based on d, the edit distance can be incrementally computed with sequence alignment, as Persing et al. [2010] do for discourse functions. Figure 7 illustrates this as a shortest-path search. The minimum distance is normalized to $[0, 1]$ by the maximum length of the flow and flow pattern.

4.2. Learning the Prediction of Output Classes and Values from the Similarities

We consider the flow patterns as a feature type for supervised machine learning such that each similarity between a flow and a flow pattern denotes a single feature. The output class or value of a text can then be predicted from a feature vector of similarity values. Thereby, we compare the global structure of a text (represented by the flow) to known global structures (the flow patterns), hence mapping the text into the feature space of global structures. Figure 6 exemplifies this approach for the prediction of neutral global sentiment from the similarity of a sentiment flow to four flow patterns.

To learn the prediction, the same training set can be used that has been processed when identifying the flow patterns. In addition, the similarity features can be combined with any other features. In Section 6, we evaluate our novel feature type in two analysis tasks. More details about its implementation are found in [Wachsmuth 2015].

5. EMPIRICAL STUDY OF FLOW PATTERNS

This section reports on the identification of flow patterns in several empirical analyses with four types of rhetorical moves on datasets with two types of argumentative texts. Section 5.1 seeks to provide evidence that our model reveals novel insights about how people argue in texts. Section 5.2 demonstrates that flow patterns generalize well across domains. An overview of all datasets used in this paper is given in Table I.[3]

5.1. Analysis of Common Flows in Reviews and Essays

Our hypothesis is that the global structure of an argumentative text impacts discourse-level properties of the text. To investigate this, we computed the most common flows for different types of argumentative texts and rhetorical moves as well as their cooccurrence with certain text classes, given the following setting:

5.1.1. Argumentative Texts. One set of argumentative texts is the hotel review corpus from [Wachsmuth et al. 2014b] in Table I. Each clause-level segment in the reviews is classified to have a negative, neutral, or positive sentiment. The reviews are balanced regarding their overall ratings between 1 and 5; here, we map the ratings to negative (1 and 2), neutral (3), and positive (4 and 5) global sentiment.

[3]All presented results have been produced with Java implementations found at http://www.arguana.com. In case you need further instructions to reproduce them, do not hesitate to contact the authors of this article.

Table I. Facts about the datasets on which we evaluated the two specified analysis tasks in the specified sections: Source, text type, number of texts, segment level of the analyzed flows, and average number of segments per text.

Task	Sections	Source of dataset	Text type	Texts	Seg' level	Seg's
Sentiment analysis	5.1, 5.2, 6.1	[Wachsmuth et al. 2014b]	Hotel reviews	2100	Clauses	14.8
	5.2, 6.1	[Täckström and McDonald 2011]	Product reviews	294	Sentences	11.5
	5.2	[Pang and Lee 2005]	Movie reviews	5098	Sentences	36.1
	5.2, 6.1	[Mao and Lebanon 2007]	Movie reviews	450	Sentences	28.9
Organization scoring	6.2	[Persing et al. 2010]	Student essays	1003	Sentences	31.4
	5.1, 6.2	[Persing et al. 2010]	Student essays	1003	Paragraphs	7.5

Second, we use the 1003 student essays from the International Corpus of Learner English [Granger et al. 2009], annotated by Persing et al. [2010] with respect to an argumentation-related quality dimension. In particular, each essay is assigned a half-point score in the range $[1.0, 4.0]$ that represents the quality of its organization. According to the authors, a high score is given to essays that introduce their topic, argue for a stance on the topic, and conclude. The mean score is 3.05, the standard deviation 0.59.

5.1.2. Rhetorical Moves. In case of the hotel reviews, we model rhetorical moves for the given clause-level segments. Essay argumentation, however, proceeds paragraph-wise, which is why we look for rhetorical moves on the paragraph level there (the impact of the chosen level is evaluated in Section 6). In total, we assess four types of moves:

(1) **Local Sentiment**. The *negative*, *neutral*, or *positive* sentiment of a text segment.
(2) **Discourse Relations**. The discourse relations between consecutive segments. We use a subset of the relations from [Mann and Thompson 1988]: *background*, *cause*, *circumstance*, *concession*, *condition*, *contrast*, *motivation*, *purpose*, and *summary*.
(3) **Discourse Functions**. The four paragraph-level functions considered by Persing et al. [2010], i.e., *introduction*, *body* (own argument), *rebuttal*, and *conclusion*.
(4) **Argument Roles**. Whether a text segment is a real *argument* (with premises and a claim or major claim), whether it serves as *premise* or a (major) *claim*, or *none*.

5.1.3. Preprocessing. Local sentiment is already given in the reviews. Due to lack of ground-truth data for the other rhetorical moves, we relied on automatic classification. Concretely, to obtain paragraph sentiment, we first apply the state-of-the-art sentence sentiment classifier of Socher et al. [2013] to all essays. A paragraph is then classified as positive if it contains a positive but no negative sentence, and vice versa. Else, it is seen as neutral. For discourse relations, we used the rule-based algorithm from [Wachsmuth et al. 2014a] that looks for high-precision indicator words, such as "but" or "because". Discourse functions were found with the heuristic algorithm of Persing et al. [2010]. Finally, we reimplemented a mining approach of Stab and Gurevych [2014] to identify major claims, claims, premises, and the none class on the sentence level. Similar to the authors, we achieve a weighted F_1-score of 74.5 on their data. Argument roles are derived as follows: All paragraphs with a claim and a premise were classified as arguments, all with neither as *none*. The remaining are claims or premises.

5.1.4. Experiments. Given all rhetorical moves, we created the different flows for all reviews and essays. As many original flows do not generalize well (recall Section 3.1), we applied the defined *Change* abstraction to each flow. Afterwards, we computed the frequency of each resulting *change flow pattern* and its distribution over all text classes. Due to the limited effectiveness of automatic classification, the patterns we found have to be interpreted with caution, because they may not fully represent actual argumentation. Still, clear correlations with text classes can be insightful.

5.1.5. Results. Table II lists the most frequent flow patterns in the hotel reviews based on local sentiment and discourse relations, respectively. While the first three sentiment patterns are trivial (i.e., without any changes), the others reveal the impact of a re-

Table II. The 12 most frequent flow patterns in the review corpus from [Wachsmuth et al. 2014b] when considering changes of clause-level sentiment and discourse relations, respectively. The relative frequency of each pattern and its distribution over the three global sentiments are given in percent (the majority global sentiment is marked bold).

