Cross-Domain Mining of Argumentative Text through Distant Supervision

Khalid Al-Khatib Henning Wachsmuth Matthias Hagen Jonas Köhler Benno Stein
Faculty of Media, Bauhaus-Universität Weimar, Germany
<firstname>.<lastname>@uni-weimar.de

Abstract

Argumentation mining is considered as a key technology for future search engines and automated decision making. In such applications, argumentative text segments have to be mined from large and diverse document collections. However, most existing argumentation mining approaches tackle the classification of argumentativeness only for a few manually annotated documents from narrow domains and registers. This limits their practical applicability. We hence propose a distant supervision approach that acquires argumentative text segments automatically from online debate portals. Experiments across domains and registers show that training on such a corpus improves the effectiveness and robustness of mining argumentative text. We freely provide the underlying corpus for research.

1 Introduction

Argumentation mining attracts much attention recently: it is an important building block of applications like automated decision making (Bench-Capon et al., 2009) or pro-and-con search engines (Cabrio and Villata, 2012c). In such applications, argumentation mining usually consists of solving three tasks for each document: (1) Identifying all argumentative text segments in the document, (2) classifying the type of each segment, and (3) classifying relations between the segments.

In this paper we focus on the first task taking on the retrieval perspective of a search engine: Given a large-scale collection of documents (e.g., the web) and a query on some topic, return all argumentative text segments relevant to the topic. Among others, a classifier is needed for this task that can distinguish argumentative from non-argumentative segments. Since we cannot presuppose a specific domain or register within a general retrieval scenario, the classifier needs to robustly deal with documents from diverse domains and registers. In this regard the following two key issues arise.

First, existing approaches to classifying argumentativeness usually focus on specific text domains (e.g., education) and registers (e.g., student essays). Therefore, many used features capture not only local linguistic properties of a text segment, but also global document properties (e.g., that a segment is part of the introduction). Such kinds of features tend to be effective only within a certain domain or a particular register while often failing for others.

Second, all major existing approaches follow a supervised learning scheme based on manual annotation of argumentative text segments. However, the annotation of arguments is particularly intricate and thus expensive due to the complex linguistic structure and the partly subjective interpretation of argumentativeness. Different types of argumentative and non-argumentative segments may come in any order, segment boundaries are not always unambiguous, and parts of an argument may be implicit. Studies reveal that annotators need multiple training sessions to identify and classify argumentative segments with moderate inter-annotator agreement, and crowdsourcing-based annotation does not help notably (Habernal et al., 2014). I.e., a high-quality manual annotation will not scale to large numbers of documents from diverse domains and registers.

Proceedings of NAACL-HLT 2016, pages 1395–1404,
San Diego, California, June 12-17, 2016.

We propose a solution to the outlined issues. In particular, we follow the idea of distant supervision to construct a large-scale corpus of text segments from diverse domains and registers annotated with respect to argumentativeness. Distant supervision is a well-known idea for training robust statistical classifiers. Here, we exploit online debate portals that (1) contain argumentative and non-argumentative text segments for several controversial topics, and that (2) are organized in a semi-structured form, allowing to derive annotations from it.

In several experiments we compare classifiers trained on the constructed corpus to those trained on existing corpora for argumentation mining. We classify argumentativeness using a rich set of lexical, syntax, and indicator feature types. Our results suggest that the new corpus is the most robust resource for classifying argumentative text segments across domains and registers. In addition, we observe that n-grams seem to be most domain-dependent, while syntax features turn out to be more robust.

The contribution of this paper is three-fold: First, through distant supervision we acquire a large corpus with 28,689 argumentative text segments from the online debate portal idebate.org. The corpus covers 14 separate domains with strongly varying feature distributions. It will be made freely available to other researchers.[1] Second, we obtain a robust classifier for argumentativeness, providing evidence that distant supervision does not only save money and time, but also benefits the effectiveness of cross-domain and cross-register argumentation mining. Third, we evaluate—for the first time—the robustness of several features in classifying argumentativeness across domains and registers.

