Modeling Deliberative Argumentation Strategies on Wikipedia

Khalid Al-Khatib[†] **Henning Wachsmuth**[‡] **Kevin Lang**[†] **Jakob Herpel**[†]
Matthias Hagen[§] **Benno Stein**[†]

[†] Bauhaus-Universität Weimar
Webis Group, Faculty of Media
<firstname>.<lastname>@uni-weimar.de

[‡] Paderborn University
Computational Social Science Group
henningw@upb.de

[§] Halle University
matthias.hagen@informatik.uni-halle.de

Abstract

This paper studies how the argumentation strategies of participants in deliberative discussions can be supported computationally. Our ultimate goal is to predict the best next deliberative move of each participant. In this paper, we present a model for deliberative discussions and we illustrate its operationalization. Previous models have been built manually based on a small set of discussions, resulting in a level of abstraction that is not suitable for move recommendation. In contrast, we derive our model statistically from several types of metadata that can be used for move description. Applied to six million discussions from Wikipedia talk pages, our approach results in a model with 13 categories along three dimensions: discourse acts, argumentative relations, and frames. On this basis, we automatically generate a corpus with about 200,000 turns, labeled for the 13 categories. We then operationalize the model with three supervised classifiers and provide evidence that the proposed categories can be predicted.

1 Introduction

Deliberation is the type of discussions where the aim is to find the best choice from a set of possible actions (Walton, 2010). This type is influential for making decisions in different processes including *collaborative writing*. Studies have shown the positive impact of deliberation on the quality of several document types, such as scientific papers, research proposals, political reports, and Wikipedia articles, among others (Kraut et al., 2012).

However, deliberative discussions may fail, either by agreeing on the wrong action, or by reaching no agreement. While the former is hard to measure, the latter is, for example, clearly reflected in the number of disputed discussions on Wikipedia (Wang and Cardie, 2014).

Although agreement can never be guaranteed, a deliberative argumentation strategy of a discussion's participants makes it more likely (Kittur et al., 2007). With *strategy*, we here mean the sequence of moves that participants take during the discussion. Such a sequence is effective if it leads to a successful discussion. To achieve effectiveness, every participant has to understand the current state of a discussion and to come up with a next deliberative move that *best* serves the discussion. For newcomers, this requires substantial effort and time, especially when a discussion grows due to conflicts and back-and-forth arguments. Here, automated tools can help by annotating ongoing discussions with a label for each move or by providing a textual summary of past moves (Zhang et al., 2017a,b). A way to go beyond that is to let the tool *recommend the best possible moves* according to an effective strategy. This is the ultimate goal of our research.

As a substantial step towards this goal, two fundamental research questions are addressed in the paper at hand: (1) How to model deliberative discussions in light of the aim of agreement, and (2) how to operationalize the model in order to identify different argumentation strategies and to learn about their effectiveness.

Different models of deliberative discussions have been proposed in previous studies. These models were developed based on expert analyses of a *small* set of sampled discussions (see Section 2). However, the small size, in fact, confines the ability to develop a *representative* model, which should ideally cover a wide range of moves while being abstract to fit the majority of discussions.

To overcome this limitation, we propose to derive a model statistically from a large set of discussions. We approach this based on different types of

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 2545–2555
Melbourne, Australia, July 15 - 20, 2018.

metadata that people use to describe their moves on Wikipedia talk pages, the richest source of deliberative discussions on the web. Particularly, we extract the entire set of about six million discussions from all English Wikipedia talk pages. We parse each discussion to identify its structural components, such as turns, users, and time stamps. Besides, we store four types of *metadata* from the turns: the user tag, a shortcut, an in-line template, and links. To learn from the metadata, we cluster the types' instances based on their semantic similarity. Then, we map each cluster to a specific concept (e.g., 'providing a source'), and the related concepts into a set of categories (e.g., 'providing evidence'). Table 2 shows the categories of our model.

Analyzing the distribution of these categories, we find that each turn ideally should have (1) one of six categories that we call *discourse acts*, (2) one of three categories that we call *argumentative relations*, and (3) one of four categories that we call *frames*. As such, our model is in line with three well-established theories: *speech act theory* (Searle, 1969), *argumentation theory* (Peldszus and Stede, 2013), and *framing theory* (P. Levin et al., 1998). A model instance is sketched in Figure 1.

Based on the model, we generate a new large-scale corpus using the metadata automatically: *Webis-WikiDebate-18* corpus. Basically, if a turn in a discussion has metadata that belongs to a specific category according to the above-mentioned analysis, it is labeled with that category. The corpus includes 2400 turns labeled with a discourse act, 7437 turns labeled with a relation, and 182,321 turns labeled with a frame.