Type	#	Clause-level Change Flow Pattern	Rel. Freq.	Positive	Neutral	Negative
Local sentiment	1	(positive)	7.7	**87.7**	7.4	4.9
	2	(neutral)	5.2	**62.7**	20.0	17.3
	3	(negative)	3.5	1.4	9.6	**89.0**
	4	(positive, neutral, positive)	3.0	**93.5**	6.5	0.0
	5	(neutral, positive)	2.7	**91.2**	7.0	1.8
	6	(positive, negative, positive)	2.1	**72.7**	11.4	15.9
	7	(neutral, positive, neutral, positive)	1.9	**94.9**	5.1	0.0
	8	(negative, neutral, negative)	1.7	0.0	2.8	**97.2**
	9	(positive, negative)	1.7	19.4	33.3	**47.2**
	10	(neutral, positive, negative, positive)	1.5	**64.5**	32.3	3.2
	11	(negative, positive, negative)	1.5	0.0	12.9	**87.1**
	12	(neutral, negative)	1.1	0.0	0.0	**100.0**
Discourse Relations	1	(contrast)	25.2	30.8	25.5	**43.8**
	2	(circumstance)	3.7	20.8	5.2	**74.0**
	3	(concession)	2.6	**40.0**	30.9	29.1
	4	(motivation)	2.4	**60.0**	18.0	22.0
	5	(contrast, circumstance)	1.9	20.0	22.5	**57.5**
	6	(circumstance, contrast)	1.5	28.1	12.5	**59.4**
	7	(contrast, concession)	1.4	33.3	13.3	**53.3**
	8	(concession, contrast)	1.3	32.1	28.6	**39.3**
	9	(contrast, motivation)	1.3	32.1	**35.7**	32.1
	10	(cause)	1.3	**44.4**	11.1	**44.4**
	11	(motivation, contrast)	1.0	27.3	22.7	**50.0**
	12	(contrast, circumstance, contrast)	1.0	15.0	15.0	**70.0**
Average				**40.0**	20.0	**40.0**

view's sentiment flow on its global sentiment. E.g., the flow *(positive, negative, positive)* (line 6, upper part) is mostly positive while *(negative, positive, negative)* in line 11 is mostly negative. The 1.7% of all reviews, which begin positive and turn negative, typically result in negative (47.2%) or neutral (33.3%) global sentiment. Discourse relation flows seem to play a limited role. E.g, the circumstance relation indicates a rather negative review—irrespective of its position in a flow (lines 2, 5, 6, and 12 in the lower part). Some impact can be seen, though. For example, *motivation* (indicated by second person voice) in isolation is positive in 60% of the cases, whereas *(contrast, motivation)* rather occurs in neutral reviews and *(motivation, contrast)* in negative reviews.

Turning to the essays, Table III shows the most frequent change flow patterns for three rhetorical move types. The average distribution at the bottom helps to interpret their score distribution. In case of local sentiment, the shortest pattern found, *(neu.)*, often results in score 1.0 (24.4%), whereas the longest has the highest correlation with 4.0 (17.4%) among all listed sentiment flow patterns. These results suggest that organization scores correlate with the length of essays. Hardly any patterns contains the move *positive*, which may be due to the fact that the employed sentiment classifier was trained out-of-domain (on movie reviews). For discourse functions, the listed patterns provide strong evidence that people tend to (or are taught to) follow a particular argumentative structure when writing an essay. Most dominantly, *(intro., body, conclusion)* represents every fifth essay (19.6%). This and some other patterns support a high organization score. Especially, the pattern in line 12 (middle part) results in 3.5 or 4.0 in about 60% of the cases, suggesting that it is clever to first rebut con arguments before presenting pro arguments. Similar observations can be made for argument roles. Among others, the most frequent pattern *(claim, argument)* reflects a good organization, whereas the scores of *(claim, argument, claim, argument, claim)* indicate that multiple switches between claim and argument paragraphs are not optimal.

We conclude that several common flow patterns exist in the given reviews and essays. Also, some types of rhetorical moves have more impact on discourse-level properties than others. Their actual use for analysis tasks is evaluated in Section 6.

Table III. The 12 most frequent flow patterns in the essay corpus of Persing et al. [2010] when considering changes of paragraph-level sentiment, discourse functions, and argument roles, respectively. All values in percent.

Type	#	Paragraph-level Change Flow Pattern	Rel. F.	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Local sentiment	1	(neu., negative, neu.)	8.9	0.0	0.0	0.0	4.5	38.2	**43.8**	13.5
	2	(neu.)	8.6	**24.4**	8.1	7.0	3.5	**24.4**	**24.4**	8.1
	3	(neu., negative, neu., negative)	6.5	0.0	0.0	3.1	12.3	**38.5**	36.9	9.2
	4	(negative, neu., negative)	6.2	0.0	0.0	1.6	11.3	**45.2**	33.9	8.1
	5	(neu., negative)	6.1	0.0	0.0	6.6	18.0	**32.8**	**32.8**	9.8
	6	(negative, neu.)	5.5	1.8	1.8	0.0	7.3	**41.8**	34.5	12.7
	7	(negative, neu., negative, neu.)	5.3	0.0	0.0	1.9	15.1	**43.4**	34.0	5.7
	8	(neu., negative, neu., negative, neu.)	3.9	0.0	0.0	0.0	10.3	**38.5**	35.9	15.4
	9	(negative, neu., negative, neu., negative)	3.8	0.0	0.0	0.0	10.5	**52.6**	23.7	13.2
	10	(neu., negative, neu., negative, neu., negative)	2.3	0.0	0.0	4.3	26.1	**34.8**	17.4	17.4
	11	(neu., positive, neu.)	1.6	0.0	0.0	0.0	18.8	37.5	31.3	12.5
	12	(negative)	1.4	0.0	0.0	7.1	7.1	64.3	7.1	14.3
Discourse functions	1	(intro., body, conclusion)	19.6	0.5	0.0	1.0	8.6	**42.6**	37.1	10.2
	2	(intro., body)	6.7	1.5	1.5	1.5	16.4	**34.3**	**34.3**	10.4
	3	(intro., body, rebuttal, body, conclusion)	4.3	0.0	2.3	4.7	34.9	**44.2**	14.0	0.0
	4	(intro., body, intro.)	4.2	0.0	0.0	0.0	7.1	**42.9**	38.1	11.9
	5	(intro., body, intro., body, conclusion)	3.5	0.0	0.0	2.9	20.0	**37.1**	31.4	8.6
	6	(intro., body, intro., conclusion)	3.4	0.0	0.0	0.0	14.7	**41.2**	26.5	17.6
	7	(intro.)	3.3	**60.6**	21.2	6.1	3.0	3.0	6.1	0.0
	8	(conclusion, body, conclusion)	3.1	0.0	0.0	0.0	3.2	35.5	**54.8**	6.5
	9	(intro., body, conclusion, body, conclusion)	2.7	0.0	0.0	0.0	11.1	29.6	**48.1**	11.1
	10	(intro., conclusion)	1.6	0.0	0.0	12.5	18.8	25.0	**31.3**	12.5
	11	(intro., body, conclusion, body)	1.2	0.0	0.0	0.0	16.7	66.7	16.7	0.0
	12	(intro., rebuttal, body, conclusion)	1.2	0.0	0.0	8.3	0.0	33.3	33.3	25.0
Argument roles	1	(claim, argument)	10.2	0.0	2.0	1.0	9.8	31.4	**38.2**	17.6
	2	(none, argument)	9.2	1.1	2.2	2.2	7.6	30.4	**45.7**	10.9
	3	(claim, argument, claim)	5.2	0.0	0.0	0.0	9.6	**46.2**	38.5	5.8
	4	(none, argument, claim)	4.9	0.0	0.0	0.0	10.2	**42.9**	34.7	12.2
	5	(claim, argument, claim, argument, claim)	4.1	0.0	0.0	4.9	22.0	**48.8**	24.4	0.0
	6	(none, argument, premise, argument)	3.3	0.0	0.0	0.0	12.1	**36.4**	33.3	18.2
	7	(claim, argument, claim, argument)	3.2	0.0	0.0	0.0	9.4	**53.1**	34.4	3.1
	8	(none, argument, claim, argument)	3.0	0.0	0.0	3.3	10.0	**63.3**	23.3	0.0
	9	(claim, argument, premise, argument)	3.0	0.0	0.0	0.0	3.3	40.0	**43.3**	13.3
	10	(premise)	2.1	**90.5**	9.5	0.0	0.0	0.0	0.0	0.0
	11	(none, argument, claim, argument, claim)	2.1	0.0	0.0	0.0	9.5	61.9	23.8	4.8
	12	(claim, argument, premise, argument, claim)	2.0	0.0	0.0	5.0	0.0	50.0	30.0	15.0
Average				2.4	1.4	3.4	14.6	**41.6**	28.8	7.9