Altogether, the paper serves as a starting point for bringing argumentation mining to practice. We expect that a robust identification of arguments will be a core module of future search engines, as it allows to provide rationales for retrieved documents. To this end, the search engines also need to identify the most *relevant* arguments for a given topic. The paper concludes with ideas on how to assess argument relevance with resources that are obtained through applying our proposed distant supervision technique to other datasets.

[1] http://www.uni-weimar.de/medien/webis/corpora

2 Related Work

Argumentation mining is still in an early stage of investigation, although several promising approaches have been proposed in the last years. Our survey of the argumentation mining literature especially covers three respects: (1) favored domains and registers, (2) techniques for annotation acquisition, and (3) the exploitation of debate portals. We combine these research lines in our approach to tackle argumentativeness classification across domains.

The existing argumentation mining approaches achieve classification accuracies ranging from 73% and 86% (Stab and Gurevych, 2014b; Levy et al., 2014; Palau and Moens, 2009) but they deal with texts from one register or one narrow domain only. For instance, Palau and Moens (2009) address the legal domain, Cabrio and Villata (2012b) as well as Boltužić and Šnajder (2014) investigate online debates and discussions, Aharoni et al. (2014) examine Wikipedia articles, Villalba and Saint-Dizier (2012) as well as Wachsmuth et al. (2014a) work on product reviews, Stab and Gurevych (2014a) focus on persuasive essays, and Peldszus (2014) on microtext. In (Wachsmuth et al., 2015), we studied the generality of sentiment-related argumentative structures across domains. In contrast, here we aim at effectiveness in cross-domain argumentation mining, which is useful for practical applications such as argument retrieval from diverse web-scale document collections.

All mining approaches above proceed as follows. Starting point is a complex and often expensive manual annotation of argumentative text segments in a collection of documents, including the segments' roles (e.g., premise or conclusion) and their relations (e.g., support or attack). Then, the classification of argumentativeness, roles, and relations is achieved via supervised machine learning using different linguistic and statistical features. Our approach avoids manual annotation. Instead, we apply distant supervision to automatically acquire annotations.

Distant supervision is a technique to automatically harvest annotations from data that has been compiled and structured intentionally by a user community on the web. Most approaches employing distant supervision so far address the problems of relation extraction (Mintz et al., 2009; Hoffmann et al., 2011) or event extraction (Reschke et al., 2014). A

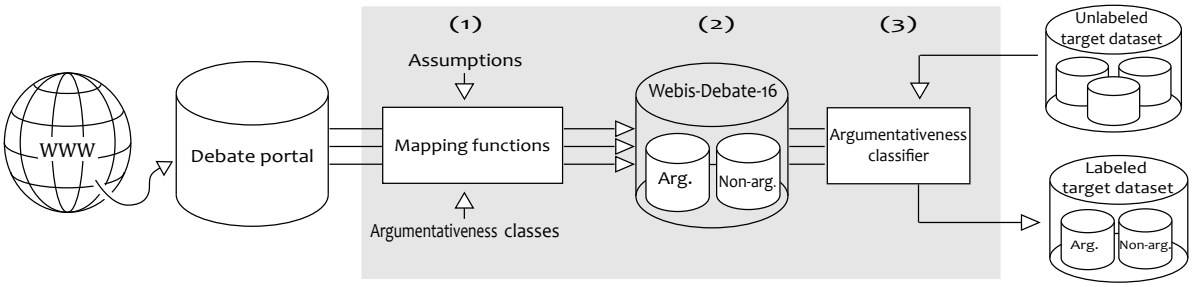

Figure 1. Overview of our distant supervision approach: The mapping functions transform the debate portal content into an annotated corpus for argumentativeness. This corpus is then used to train an argumentativeness classifier.

few others target at sentiment analysis (Marchetti-Bowick and Chambers, 2012) and emotion detection (Purver and Battersby, 2012). In case of the latter, annotations are derived from strong textual indicators like emoticons. In this paper, we exploit metadata from the debate platform idebate.org for mapping texts from the platform to argumentative and non-argumentative classes.