To operationalize our model, we train three supervised classifiers for acts, relations, and frames on the corpus. The classifiers employ a rich set of linguistic features that has been shown to be effective in similar tasks (Ferschke et al., 2012). The results of our experiments suggest that we are able to predict the labels with a comparable performance to the one achieved in similar tasks.

Overall, the contribution of this paper is threefold: (1) A data-driven approach for creating a new model of deliberative discussions that is aligned with well-established theories, (2) a corpus with about 200,000 turns labeled for 13 different categories, and (3) a classification approach that predicts the labels of turns. All developed resources are freely available at `https://www.webis.de/data/data.html`.

2 Related Work

Modeling deliberative discussions in Wikipedia has been already addressed in different studies. The central goal of these studies is to minimize the coordination effort among discussion participants. In particular, Ferschke et al. (2012) have proposed a model of 17 dialogue acts, each belonging to one of four categories: article criticism, explicit performative, information content, and interpersonal. The model was derived by performing a manual analysis of 30 talk pages in the Simple English Wikipedia. Based on the model, a new corpus of 1367 turns has been created and used to train and evaluate a multi-label classifier for predicting the model's acts. Another model is the one proposed by Viegas et al. (2007). The model consists of 11 different dialogue acts. These acts have been used to manually label 25 talk pages from the English Wikipedia. Furthermore, Bender et al. (2011) have developed a model for authority claims and alignment moves in Wikipedia discussions. The model then has been used to label 47 talk pages.

Rooted in the limitation of being derived from a small sample, these models obtain low coverage and/or are over-abstracted. This is indicated by labels such as 'other' (Viegas et al., 2007) or by a very abstract 'information providing' act (Ferschke et al., 2012), which covers 78% of the turns. We argue that recommending such moves for new participants will not be useful. On the other hand, the model of Ferschke et al. (2012) does not include anything similar to 'propose alternative action', for example, although such a concept was shown to be important in deliberative dialogues (Walton, 2010).

Moreover, no existing model distinguishes the three dimensions of turns: act, relation, and frame. They either consider only one dimension or mix an act with a relation, such as in the label: 'criticizing unsuitable or unnecessary content' (Ferschke et al., 2012). This is a problem for predicting the next best deliberative move. For example, consider a discussion about adding new content to an article, where the participants support the action with different acts (e.g., 'providing evidence'), but all of them consider the 'writing quality' frame. A new turn attacks the action by providing evidence that the action would violate the 'neutral point of view'. The best next move should actually consider this frame, since no content that violates 'neutral point of view' policy should be added, regardless of its adherence to the 'writing quality'.

(Computational Linguistics) Merge

I think that this article should probably be merged with Computational linguistics, but I'm fairly new to the Wikipedia, so I'm not sure. Lambda 22:55, 22 Feb 20164 (UTC)

	Act	Relation	Frame
Disagree While they're related, they're not really the same thing. Computational linguistics tries to use computer techniques to better understand linguistics as a discipline, while NLP tries to build ways for a computer to understand language. See the top answer here: www.quora.com/How-is-computational-linguistics-different-from-natural-language-processing. It is a nice explanation from an expert. Delirium 22:58, Feb 22, 20116 (UTC)	Providing evidence	Attack	Verifiability and factual accuracy
proposal I think we can merge them and call the article' Computational linguistics and Natural Language processing'. That solve the problem. It doesn't violate any rule, I guess. 21.59.174.21 13:26, 23 June 2016 (UTC)	Recommending an act	Support	Writing quality
Based on WP:MOS, they should be merged on one article with the name of the most used term (if they are similar)24.59.194.44 13:27, 23 June 2016 (UTC)	Enhancing the understanding	Attack	Writing quality
Do CL and NLP have separate conferences? 24.59.194.44 13:28, 23 June 2016 (UTC)	Asking a question	Neutral	Verifiability and factual accuracy
I think ACL conferences and Coling have both CL and NLP papers. 24.59.194.4 13:296, 23 June 2016 (UTC)	Enhancing the understanding	Neutral	Verifiability and factual accuracy
Thanks for your answer. 24.59.194.44 13:29, 23 June 2006 (UTC)	Socializing	Neutral	Dialogue management

Figure 1: Left: An excerpt of a discussion in a Wikipedia talk page. Right: The labels of each turn in the discussion according to our proposed model.

In contrast, our approach of deriving the model using thousands of different 'descriptions' of moves written by the numerous Wikipedia users is, in our view, more likely to give a representative picture of how people argue in deliberative discussions. This, in turn, leads not only to high coverage, but also to better abstraction. Our model is in line with three well-known theories, which we summarize in the next paragraph.