5.2. Analysis of the Generality of Flow Patterns across Domains

We have hypothesized above that similar flows can be found across domains of texts—provided an adequate abstraction. In [Wachsmuth et al. 2015], we systematically analyze the generality of combinations of the abstractions outlined in Section 3.1. Since we also evaluate the domain robustness of using flows for analysis tasks (in Section 6), we study our hypothesis only exemplarily here for sentiment analysis. In particular, we bring together two experiments from [Wachsmuth et al. 2014a] and [Wachsmuth et al. 2015], allowing to compare normalized and abstracted flows:

5.2.1. Argumentative Texts. For the case of normalization, we consider all 900 training hotel reviews from [Wachsmuth et al. 2014b] as one domain (cf. Table I). The other domain is defined by movie reviews from [Pang and Lee 2005], where we use the 1302 reviews of author c and the 1027 reviews of author d as two separate datasets. We map their sentiment scale $[0, 2]$ to negative (0), neutral (1), and positive (2) global sentiment, and we assume each sentence as one text segment. For the abstraction case, we rely on ground-truth local sentiment. Besides the hotel reviews, we consider the 450 movie reviews from [Mao and Lebanon 2007], a subset of the dataset from [Pang and Lee 2005]. In these, all sentences are classified as negative, neutral, or positive. The annotated reviews lack global sentiment. We could recover it from the original dataset (178 positive, 139 neutral, and 133 negative reviews). We used the 201 reviews of one of author here. Finally, we added the 294 product reviews from [Täckström and McDonald

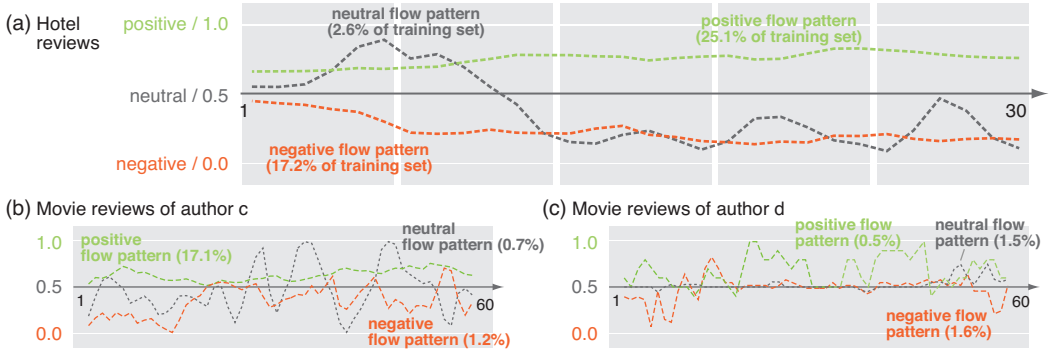

Fig. 8. The three most common normalized sentiment flow patterns found in (a) the processed hotel reviews and (b–c) the movie reviews of author c and d. All patterns are labeled with the associated global sentiment.

2011] where every sentence is classified as positive, negative, neutral, mixed, or irrelevant. We mapped the last three to *neutral*. The reviews are roughly balanced with respect to global sentiment and five product categories (books, DVD, electronics, music, and videogames) from which we used the 175 book, DVD, and electronics reviews. The remaining parts of the corpora are left for testing our approach in Section 6. Besides the length of reviews, the given domains largely differ in the distribution of local sentiment: 38% of the ground-truth segments in hotel reviews are positive, 41.7% are negative. In contrast, only 24.1% (34.4%) of the product sentences are positive (negative) and only 17.6% (21.2%) of the movie sentences.

5.2.2. Rhetorical Moves. We model negative, neutral, and positive local sentiment as task-related rhetorical moves for the prediction of global sentiment.

5.2.3. Preprocessing. For normalized sentiment flows, we trained linear support vector machines that classify local sentiment of text segments [Wachsmuth et al. 2014a]. Based on their output, flows are created and normalized to 30 (hotel reviews) and 60 (movie reviews) dimensions, respectively. While the normalization length itself is an optimization parameter that is not studied here, the chosen lengths capture about 90% of the original flows without loss according to the respective training sets. To obtain flow patterns as defined in Section 3.2, we developed an agglomerative hierarchical clusterer that employs the Manhattan distance between flows and that operationalizes the merging with group-average link. The dendrogram cuts are made for a purity threshold of 0.8; the centroid of each cluster with at least three flows became a flow pattern.[4] For the abstracted sentiment flows, no preprocessing was performed, but the flows created from the given ground-truth local sentiment were used.

5.2.4. Experiments. In the supervised normalization setting, we computed the most common flow pattern for each global sentiment in the respective dataset and compared the different patterns. In the unsupervised abstraction setting, we first determined the most frequent abstracted flow in every domain for combinations of the abstractions from Section 3.1. To assess their generality, we analyzed the frequency and global sentiment distribution of each flow in all three domains.[5]

5.2.5. Results. Figure 8 shows the most common normalized sentiment flow patterns in the hotel and the movie domain. The three hotel patterns represent between 2.6%

[4]The chosen parameter values of the clusterer worked best in initial experiments. In principle, these values are also subject to optimization, as discussed in [Wachsmuth 2015].
[5]For comparability, we balanced the training flows before according to their global sentiment distribution. E.g., if 40% are positive, 30% neutral, and 30% negative, then the positive flows are weighted by 0.75.

Table IV. The most frequent flow pattern in each evaluated training domain (marked bold) for five combinations of abstractions. For each test domain, the relative frequency and global sentiment distribution are given in percent.

Abstractions	Domain	Rank	Clause-level Sentiment Flow Pattern	Rel. F.	Pos.	Neu.	Neg.
Change, Fewer classes, No loops	**Product**	**1st**	(negative, negative)	13.6	0.0	25.0	75.0
	Hotel	9th		4.2	0.0	8.9	91.1
	Movie	4th		6.7	0.0	0.0	100.0
	Product	8th	(positive, negative, positive)	3.4	34.1	65.9	0.0
	Hotel	**1st**		10.5	45.1	49.6	5.3
	Movie	14th		2.3	0.0	73.3	26.7
	Product	10th	(negative, negative, positive, negative)	3.4	0.0	33.3	66.7
	Hotel	15th		2.2	0.0	16.7	83.3
	Movie	**1st**		8.6	0.0	57.9	42.1
Fewer classes, No loops	**Product**	**1st**	(negative, negative)	15.3	7.6	29.6	62.8
	Hotel	5th		4.4	0.0	8.5	91.5
	Movie	**1st**		6.7	0.0	0.0	100.0
	Product	3rd	(positive, positive)	12.8	86.8	8.8	4.4
	Hotel	**1st**		7.6	87.8	12.2	0.0
	Movie	10th		2.1	61.1	38.9	0.0
Change, No loops	**Product**	**1st**	(positive, neutral, positive)	10.5	94.6	0.0	5.4
	Hotel	6th		3.3	88.9	11.1	0.0
	Movie	23rd		1.3	100.0	0.0	0.0
	Product	22nd	(positive)	1.1	50.9	49.1	0.0
	Hotel	**1st**		6.6	83.1	14.1	2.8
	Movie	–		–	–	–	–
	Product	2nd	(neutral, negative)	9.1	6.5	24.9	68.6
	Hotel	5th		3.5	0.0	10.5	89.5
	Movie	**1st**		7.6	8.6	0.0	91.4
No loops, Fewer classes	**Product**	**1st**	(negative, negative)	10.2	11.5	33.2	55.3
	Hotel	5th		2.0	0.0	–	100.0
	Movie	–		–	–	–	–
	Product	3rd	(positive, positive, positive)	5.3	100.0	0.0	0.0
	Hotel	**1st**		5.6	90.0	10.0	0.0
	Movie	–		–	–	–	–
	Product	22nd	(6x negative)	1.1	0.0	0.0	100.0
	Hotel	65th		0.3	0.0	0.0	100.0
	Movie	**1st**		2.7	0.0	0.0	100.0
Fewer classes	**Product**	**1st**	(negative, negative, negative, negative)	5.1	0.0	22.2	77.8
	Hotel	13th		0.9	0.0	20.0	80.0
	Movie	–		–	–	–	–
	Product	12th	(6x positive)	4.6	75.6	12.2	12.2
	Hotel	**1st**		2.3	84.0	16.0	0.0
	Movie	–		–	–	–	–
	Product	–	(16x negative)	–	–	–	–
	Hotel	–		–	–	–	–
	Movie	**1st**		1.5	0.0	0.0	100.0

and 25.1% of the 900 training reviews. The positive and the negative pattern both indicate rather simple one-sided argumentations, while the neutral pattern captures the intuitive argumentation of starting with positive but soon arriving at negative aspects. In total, the clustering produced 38 hotel patterns. In contrast, 75 patterns resulted from the movie reviews of author c and 41 from those of author d. The depicted movie patterns show less clear sentiment but more changes than the hotel patterns. While the global behavior of the hotel patterns and those of author c are partly comparable, two patterns of author d contain only little sentiment at all, especially in the middle. Although exemplary only, these observations suggest that normalized sentiment flow patterns do not sufficiently generalize and that abstraction is needed.

Table IV shows the most frequent abstracted flow from each review domain for five selected combinations of abstractions. Frequent flows naturally tend to be simple, and this tendency is reinforced by the abstractions. While less insightful, the respective patterns may still capture decisive discourse-level differences between flows. Besides, we also see more complex patterns: In case of the first combination, e.g., the top movie

Table V. Rhetorical moves and flow types of the flow patterns evaluated in the specified sections. For each type of rhetorical move, both normalization and abstraction are used, based on clustering and frequencies, respectively.

Rhetorical Moves	Flow type	Sections	Unification	Identification	Similarity function
Argument roles	Nominal	6.2	Normalization Abstraction	Cluster centroids Frequency $\geq 1\%$	Manhattan distance Edit distance
Local sentiment	Numeric	6.1, 6.2	Normalization Abstraction	Cluster centroids Frequency $\geq 1\%$	Manhattan distance Edit distance
Discourse functions	Both	6.2	Normalization Abstraction	Cluster centroids Frequency $\geq 1\%$	Manhattan distance Edit distance

flow *(negative, negative, positive, negative)* also constitutes the 10th and 15th most frequent flow in the product and hotel domain, respectively. For *Fewer classes, No Loops*, two domains even share the first flow. At least the flows of the upper three combinations are common across domains with a frequency of over 2% in most cases. While only few flows clearly predict one global sentiment, almost all behave similar across domains. An exception is *(positive, negative, positive)* in *Change, Fewer classes, No Loops*, which is rather positive in product reviews but negative in movie reviews.

To summarize, our analyses yield flow abstractions that prove to be general across review domains, although not being a perfect model of review argumentation. The discussed combinations were selected since we found in [Wachsmuth et al. 2015] that they abstract neither too little, such as the original flows, nor too much, such as *Fewer classes, Change, No Loops*, which creates at most seven flows per domain. In Section 6, we evaluate the selected combinations for analyzing discourse-level argumentation.

6. EXPERIMENTS WITH DISCOURSE-LEVEL ARGUMENTATION ANALYSIS

This section presents the evaluation of our approach in two argumentation-related analysis tasks. We demonstrate the effectiveness and domain robustness of modeling discourse-level argumentation with flows and we examine major flow parameters. An overview of all evaluated features based on flow patterns is given in Table V.[6]

6.1. Domain-Robust Sentiment Analysis of Reviews

Given the generality results from Section 5.2, we hypothesize that our features based on flow patterns allow for domain-robust argumentation analyses. We tested this hypothesis for the global sentiment analysis of reviews in the following experiments.

6.1.1. Argumentative Texts. As in Section 5.2, we resort to the product, hotel, and movie reviews in Table I from [Täckström and McDonald 2011; Wachsmuth et al. 2014b; Mao and Lebanon 2007]. While all three datasets contain ground-truth sentiment on the clause- or sentence-level, below we analyze the effect of using self-classified rhetorical moves. Precisely, we classify local sentiment with the algorithm of Socher et al. [2013].

6.1.2. Approach and Baselines. All approaches evaluated here represent features for supervised learning with normalized values in $[0, 1]$. We consider both pattern variants:

- a_1 **Normalized Sentiment Flows.** The Manhattan distance of the normalized flow of a review to all normalized flow patterns, identified through clustering as introduced in Section 3.2 and concretized in Section 5.2.
- a_2 **Abstracted Sentiment Flows.** The minimum edit distance of the abstracted flow of a review to the most frequent abstracted flow patterns, namely, those that represent at least 1% of the respective training set. The distances are computed for all five combinations of the abstractions from Section 3.1 that are shown in Table IV.

We compare the accuracy of our approach to previous approaches for the given corpora. To assess its domain robustness, we also evaluate two additional baselines:

[6]Again, results can be reproduced with the code at http://www.arguana.com and http://tinyurl.com/toit2016.

Table VI. Accuracy in 3-class global sentiment analysis using our approach and the baselines for each combination of training and test review domain, based either on self-classified or on ground-truth local sentiment.