Speech act is a widely accepted theory in pragmatics (Searle, 1969). Based on this theory, many research papers have been proposed for modeling different domains, such as one-on-one live chat (Kim et al., 2010), persuasiveness in blogs (Anand et al., 2011), twitter conversations (Zarisheva and Scheffler, 2015), and online dialogues (Khanpour et al., 2016). In the context of *argumentation theory* (Peldszus and Stede, 2013), agreement detection is a related direction of work which has been studied in discussions (Rosenthal and McKeown, 2015). Notably, Andreas et al. (2012) annotated 822 turns from 50 talk pages with three labels: 'agreement', 'disagreement', and 'non'. Anyhow, over the last few years, argumentation mining became a hot topic in our community, where several studies have went beyond the agreement detection to investigate the identification of the 'support' and 'attack' relations in argumentation discourses (Peldszus and Stede, 2013). Finally, *framing* is one of the important theories in discourse analysis (Entman, 1993). This theory has been studied widely in different domains, such as news article (Naderi and Hirst, 2017) and political debates (Tsur et al., 2015). These three theories back up the essence of our proposed model. We found that a participant in a discussion writes her text considering a specific act, an argumentative relation, and a frame.

The metadata in Wikipedia have been used for different tasks. The 'infobox' has been exploited in the tasks of question answering (Morales et al., 2016) and summarization (Ye et al., 2009), among others. Moreover, Wang and Cardie (2014) have used specific discussion templates to identify discussions that are disputed. Besides Wikipedia, metadata such as 'point for', 'point against', and 'introduction' have been used successfully for modeling argumentativeness in debate platforms (Al-Khatib et al., 2016a). Also, The metadata for user interactions, such as the 'delta indicator' and users votes in Reddit ChangeMyView discussions have been used to model the persuasiveness of a text (Tan et al., 2016).

We started the investgation of strategies for writing argumentative texts in previous work. In (Al-Khatib et al., 2016b), we have presented a corpus for argumentation strategies in news editorials. We then used this corpus and other data in (Al-Khatib et al., 2017) to identify patterns of strategies across different general topics. In contrast to those two studies targeting monological texts, here we address argumentation strategies in dialogical texts.

3 Modeling Deliberative Discussions

The web is full of platforms where users can share and discuss opinions, beliefs, and ideas. In case of deliberative discussions, in particular, participants try to find the best action from several choices. Apparently, the participants there follow a strategy to achieve an effective discussion, i.e., each participant tries to come with the best deliberative move that leads to achieve the goal of discussion.

The numerous deliberative discussions on these platforms do not only include user-written text, but also different types of metadata that users add to benefit the coordination between them. For example, users vote for specific posts, summarize texts, include references to the sources they use, refer to the discussion policies of a platform, or report bad behavior of others. Overall, the available metadata represents a valuable resource that provides insights into three main aspects of a discussion: The functions of users' moves, the users' roles, and the discussion topics along with their flows. We propose to exploit the metadata for modeling argumentation strategies in deliberative discussions.

To this end, we proceed in four general steps: (1) *metadata inspection*, which includes investigating the used metadata and its functions, (2) *concept origination*, where clusters of similar metadata are created and mapped to corresponding concepts, (3) *concept categorization*, where similar concepts are abstracted into a defined set of categories, and (4) *category composition*, where possible overlaps between categories should be identified.

The idea of this approach is not only to model the strategies, but also to allow for an operationalization of the resulting model by providing a dataset for training classifiers. In particular, the metadata can also be used to label discussions based on distant supervision (Mintz et al., 2009). In the following, we describe how we implement our approach to derive a new model of Wikipedia discussions, using the metadata provided by the participants.

3.1 Discussion Parsing

As part of the management policies of Wikipedia, each article has an associated page called 'Talk'. The main purpose of the talk page is to allow users to discuss how to improve the article through specific actions that they agree on. Most of these discussions can be seen as deliberative, since all participants share the same goal: finding the best action to improve the article.

When a user has a proposal on how to improve an article, she can open a discussion on the article's talk page, specifying a title and the main topic of discussion. Usually, the topic denotes a suggestion to perform a specific action, such as adding, merging, or deleting certain content of the article, among others. Ideally, multiple users then participate in the discussion about whether the action would improve the article or not.

Each single comment written by a user at a specific time is called a 'turn'. A turn may reply directly to the main topic of the discussion or to any other turn. Overall, a discussion consists of the title, the main topic, and a number of turns written by users with attached time stamps (see Figure 1). Based on a manual inspection of the turns' texts of 50 discussions, we found four general types of metadata used by the participants: *user tags*, *shortcuts*, *inline-templates*, and *external links*.