Training domain		Feature types	Accuracy on self-classified			Accur. on ground-truth		
			Product	Hotel	Movie	Product	Hotel	Movie
Product	a_1	Normalized sentiment flows	46.8	57.5	47.8	**77.2**	**70.1**	55.8
	a_2	Abstracted sentiment flows	50.5	51.3	42.4	70.9	58.1	55.1
	a	**Approach** (a_1+a_2)	50.9	58.2	**51.1**	73.7	64.5	**58.0**
	b_1	Bag-of-words	49.0	45.9	32.4	49.0	45.9	32.4
	b_2	Sentiment distribution	51.7	50.4	39.3	74.4	69.1	55.1
	b	**Baselines** (b_1+b_2)	**56.5**	53.0	34.7	67.3	68.8	42.7
	a+b_1	Approach + Bag-of-words	54.1	**60.7**	50.0	72.9	65.4	57.6
	a+b_2	Approach + Sentiment distribution	50.8	59.7	50.2	73.6	64.3	56.4
	a+b	**Approach + Baselines**	54.2	60.0	48.7	74.6	64.3	56.4
Hotel	a_1	Normalized sentiment flows	50.7	74.2	51.1	60.2	81.5	59.8
	a_2	Abstracted sentiment flows	53.4	69.0	**54.7**	67.0	79.3	64.0
	a	**Approach** (a_1+a_2)	53.8	75.5	53.6	65.6	81.6	66.7
	b_1	Bag-of-words	37.8	79.6	39.8	37.8	79.6	39.8
	b_2	Sentiment distribution	51.4	64.2	51.1	59.9	76.6	56.0
	b	**Baselines** (b_1+b_2)	44.9	**81.9**	41.3	50.7	**87.0**	42.4
	a+b_1	Approach + Bag-of-words	54.8	79.2	52.2	**68.0**	85.3	68.7
	a+b_2	Approach + Sentiment distribution	**57.1**	75.6	51.8	65.6	81.5	**69.3**
	a+b	**Approach + Baselines**	56.4	79.0	53.3	67.0	85.4	69.1
Movie	a_1	Normalized sentiment flows	42.2	39.5	67.2	**69.4**	70.5	74.4
	a_2	Abstracted sentiment flows	44.6	49.7	60.9	60.5	62.1	73.4
	a	**Approach** (a_1+a_2)	47.6	51.9	65.2	62.2	66.7	75.6
	b_1	Bag-of-words	35.0	41.2	64.8	35.0	41.2	64.8
	b_2	Sentiment distribution	43.2	44.2	59.0	67.4	**71.4**	72.2
	b	**Baselines** (b_1+b_2)	38.1	43.1	67.6	43.2	54.7	69.1
	a+b_1	Approach + Bag-of-words	47.6	55.2	70.6	62.9	68.5	**77.6**
	a+b_2	Approach + Sentiment distribution	**49.7**	**54.1**	65.9	62.2	68.3	75.9
	a+b	**Approach + Baselines**	48.0	52.3	**71.8**	66.0	68.2	76.7

b_1 **Bag-of-Words.** The frequencies of the tokens in a review. We consider only tokens occurring in $\geq 5\%$ of all training reviews to avoid noise induced by rare tokens. In cross-domain scenarios, such tokens behave unpredictable and usually fail.

b_2 **Sentiment Distribution.** The frequencies of positive, neutral, and negative local sentiment in a review, as well as the first and last local sentiment.

6.1.3. Experiments. Based on either self-classified or ground-truth local sentiment, we predicted 3-class global sentiment with the default configuration of the state-of-the-art ensemble classifier Random Forest [Breiman 2001]. In particular, we learned one classifier for each feature type and different feature sets for all training and test domains. No hyperparameters were optimized, so no knowledge about a test domain was used. To prevent class bias, the training sets were balanced with oversampling. We measured in-domain accuracy in 10-fold cross-validation, averaged over five runs. For out-of-domain accuracy, we applied the learned classifiers to the other corpora. Table VI lists all results for both kinds of local sentiment and all nine domain combinations.

6.1.4. Results for Self-classified Local Sentiment. Bag-of-words (b_1) proves strong in some in-domain tasks, but it consistently fails out-of-domain. Although less clear, similar holds for b_2, so a restriction to the distribution of local sentiment does not suffice to tackle domain dependence. In contrast, the normalized sentiment flows (a_1) and especially the abstracted sentiment flows (a_2) are more effective out-of-domain. Either of them is the best single feature type in all six out-of-domain experiments.

In the movie domain, we obtain an overall accuracy of 71.8. I.e., we were able to outperform Pang and Lee [2005] who report about 75 on the reviews of one author, but only 63 for the other. Partly due to an improved local sentiment classification, we also beat our highest accuracy result on the hotel reviews (78) presented in [Wachsmuth 2015]. In the product domain, we fail to compete with Täckström and McDonald [2011] who achieve 66.6 after training on large-scale corpora. Our small product training set

explains the limited in-domain accuracy; even some out-of-domain classifiers perform better. Also, our sentiment analysis approach (**a**) performs worse than the baselines within the domains, achieving an accuracy of 63.9 compared to 68.7 on average. This is significantly worse at $p < 0.05$ according to a student t-test.

Across domains, however, the bottom lines of all domains in the left part of Table VI provide evidence for the robustness of our approach: the out-of-domain accuracy consistently improves when using the flow patterns. Averaging over the six out-of-domain experiments, our approach (**a**) achieves an accuracy of 52.7 compared to 38.7 of baseline b_1, 46.6 of b_2, and 42.5 of **b**. Thus, **a** performs significantly better at $p < 0.01$ in all cases. All features together (**a+b**) obtain 53.1, which is even significant at $p < 0.005$.

6.1.5. Results for Ground-truth Local Sentiment. The availability of ground truth allows an estimation of what could be achieved on the given corpora in theory. The right part of Table VI shows the high impact of flow patterns: Only in the hotel in-domain experiment, the sentiment distribution (b_2) performs stronger. When training on product reviews, our approach even performs better alone than together with the baselines.

6.1.6. Conclusion. The latter results demonstrate that the sentiment flow is often decisive for a review's global sentiment. Practically, our flow features did not beat the baselines within the domains. Still, they added to the state-of-the-art-like overall effectiveness, indicating that the flows make usually disregarded aspects of reviews measurable. The real impact of the flows becomes obvious out-of-domain, though, where they significantly improved robustness (with slight advantages for the abstracted flows). This suggests that similar sentiment flows are used across domains of reviews. While very high effectiveness seems to require more adaptation to a target domain, the modeling of flows can thus serve as a basis for aligning effective features between domains.

6.2. State-of-the-Art Scoring of the Organization of Essays

In Section 5.1, we have seen that different types of flow patterns can be found in an essay, which indicate the quality of the essay's organization. Now, we report on our experiments with the scoring of this organization based on flow patterns. We evaluated major parameters of our approach, relying on the following set-up.

6.2.1. Argumentative Texts. For training and testing, we used the student essay dataset from Table I with organization scores from $[1.0, 4.0]$ again. As in Section 5.1, we processed all essays with the algorithms from [Persing et al. 2010; Socher et al. 2013; Stab and Gurevych 2014] to obtain local sentiment, discourse functions, and argument roles. This time, we considered these rhetorical moves also on the sentence level.

6.2.2. Approach and Baselines. We modeled three feature types for the rhetorical moves:

- a_1 **Sentiment Features.** (1) The *sentiment distribution* features already used in Section 6.1. (2) The respective numerical *normalized flows* with 15 (paragraph level) or 30 (sentence-level) dimensions. (3) The respective *abstracted flows*. For simplicity, we used all possible 16 combinations of 0–3 of the abstractions from Section 3.1.
- a_2 **Discourse Function Features.** Analog to a_1, the *discourse function distribution*, the *normalized flows*, and the *abstracted flows* (ignoring *body* in case of *Fewer Classes*). In addition, we computed the distribution of discourse function n-grams, i.e., sequences of $n \leq 3$ consecutive discourse functions. We created both nominal and numerical flows to evaluate the benefit of numerical values. To this end, we mapped *body* to 1.0, *introduction* and *conclusion* to 0.5, and *rebuttal* to 0.0.
- a_3 **Argument Role Features.** The *argument role distribution*, the *normalized flows*, and the *abstracted flows* (ignoring *none* in case of *Fewer Classes*). The paragraph-level features are analog to a_1. On the sentence level, we computed the distribution of argument unit n-grams (for major claims, claims, premises, and none). Also, the

Table VII. Mean average error (MAE) and mean squared error (MSE) in scoring essay organization for all proposed feature types. The different rhetorical moves are computed either on the sentence level or on the paragraph level.

Rhetorical Moves	#	Feature Type	Sentence level MAE	Sentence level MSE	Paragraph level MAE	Paragraph level MSE
Sentiment		Sentiment distribution	0.425 ±.016	0.349 ±.030	0.422 ±.017	0.349 ±.032
		Normalized flows (numerical)	0.425 ±.016	0.349 ±.030	0.418 ±.015	0.348 ±.038
		Abstracted flows (numerical)	0.425 ±.014	0.349 ±.030	0.371 ±.011	0.231 ±.019
	a_1	**Sentiment features**	0.423 ±.013	0.350 ±.027	**0.369 ±.014**	**0.228 ±.022**
Discourse functions		Discourse function distribution	0.421 ±.016	0.346 ±.031	0.416 ±.016	0.342 ±.032
		Normalized flows (nominal)	0.417 ±.012	0.347 ±.033	0.405 ±.016	0.305 ±.027
		Normalized flows (numerical)	0.415 ±.017	0.343 ±.033	0.398 ±.022	0.299 ±.055
		Abstracted flows (nominal)	0.413 ±.020	0.342 ±.039	0.385 ±.027	0.247 ±.031
		Abstracted flows (numerical)	0.421 ±.009	0.351 ±.031	0.376 ±.025	0.231 ±.036
	a_2	**Discourse function features**	0.420 ±.018	0.344 ±.043	**0.358 ±.028**	**0.210 ±.035**
Argument roles		Argument role distribution	0.375 ±.015	0.231 ±.019	0.388 ±.023	0.258 ±.030
		Normalized flows (nominal)	0.390 ±.015	0.302 ±.040	0.376 ±.022	0.242 ±.029
		Abstracted flows (nominal)	0.367 ±.022	0.234 ±.017	0.360 ±.025	0.225 ±.030
	a_3	**Argument role features**	**0.345 ±.021**	**0.198 ±.022**	0.361 ±.025	0.224 ±.029
	b_1	Average baseline	0.425 ±.016	0.349 ±.030	0.425 ±.016	0.349 ±.030

sentence-level flows represented only single paragraphs, hence being captured in terms of their frequency in an essay.

Based on the feature types, we predicted organization scores. To assess effectiveness, we compared the feature types to three baselines:

b_1 **Average Baseline.** This lower-bound baseline simply assigns the average score of the training essays to all test essays.
b_2 **Bag-of-Words.** The frequencies of all tokens occurring in $\geq 10\%$ of the training set.
b_3 **State of the Art.** The best approach of Persing et al. [2010]. To our knowledge, this approach denotes the state of the art in organization scoring.

6.2.3. Experiments. For direct comparison, we replicate the experimental set-up of Persing et al. [2010], who score essay organization with supervised regression based on sequences of discourse functions (the relation between ours and their approach is discussed in Section 7). In particular, we measured the mean absolute error (MAE) and the mean squared error (MSE) of regression, using cross-validation on the five predefined folds of the given corpus. Like the authors, we trained linear support vector machines, relying on LibSVM [Chang and Lin 2011]. Unlike them, we did not optimize the SVM cost parameter but simply set it to 0.1 in all cases after some initial tests.

6.2.4. Results. Table VII lists errors and their standard deviations for each proposed feature type on both granularity levels. Consistently, the abstracted flows beat the normalized flows and the latter beat the distribution of the respective rhetorical moves. The mapping from nominal discourse functions to numerical values indeed proves beneficial, reducing the MSE from 0.247 to 0.231 on the paragraph level, among others. In case of sentiment and discourse functions, the paragraph-level features clearly outperform the sentence-level features. A comparison with the average baseline (b_1) reveals that hardly any impact is achieved based on sentences. In contrast, the sentence-level argument role features succeed, even denoting the best-performing of all feature types with an MAE of 0.345 and an MSE of 0.198. We used all types whose values are shown bold in Table VII in the final evaluation of our complete approach.

The overall effectiveness of our approach is shown in Table VIII. Together, the chosen feature types (**a**) have about the same MSE as the state-of-the-art baseline (b_3). While bag-of-words (b_2) fails in isolation, it adds to our approach when combining them (**a**+b_2), reducing the MAE to 0.322 and the MSE to 0.171. These results are slightly (though not significantly) better than those of Persing et al. [2010].

Table VIII. MAE and MSE of our complete approach and two baselines in scoring essay organization.

#	Feature Type	MAE	MSE
a_1	Sentiment features (paragraph level)	0.369 $\pm$.014	0.228 $\pm$.022
a_2	Discourse function features (paragraph level)	0.358 $\pm$.028	0.210 $\pm$.035
a_3	Argument role features (sentence level)	0.345 $\pm$.021	0.198 $\pm$.022
b_1	Average baseline	0.425 $\pm$.016	0.349 $\pm$.030
b_2	Bag-of-words	0.422 $\pm$.015	0.345 $\pm$.003
b_3	State of the art [Persing et al. 2010]	0.323	0.175
a	**Approach** (a_1–a_3)	0.328 $\pm$.018	0.176 $\pm$.019
a+b_2	**Approach** + Bag-of-words	**0.322** $\pm$.013	**0.171** $\pm$.016

6.2.5. Conclusion. The flow patterns achieved the best results known so far for the given essay scoring task. Compared to the former state of the art [Persing et al. 2010], additional gains were obtained through the resort to sentence-level argument roles, underlining the connection of essay organization and argumentation quality. In general, the dominance of the abstracted over the normalized flows suggests that a computationally expensive clustering could be omitted. This would enable the use of flow patterns on big data. Also, the abstracted flows come with fewer parameters (cf. Section 3). Still, in nearly all cases, combining both flow types (a_1–a_3) performed best.

7. RELATED WORK

Argumentation is in the focus of current research [Al-Khatib et al. 2016]. We do not deal with dialogical argumentation, such as online debates [Cabrio and Villata 2012], but we target at texts that comprise a monological argumentation, such as reviews, essays, or scientific articles. Their goal is to persuade the intended reader of a thesis or opinion on some topic [Besnard and Hunter 2008]. Faulkner [2014] points out that only in this setting true argumentative structure is found, while dialogs—especially online debates—are often dominated by short, emotion-loaded text fragments.