To derive a model from Wikipedia, we need to extract and parse the whole set of discussions on all talk pages, including both ongoing and closed ones. This is all but trivial, particularly due to the fact that the creation of a discussion is solely done by the users; although Wikipedia describes the required format of the different parts of a discussion in detail, not all users follow the format, often forgetting required symbols or mistakenly confusing a symbol with another one. In the implementation of our approach, we built upon the English Wikipedia dump created on March 1st, 2017. Given a Wikipedia dump, we parse it in the following steps:

Extraction of Talk Pages First, we obtain the talk pages. We use the Java Wikipedia Library (JWPL) from Zesch et al. (2008), which converts a Wikipedia dump into a database that provides an easy-to-use access to the dump components.

Extraction of Discussions Next, we extract the discussions from the talk pages. To this end, we develop several regular expressions that capture the format for starting and ending a discussion.

Corpus Component	Instances
Page	5 807 046
Discussion	5 941 534
Discussion template	144 824
Turn	20 816 860
Registered users	739 244
Turns by registered users	10 926 670
Turns by anonymous user	9 890 190
Tag	99 889
Shortcut	425 583
Inline template	3 382 443
Links	4 824 085
Turns with tag and shortcut	2 347
Turns with tag and inline template	61 521
Turns with shortcut and inline template	170 065

Table 1: Instance counts of the different components of the Webis-WikiDiscussions-18 corpus.

Identification of Structure Given the discussion, we identify their structure. We created a specific template to mine the title. The topic of the discussion is simply given by the first turn. To identify and correctly segment all users' turns, we use several indicators, for instance, indentations.

Identification of Turn Metadata Finally, we identify the metadata of each turn. We analyzed how users include the tags in their turns, finding that they usually start a turn with a user tag in triple quotation marks. A shortcut starts with 'WP:', followed by a name for the shortcut, together encapsulated by brackets. Also templates are placed between double parentheses, but they do not start with 'WP:'. Links are simply identified by either of the affixes 'www.' and 'http:'.

3.2 The Webis-WikiDiscussions-18 Corpus

The result of the parsing process is a large-scale corpus of Wikipedia discussions. In particular, the *Webis-WikiDiscussions-18* corpus we created contains about six million discussions, consisting of about 20 million turns. The turns comprise around 74,000 different tags with a total of about 100,000 instances, around 7000 different shortcuts with about 400,000 instances, and around 51,000 different inline templates with about 3.3 million instances. Half of the turns are written by registered users. Table 1 lists the exact instance counts.

3.3 Model Derivation

We now explain how we derive a model of deliberative discussions from the metadata obtained in the previous subsection. The derivation process includes the four steps outlined in the beginning of this section.

Metadata Inspection As mentioned before, a turn on Wikipedia includes up to four types of metadata: user tag, shortcut, inline template, and external link. Each type has a specific definition, a suggested usage, and properties that we discuss in the following paragraphs.

A *user tag* is a short text that a discussion participant uses to describe or summarize her contribution. Most tags indicate the main function of the contribution, such as 'proposal' and 'question'. Users can define any free-text tag they want using a noun, verb, etc. Analyzing the tags in the crawled discussions, we found the most frequent tags to be rather general and meaningful, whereas less frequent tags often capture aspects of the topic of discussion, such as 'Israel-Venezuela relations' in the discussion about 'Foreign relations of Israel'. Sometimes, tags are used to get the attention of specific users, such as 'For who reverted my change'. Unfortunately, many users also misuse tags, for example, by including the whole turn's text there or by encoding meaningless information.

A *shortcut* is an abbreviation text link that redirects the user to some page on Wikipedia. Although shortcuts may link to any Wikipedia page, they are often used to link to rules or policies. The respective pages belong to one of five categories:
(1) Behavioral guidelines: Pages that describe how users should interact with each other (e.g., during a discussion). This includes that users should be "good-faith" (WP:AGF), among others.
(2) Content guidelines: Pages that describe how to identify and include information in the articles, such as those about how an article should have reliable and accepted sources (WP:RELIABLE).
(3) Style guidelines: Pages that contain advice on writing style, formatting, grammar, and similar. This includes how to write the introduction (WP:LEAD) and headings (WP:HEADINGS), and what style to use for the content (WP:MOS).
(4) Notability guidelines: Pages that illustrate the conditions of testing whether a given topic warrants its own article. The most common shortcut in this category is (WP:N).
(5) Editing guidelines: Pages that provide information on the metadata of articles, such as the articles' categories (WP:CAT).