Several approaches have been proposed to mine argumentative structure from texts, often grounded in models from theory. Mochales and Moens [2011] seek for conclusions and premises in legal cases based on [Walton et al. 2008]. Relying on [Freeman 2011], Peldszus and Stede [2015] capture support and attack relations in microtexts. For the web, Habernal and Gurevych [2015] adapt the model of Toulmin [1958], which defines facts and warrants that support a claim unless a rebuttal applies. Shallower models consider only claims and evidence [Rinott et al. 2015]. Common to these approaches is the focus on the local structure of single *arguments* and their relations. In contrast, we model and investigate the global structure of complete *argumentations*. An argumentation composes a set of arguments to justify a thesis. The thesis may be explicit, implicit, or encoded in an overall judgment (such as the rating of a review). In experiments with the argument types from [Stab and Gurevych 2014], we have also modeled the completeness of local structure inspired by [Park and Cardie 2014].

From a linguistics perspective, an argumentation can be viewed as a "regulated" sequence of speech acts [Walton and Godden 2006]. As related research, we model it as a sequence of rhetorical moves [Swales 1990]. For instance, Teufel et al. [2009] capture argumentative zones of scientific articles for retrieval purposes, such as the statement of a research goal. Our model can integrate such zones—e.g., to learn about clear or well-organized ways of arguing scientifically. Generally, our model allows a flexible choice of rhetorical moves with regard to the requirements of a given analysis task. As shown, we represent a review's argumentation regarding its global sentiment by the sequence of local sentiments in the review, among others. The correlation of local and global sentiment has already been analyzed (e.g., by Täckström and McDonald [2011]), but not for the impact of structure. Further rhetorical moves have been evaluated in our experiments, partly derived from argumentation mining.

Our focus is not argumentation *mining* in terms of capturing argumentative structure, but it is the argumentation *analysis* of the previously captured structure (here, in the form of a flow). Following Mochales and Moens [2011], a complete argumentation analysis investigates the content and linguistic structure of composed arguments, relations between the arguments, underlying beliefs, and the coherence of the discussed topic. E.g., Brüninghaus and Ashley [2003] analyze the types of reasoning used in individual arguments to predict the outcomes of legal cases. Here, we have identified patterns in the argumentative structure of complete texts at the discourse level. While Feng and Hirst [2011] perform related analyses to classify argumentation schemes, we have used the structure to address specific argumentation-related analyses.

In related work, Faulkner [2014] uses a proprietary model of single arguments, derived from dependency trees, to classify stance. Sobhani et al. [2015] investigate the use of argumentation mining for the same task. Ong et al. [2014] exploit heuristically found argument units to assess general essay quality, and Persing and Ng [2015] do similar to score an essay's argument strength. Based on [Mann and Thompson 1988], Feng et al. [2014] extract long-distance discourse relations to measure text coherence, but even they do not capture discourse-level argumentation.

In contrast, Persing et al. [2010] rely on the sequence of all discourse functions in an essay (both on the sentence level and on the paragraph level) to score the essay's organization. Given an unknown sequence, they first determine its k nearest-neighbor sequences in a training set and derive scores from these sequences. The scores as well as the representations of discourse function subsequences are then encoded in linear, string, and alignment kernels. Finally, the kernels are combined in a composite kernel that is fed to a supervised learning algorithm. Similar to Persing et al. [2010], we align unknown and known discourse-level sequences. We identify and consider only common known sequences, though, which prevents an overfitting to noise in the training set. Unlike Persing et al. [2010], we directly use the sequences as similarity features, making fully transparent what our approach actually measures. At the same time, we can easily capture any type of information in the sequences on any level in order to find the most decisive sequences, as shown. Moreover, the resulting features can be immediately integrated with any other features, thus further improving flexibility. Altogether, our approach can hence be seen as generalization of [Persing et al. 2010], which we can approximate if needed, as we have demonstrated in Section 6.

The core idea of our approach is to analyze the flow of rhetorical moves in a text. This idea is derived from empirical analyses of a review corpus [Wachsmuth et al. 2014b]. There, we identify frequent *local sentiment flows* in reviews that cooccur with particular global sentiment. Such flows have been introduced by Mao and Lebanon [2007], who use conditional random fields to classify sentence sentiment in reviews depending on preceding sentence sentiment. They also predict a review's global sentiment from all local sentiments. However, they represent each position in a flow as a single feature. Thereby, they disregard the flow's ordering and, thus, do not capture overall structure. We overcome this limitation by comparing a complete flow to a set of known flows. This resembles explicit semantic analysis [Gabrilovich and Markovitch 2007], which compares texts based on their similarity to complete Wikipedia articles. While the content of each article represents one concept, our flows abstract from content.

We claim that our model of discourse-level argumentation benefits effectiveness in argumentation-related analysis tasks. To support this claim, we have evaluated the model in sentiment analysis and essay scoring. We do not compete with state-of-the-art sentiment analysis, which learns the composition of words in sentences [Socher et al. 2013], but we use it to create local sentiment flows from which we infer global sentiment. Regarding essay scoring, we are not interested in classical aspects like grammar

or vocabulary usage [Dikli 2006]. Instead, we look at an important aspect of the argumentation of essays, such as the evaluated quality of organization.

Finally, we aim at domain robustness: Most text analyses use, more or less, domain-specific features and, thus, they tend to fail in other domains [Daumé and Marcu 2006]. Domain adaptation tackles this problem, e.g., by learning structural correspondences between domains based on a few domain-independent pivot features [Blitzer et al. 2006]. While adaptation requires some texts from the target domain for training, the proposed flow features strive for domain independence directly and, so, could serve as pivot features. Among others, such domain independence is analyzed by Menon and Choi [2011] for function words in authorship attribution. To our knowledge, however, we are the first to model the discourse-level structure of texts for robustness.

Originally, we introduced our approach in previous work for one type of argumentative texts, namely, reviews [Wachsmuth et al. 2014a]. There, we modeled global structure as a flow of clause-level sentiments. More details and extended evaluations followed in [Wachsmuth 2015]. Later, we provided empirical evidence that sentiment flows generalize across topical domains [Wachsmuth et al. 2015]. While we have built on these works here, we have also gone significantly further. In particular, we have created flows based on types of rhetorical moves never used before (argument roles and discourse functions), we have explicitly compared different ways to represent (nominal vs. numeric) flows and to identify patterns (clustering vs. frequency counts) for the first time, and we have addressed another argumentation-related analysis task (organization scoring). In doing so, we laid the ground for our final discussion of the suitability of flows as a universal model for analyzing the discourse-level argumentation of texts.

8. DISCUSSION AND CONCLUSION

In this paper, we propose to model an argumentative text as a flow of rhetorical moves. The flow representation matches the view of argumentation as a regulated sequence of speech acts [Walton and Godden 2006]. It builds upon our observation that many user-generated argumentative texts are written in an ad-hoc fashion rather than comprising a well-planned argumentation. Based on the flow, the global structure of a text can be analyzed by comparing it to common flow patterns. We even claim that the flow may serve as a universal model for discourse-level argumentation analysis.