Overall, we found that shortcuts are used particularly frequently for style, content, and behavioral

guidelines in Wikipedia discussions. The participants mainly use them to discuss the impact of applying an action that has been proposed to be performed on a Wikipedia article. For example, adding a lot of content to the introduction of an article may violate the style guidelines. A user can indicate this by referring to the style rules using the shortcut (WP:LEAD).

An *inline template* is a Wikipedia page that has been created to be included in other pages. Inline templates usually comprise specific patterns that are used in many articles, such as standard warnings or boilerplate messages. For example, there are templates for including a quotation, citation, or code, among others. Templates are used frequently in Wikipedia discussions, with the objective of writing readable and well structured turns.

An *external link*, finally, points to a web page outside Wikipedia. External links occur both in Wikipedia articles and in Wikipedia discussions. While there are some restrictions for using them in articles, they can be used without restriction in discussions. We found that these links are used in Wikipedia discussions to point to evidence on the linked web pages. In particular, they often link to research, news, search engines, educational institutions, and blogs.

Concept Origination We analyzed the usage of the four types of metadata in Wikipedia discussions and identified a set of concepts. Each concept primarily describes the turn that a participant writes:

User tags: We explored all 376 tags that occurred at least 35 times. As discussed before, the tags could be seen as a keywords that describe the turns. Often, different tags refer to the same concept, for example, 'conclusion', 'summary', and 'overall' all capture the concept of 'summarization', i.e., the main function of the respective turns is to summarize the discussion. As a result, we identified 32 clusters. We examined some turns belonging to each cluster, and mapped each cluster to a specific concept that describes it.

Shortcuts: Analogously, we explored all 99 shortcuts that occurred at least 900 times. Since the shortcuts themselves do not describe the turn, but rather the policy pages they refer to, we analyzed these pages by reading their first paragraphs and by checking their relation to the pages of the five shortcut categories we discussed before (e.g., 'behavioral'). This resulted in the identification of 12 concepts. We found that each shortcut concept describes the main quality aspect that a turn addresses. For example, 'writing content' specifies how a proposed action influences the quality of the writing of the associated article.

Inline-templates: Our investigation of this type led only to concepts that we already found before for the tags and shortcuts, such as 'stating a fact'.

External links: Similar to the templates, we identified concepts in the links that we also observed in the tags, such as 'providing source'.

Concept Categorization The concepts that we identified in the user tags can be grouped into six categories that we see as 'discourse acts':

1. *Socializing:* All concepts related to social interaction, such as thanking, apologizing, or welcoming other users.
2. *Providing evidence:* All concepts concerning the provision of evidence. Evidence may be given in form of a quote, an example, a fact, references, a source, and similar.
3. *Enhancing the understanding:* All concepts related to helping users understand the topic of discussion or a discussion itself. This can be done by giving background information, by clarifying misunderstandings, or by summarizing the discussion, among others.
4. *Recommending an act:* All concepts proposing to add a new aspect to the discussion, to ask more users to participate in the discussion, or to come up with an alternative to the proposed action.
5. *Asking a question:* All concepts related to questions serving different purposes, such as obtaining information on the topic of discussion, requesting reasons of specific decisions, and similar.
6. *Finalizing the discussion:* All concepts related to the decision of a discussion, including reporting the decision, committing it, or closing the discussion to move it to the archive.

In addition, we identified three further categories based on the user tags, which we see as relevant to 'argumentation theory'. Each represents a relation between the turn and the topic of discussion or between the turn and another turn:

1. *Support relation:* The turn agrees with or supports another turn or the topic of discussion,

for instance, by providing an argument in favor of the one in the 'supported' turn.

2. *Attack relation:* The opposite of the 'support relation', i.e., the turn disagrees or attacks another turn or the topic of discussion.
3. *Neutral relation:* The turn has a neutral relation to another turn or the topic of discussion when it neither support nor attack it.

Finally, we identified four categories based on the shortcuts that we see as relevant to 'framing theory'. They target a quality dimension of the article or of the discussion itself:

1. *Writing quality:* Turns that mainly address issues related to the quality of writing of an article, such as whether adding new content complies with the style guidelines for lead sections, the layout, or similar.
2. *Verifiability and factual accuracy:* Turns that address issues related to the quality of references, the reliability of sources, copyright violations, plagiarism, and similar.
3. *Neutral point of view:* Turns that focus on a fair representation of viewpoints and on how to avoid bias.
4. *Dialogue management:* Turns that concentrate on issues related to managing the discussion, such as reporting abusive language, preserving respect between users, encouraging newcomer participants, and similar.

Category Composition Given these categories, we investigated the interaction between them in 20 discussions, for instance, to see whether the categories are orthogonal. We found that each turn may have one discourse act, one relation, and one frame at the same time. For example, a turn may support another turn by providing evidence (say, of the type 'source'), while focusing on the writing quality frame. Table 2 shows the categories of our model and their concepts.

4 Model Operationalization

In this section, we present the operationalization process of our proposed model for deliberative argumentation strategies. First, we explain the construction of *Webis-WikiDebate-18*: a large-scale corpus for our model that we generated automatically based on the metadata in discussions. Then, we discuss the development and evaluation of a classification approach which we use for predicting the model's categories.

4.1 The Webis-WikiDebate-18 Corpus

To create a corpus for our model, we decided to rely again on the metadata. In particular, for each category in our model, we retrieved the metadata instances that had been used to derive the category, and then labeled any turn that included any metadata with this category. For example, the user tag 'overall' was used to originate the concept 'summarization', which was abstracted into the category 'enhancing the understanding'. Accordingly, all the turns that included this tag were labeled with the category 'enhancing the understanding'. This process is in line with the distant supervision paradigm. In case a turn contained metadata belonging to two categories, we excluded it from the corpus. This happened with some shortcuts in particular. Basically, such cases indicate that some turns address more than one frame.

Overall, the corpus comprises 2400 turns labeled with one of the six discourse act categories, 7437 turns with one of the relation categories, and 182,321 turns with one of the frame categories. In order to verify the reliability of the corpus, we randomly sampled about 100 turns from each category, ensuring that all the category's concepts are taken into consideration. The turns in the samples were verified (i.e., whether they belong to the assigned category) by a worker hired from the freelancing platform `upwork.com`. The worker was a native speaker of English with deep expertise in writing. Table 3 shows statistics of the corpus, including the percentage of turns in each sample that belong to the assigned category according to the expert. In general, this verification result is comparable to the inter-annotator agreement achieved in some related studies (Ferschke et al., 2012).

4.2 Classification Approach

Based on the Webis-WikiDebate-18 corpus, we develop three supervised classifiers: one for the discourse acts, one for the relations, and one for the frames. Since this paper does not aim at proposing a novel approach for the classification tasks, but rather at showing the ability to operationalize the model, we follow existing work that has proposed methods for the tasks at hand. Particularly, we implement a rich set of features that have been used by others before. These features capture lexical, semantic, style, and pragmatic properties of turns.

Dimension	Category	Concepts
Discourse act	Socializing	(1) Thank a user, (2) Apologize from a user, (3) Welcome a user, (4) Express anger
	Providing evidence	(1) Provide a quote, (2) Reference, (3) Source, (4) Give an example, (5) State a fact, (6) Explain a rational
	Enhancing the understanding	(1) Provide background info, (2) Info on the history of similar discussions, (3) Introduce the topic of discussion, (4) Clarify a misunderstanding, (5) Correct previous own or other's turn, (6) Write a discussion summary, (7) Conduct a survey on participants, (8) Request info
	Recommending an act	(1) Propose alternative action on the article, (2) Suggest a new process of discussion, (3) Propose asking a third party
	Asking a question	(1) Ask a general question about the topic, (2) Question a proposal or arguments in a turn
	Finalizing the discussion	(1) Report the decision, (2) Commit the decision, (3) Close the discussion
Argumentative relation	Support	(1) Agree, (2) Support
	Neutral	(1) Be neutral.
	Attack	(1) Disagree, (2) Attack, (3) Counter-attack
Frame	Writing quality	(1) Naming articles, (2) Writing content, (3) Formatting, (4) images, (5) Layout and list
	Verifiability and factual accuracy	(1) Reliable sources, (2) Proper citation (3) Good argument
	Neutral point of view	(1) Neutral point of view
	Dialogue management	(1) Be bold. (2) Be civil, (3) Don't game the system

Table 2: The concepts covered by each category of each of the three principle dimensions of our model.

Dimension	Category	Turns	Prec.
Discourse act	Socializing	83	0.71
	Providing evidence	781	0.49
	Enhancing the understanding	671	0.56
	Recommending an act	137	0.82
	Asking a question	106	0.71
	Finalizing the discussion	622	0.71
Argumentative relation	Support	2895	1.00
	Neutral	1937	0.63
	Attack	2605	1.00
Frame	Writing quality	19893	0.51
	Verifiability and factual ac.	72049	0.89
	Neutral point of view	60007	0.89
	Dialogue management	30372	0.74

Table 3: Number of turns in each category of Webis-WikiDebate-18 corpus and the precision of sampled turns for each category according to an expert.