To achieve universality, the flow model should apply to all (or at least most) analysis tasks that relate to the discourse-level argumentation of texts. We have considered the modeling and analysis of the local sentiment flow of a review for global sentiment classification as a running example to illustrate our ideas. Clearly, sentiment will often not be in the focus of an argumentation analysis, although it is closely connected to an author's stance in case of reviews [Wachsmuth et al. 2014b]. However, we have presented alternatives to sentiment, and we have evaluated the existence and benefit of flow patterns for other rhetorical moves, such as discourse functions and argument roles. Our results for sentiment analysis and the scoring of essay organization indicate that flows enable an effective analysis and generalize well across domains.

Still, several other analysis tasks exist that address discourse-level argumentation to a larger or minor extent. Besides real stance classification and other argumentation-related essay scoring tasks, the relevance or convincingness of an argumentative text may be analyzed based on the flow of different evidence types used in the text [Rinott et al. 2015]. Also, argumentative zones [Teufel et al. 2009] in a scientific article seem useful to assess its quality (e.g., in terms of clarity) or to identify its type (e.g., approach vs. resource vs. survey). We plan to study such tasks in future work. In contrast, we expect that there will also be argumentation-related tasks where our approach hardly helps, especially tasks that target at the content of arguments. We see this not as a contradiction to the universality claim: When referring to an analysis at the *discourse*

level, we mean the assessment of argumentation properties that emerge from the composition of arguments in a text along with the way this composition is conveyed. We believe that such properties are always affected by the global structure of the text.

A question is when to prefer our model over classical representations of argumentative structure, such as a tree-like graph built from the argument units and relations in a text. The latter complies with concepts from argumentation theory and can express more complex, non-linear interactions [Toulmin 1958; Walton et al. 2008]. This can be important for well-planned argumentation, as e.g. found in legal cases. Still, we expect our model to be able to capture at least sequential argument chains in such argumentation. For discourse-level analyses, such as the classification or scoring of a text, we argue that the abstraction towards sequential structures is favorable: First, it reduces the search space of global structures and hence facilitates a reliable determination of patterns that are discriminative for output classes or values. Second, the separation of a fixed structure (the flow) and flexible content (task-related rhetorical moves) in our model allows an adaptation to the specific needs in an analysis task at hand, which in the end improves expressiveness over classical representations. And third, the creation of a flow will often be more efficient than the construction of an argument graph. This becomes decisive in practical applications that deal with big data.

Altogether, this article provides a comprehensive picture of our model. The general use of flows to analyze the global structure of argumentative texts is straightforward. We have outlined major factors for the success of flows, though not all have been evaluated here. For sentiment analysis, further experiments can be found in the publications this article builds on [Wachsmuth et al. 2014a; Wachsmuth et al. 2015; Wachsmuth 2015]. For other tasks, the benefits and limitations of flows are still to be explored.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-Domain Mining of Argumentative Text through Distant Supervision. In *Proc. of the 15th NAACL: HLT*. 1395–1404.

Philippe Besnard and Anthony Hunter. 2008. *Elements of Argumentation*. The MIT Press.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. In *Proc. of the 2006 EMNLP*. 120–128.

Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

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*. 233–242.

Elena Cabrio and Serena Villata. 2012. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In *Proc. of the 50th ACL: Short Papers*. 208–212.

Sung-Hyuk Cha. 2007. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *Int. J. of Mathematical Models and Methods in Applied Sciences* 1, 4 (2007), 300–307.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3.

Hal Daumé, III and Daniel Marcu. 2006. Domain Adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research* 26, 1 (2006), 101–126.

Semire Dikli. 2006. An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment* 5, 1 (2006).

Adam Robert Faulkner. 2014. *Automated Classification of Argument Stance in Student Essays*. Dissertation. City University of New York.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In *Proc. of the 25th COLING: Technical Papers*. 940–949.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying Arguments by Scheme. In *Proc. of the 49th ACL: HLT - Volume 1*. 987–996.

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

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proc. of the 20th IJCAI*. 1606–1611.

Sylviane Granger, Estelle Dagneaux, and and Magali Paquot Fanny Meunier. 2009. International Corpus of Learner English (Version 2). (2009).

Ivan Habernal and Iryna Gurevych. 2015. Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse. In *Proc. of the 2015 EMNLP*. 2127–2137.

Johannes Kiesel, Khalid Al-Khatib, Matthias Hagen, and Benno Stein. 2015. A Shared Task on Argumentation Mining in Newspaper Editorials. In *Proc. of the 2nd Workshop on Argumentation Mining*. 35–38.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8, 3 (1988), 243–281.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Yi Mao and Guy Lebanon. 2007. Isotonic Conditional Random Fields and Local Sentiment Flow. *Advances in Neural Information Processing Systems* 19 (2007), 961–968.

Rohith Menon and Yejin Choi. 2011. Domain Independent Authorship Attribution without Domain Adaptation. In *Proc. of the RANLP 2011*. 309–315.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation Mining. *AI and Law* 19, 1 (2011), 1–22.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-Based Argument Mining and Automatic Essay Scoring. In *Proc. of the First Workshop on Argumentation Mining*. 24–28.

Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proc. of the 43rd ACL*. 115–124.

Joonsuk Park and Claire Cardie. 2014. Identifying Appropriate Support for Propositions in Online User Comments. In *Proc. of the First Workshop on Argumentation Mining*. 29–38.

Andreas Peldszus and Manfred Stede. 2015. Joint Prediction in MST-style Discourse Parsing for Argumentation Mining. In *Proc. of the 2015 EMNLP*. 938–948.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling Organization in Student Essays. In *Proc. of the 2010 EMNLP*. 229–239.

Isaac Persing and Vincent Ng. 2015. Modeling Argument Strength in Student Essays. In *Proc. of the 53rd ACL and the 7th IJCNLP*. 543–552.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show Me Your Evidence – An Automatic Method for Context Dependent Evidence Detection. In *Proc. of the 2015 EMNLP*. 440–450.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From Argumentation Mining to Stance Classification. In *Proc. of the 2nd Workshop on Argumentation Mining*. 67–77.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proc. of the 2013 EMNLP*. 1631–1642.

Christian Stab and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proc. of the 2014 EMNLP*. 46–56.

John M. Swales. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Oscar Täckström and Ryan McDonald. 2011. Discovering Fine-grained Sentiment with Latent Variable Structured Prediction Models. In *Proc. of the 33rd ECIR*. 368–374.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards Discipline-independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proc. of the 2009 EMNLP*. 1493–1502.

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

Henning Wachsmuth. 2015. *Text Analysis Pipelines—Towards Ad-hoc Large-scale Text Mining*. Lecture Notes in Computer Science, Vol. 9383. Springer.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment Flow – A General Model of Web Review Argumentation. In *Proc. of the 2015 EMNLP*. 601–611.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014a. Modeling Review Argumentation for Robust Sentiment Analysis. In *Proc. of the 25th COLING: Technical Papers*. 553–564.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014b. A Review Corpus for Argumentation Analysis. In *Proc. of the 15th CICLing*. 115–127.

Douglas Walton and David M. Godden. 2006. *Considering Pragma-Dialectics*. Erlbaum, Chapter The Impact of Argumentation on Artificial Intelligence, 287–299.

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