In short, we used the following features: The frequency of word 1–3-grams, character 1–3-grams, chunk 1–3-grams, function word 1–3-grams, and of the first 1–3 tokens in a turn. The number of characters, syllables, tokens, phrases, and sentences in a turn. the frequencies of part-of-speech tag 1–3-grams. The mean SentiWordNet score of the words in a turn (http://sentiwordnet.isti.cnr.it). The frequency of each word class of the General Inquirer (http://www.wjh.harvard.edu/~inquirer). The depth level of turns in the discussion. For the relation classifier, we had additional features that consider the target of the relation (the parent turn), namely, the cosine, euclidean, manhattan, and jaccard similarity between turn and parent turn.

4.3 Experiments and Results

As a preprocessing step, we cleaned the turns in the *Webis-WikiDebate-18* Corpus by removing all the metadata: user tags, shortcuts, user and time stamps, etc. Then, we grouped the turns that belong to the discourse act categories in a single dataset (say, the 'discourse act dataset'). The same was performed for the turns belonging to relations and frames. We then split each of the three datasets randomly into training (60%), development (20%), and test (20%) sets. We ensured that turns from the same discussion should appear only in either of the split sets, in order to avoid biasing the classifiers by topical information.

We trained different machine learning models on the training sets and evaluated them on the development sets. The models included those which had been used before in similar tasks, such as naive bayes, logistic regression, support vector machine, and random forest. We tried both under and over-sampling on the training sets. The best results in the three tasks were achieved by using support vector

Dimension	Category	Prec.	Rec.	F_1
Discourse act	Socializing	0.14	0.11	0.13
	Providing evidence	0.63	0.77	0.69
	Enhancing the understand.	0.62	0.55	0.58
	Recommending an act	0.13	0.09	0.10
	Asking a question	0.80	0.19	0.31
	Finalizing the discussion	0.67	0.74	0.71
Argumentative relation	Support	0.53	0.59	0.56
	Neutral	0.55	0.50	0.52
	Attack	0.50	0.49	0.50
Frame	Writing quality	0.74	0.47	0.57
	Verifiability and factual ac.	0.62	0.74	0.67
	Neutral point of view	0.59	0.56	0.58
	Dialogue management	0.64	0.56	0.60

Table 4: The precision, recall, and F_1-score of our classifiers for all categories of the three dimensions.

machine without sampling the training sets.

We used the support vector machine implementation from the LibLinear library (Fan et al., 2008) on the test sets and report the results in Table 4. Overall, the three classifiers achieved results that are comparable to the results of previous methods on the corresponding tasks (Ferschke et al., 2012; Zhang et al., 2017a). We obtained the best results in the frame task, followed by relations and then discourse acts. Apparently, the results correlate with the size of the datasets. In case of discourse acts, the classifier achieves low F_1-scores for 'socializing', 'recommending an act', and 'asking a question'. These categories have a significantly smaller number of turns compared to other categories, which makes identifying them harder. The effectiveness of classifying the relation and frame categories, on the other hand, appears promising given the difficulty of these tasks.

We point that we considered mainly the turns' texts in our experiments. In principle, this helps to get an idea about the effectiveness of our approach in Wikipedia as well as other registers for discussions. Nevertheless, including the metadata and structural information of the analyzed discussions is definitely worthwhile in general, and will naturally tend to lead to notably higher effectiveness.

5 Discussion and Conclusion

While our approach to modeling argumentation strategies in deliberative discussions may seem Wikipedia-specific, the derivation of concepts and categories from metadata can be transferred to other online discussion platforms. We expect the general derivation steps to be the same, whereas the techniques applied within each step may differ depending on the types, frequency, and quality of metadata. For example, the consistent usage of the most common user tags in Wikipedia discussions helps originating concepts manually. In contrast, other metadata might require the use of computational methods, such as clustering, keyphrase extraction, and textual entailment.

Unlike previous approaches to the modeling of discussions on Wikipedia, our model decouples the three principle dimensions of discussions: *discourse acts*, *argumentative relations*, and *frames*. We argue that the distinction of these dimensions is key to develop tool support for discussion participants, for example, for recommending the best possible move in an ongoing discussion.

Also, our model helps analyzing the influence of user interaction and behavior on the effectiveness of discussion decisions. For example, some Wikipedia users focus on the frame 'well written' while ignoring others, which may negatively affect the accuracy of an article's content. Also, users often attack other turns, instead of considering neutral acts such as clarifications of misunderstandings.

Many categories in our model will apply to deliberative discussions in general, particularly the discourse acts and argumentative relations. While the found frames are more Wikipedia-specific, similar play a role on collaborative writing platforms. For example, when writing a scientific paper, possible frames are the 'writing quality' or the 'verifiability of content and citations'.

Besides the model, we created two large-scale corpora: The *Webis-WikiDiscussions-18* corpus, including the entire set of Wikipedia discussions (at the time of parsing) with annotated discussion structure and metadata, and the *Webis-WikiDebate-18* corpus, where turns are labeled for their discourse acts, argumentative relations, and frames. We believe that these corpora will help foster research on tasks such as argument mining, among others.

Finally, we operationalized our Wikipedia discussion model in three support vector machine classifiers with tailored features. Our experiment results confirm that categories of our model can be predicted successfully. In future work, we plan to study how to distinguish effective from ineffective discussions based on our model as well as how to learn from the strategies used in successful discussions, in order to predict the best next deliberative move in an ongoing discussion.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Kohler, and Benno Stein. 2016a. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, pages 1395–1404. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of Argumentation Strategies across Topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1351–1357. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016b. A News Editorial Corpus for Mining Argumentation Strategies. In *Proceedings of the 26th International Conference on Computational Linguistics, COLING*, pages 3433–3443. Association for Computational Linguistics.

P. Anand, J. King, Jordan Boyd-Graber, E. Wagner, C. Martell, Douglas Oard, and Philip Resnik. 2011. Believe Me—We Can Do This! Annotating Persuasive Acts in Blog Text. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. 2012. Annotating Agreement and Disagreement in Threaded Discussion. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC*, pages 818–822.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating Social Acts: Authority Claims and Alignment Moves in Wikipedia Talk Pages. In *Proceedings of the Workshop on Languages in Social Media*, pages 48–57. Association for Computational Linguistics.

Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58.

Rong En Fan, Kai-Wei Chang, Cho-Jui Hsieh, X.-R. Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *JMLR*.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 777–786. Association for Computational Linguistics.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. In *Proceedings of 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, COLING*, pages 2012–2021.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying Dialogue Acts in One-on-one Live Chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 862–871. Association for Computational Linguistics.

Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI*, pages 453–462. ACM.

Robert E. Kraut, Paul Resnick, Sara Kiesler, Yuqing Ren, Yan Chen, Moira Burke, Niki Kittur, John Riedl, and Joseph Konstan. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing, ACL*, pages 1003–1011. Association for Computational Linguistics.

Alvaro Morales, Varot Premtoon, Cordelia Avery, Sue Felshin, and Boris Katz. 2016. Learning to Answer Questions from Wikipedia Infoboxes. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1930–1935. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. 2017. Classifying Frames at the Sentence Level in News Articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*, pages 536–542.

Irwin P. Levin, Sandra L. Schneider, and Gary J. Gaeth. 1998. All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects. *Organizational Behavior and Human Decision Processes*, 76(2):149 – 188.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31.

Sara Rosenthal and Kathy McKeown. 2015. I Couldn't Agree More: The Role of Conversational Structure in Agreement and Disagreement Detection in Online Discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL*, pages 168–177.

J.R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cam: Verschiedene Aufl. Cambridge University Press.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International World Wide Web Conference*, pages 613–624.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL-IJCNLP*, pages 1629–1638. Association for Computational Linguistics.

Fernanda B. Viegas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. 2007. Talk Before You Type: Coordination in Wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, HICSS '07, pages 78–. IEEE Computer Society.

Douglas Walton. 2010. Types of Dialogue and Burdens of Proof. In *Frontiers in Artificial Intelligence and Applications*, volume 216, pages 13–24.

Lu Wang and Claire Cardie. 2014. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. In *Proceedings of the52nd Annual Meeting of the Association for Computational Linguistics, ACL*, volume 2, pages 693–699. Association for Computational Linguistics.

Shiren Ye, Tat-Seng Chua, and Jie Lu. 2009. Summarizing Definition from Wikipedia. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing, ACL*, pages 199–207. Association for Computational Linguistics.

Elina Zarisheva and Tatjana Scheffler. 2015. Dialog Act Annotation for Twitter Conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL*, pages 114–123.

Torsten Zesch, Christof Muller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC*. European Language Resources Association (ELRA).

Amy X. Zhang, Bryan Culbertson, and Praveen Paritosh. 2017a. Characterizing Online Discussion Using Coarse Discourse Sequences. In *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media, ICWSM*, pages 357–366.

Amy X. Zhang, Lea Verou, and David Karger. 2017b. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW*, pages 2082–2096. ACM.