The idea of relying on idebate.org for argument annotation acquisition is in line with related research of Cabrio and Villata (2012c) and Gottipati et al. (2013). In these papers, however, the debate portal is used to infer text-level knowledge only (e.g., stances in debates), but not to generate a complete annotated dataset for argumentativeness.

The work that is most related to ours is the proposal of a method to exploit debate portals for semi-supervised argumentation mining by Habernal and Gurevych (2015). In particular, the authors use word embedding techniques for projecting the texts from debate portals into an annotated argument space, relying on the argument model of Toulmin (1958). On this basis they identify argumentative text segments and their roles. A clear difference to our approach is that Habernal and Gurevych (2015) consider all content of debate portals as argumentative. As a consequence, their approach concentrates mainly on exploiting the debate portals for improving the classification of segment roles, with minor impact on argumentativeness. Moreover, while being comparably effective, our approach aims for simplicity. The reason is that we apply distant supervision to derive a robust resource from the metadata of debate portals only. Thus, we allow for a rich feature space without requiring to use advanced machine learning methods. Finally, Habernal and Gurevych (2015) evaluate their approach only on one dataset from the educational domain, whereas we explicitly aim at robustness across domains. Accordingly, we conduct several experiments on different available corpora (including theirs).

3 Mining Argumentative Text through Distant Supervision

We propose an approach based on the distant supervision paradigm. Our goal is to obtain a classifier that can robustly mine argumentative texts across domains. More precisely, we focus on the task of classifying each segment of a text as being argumentative or not. We assume the text to be separated into segments already.

Our approach consists of three high-level building blocks: (1) Mapping functions that allows an automatic acquisition of argumentativeness annotations from debate portals. (2) A corpus with argumentative and non-argumentative text segments created using the functions. (3) A classifier that can distinguish the two classes of text segments. All building blocks are detailed in the following. Figure 1 depicts an overview of the approach.

3.1 Argumentativeness Mapping Functions

The basic idea of distant supervision is to generate annotations by automatically mapping unlabeled source data to a set of predefined class labels. This requires resources that are related to the given task as well as effective heuristic labeling functions. Typical resources comprise large amounts of data, often in form of user-generated content with semi-structured or structured metadata. Ideally, the resource's metadata substantially eases the mapping to the predefined labels.

In the context of argumentation mining, online debate portals serve as a rich source of argumentative

Class	Metadata	Text
–	Stance	This house believes single-sex schools are good for education.
Non-Argumentative	Introduction	Single-sex schools are schools that only admit those of one specific gender, believing that the educational environment fostered by a single gender is more conducive to learning than a co-educational school. Studies conducted have shown that boys gain more academically from studying in co-education schools, but that girls find segregated schools more conducive to achievement.
Argumentative	Points for	Boys and girls are an unwelcome distraction to each other.
Argumentative	Point	Boys and girls distract each other from their education, especially in adolescence as their sexual and emotional sides develop.
Argumentative	Counterpoint	Any negative effects of co-educational schools have been explained away by studies as the result of other factors, such as classroom size and cultural differences [1]. [1] *Bronski, M., 'Single-sex Schools'. Znet, 25 October 2002.*
Argumentative	Points against	Children need to be exposed to the opposite sex in preparation for later life.
Argumentative	Point	The formative years of children are the best time to expose them to the company of the other gender, in order that they may learn each others' behaviour.
Argumentative	Counterpoint	Children will gain exposure to the opposite sex when they reach adult life; whilst they are young, they should be around those who they feel most comfortable with.

Table 1. Excerpt of a sample discussion from idebate.org: the stance, the introduction, and some points for and against the stance. Except for parts shown in grey, all listed text segments are mapped to the listed classes.

texts on diverse topics. These portals are typically managed by user communities. Textual content can be added via a structured interface that already specifies metadata (e.g., what constitutes a topic or an argument). Thus, mapping text segments from debate portals to classes for argumentation mining is a promising instance of distant supervision.

In particular, we rely on idebate.org. This debate portal has an established community of experienced debaters and volunteers who take care of editing and monitoring semi-structured discussions on various controversial topics, subsumed under 14 high-level themes. A discussion (called "house" in the portal's terminology) starts with a one-sentence *stance* on the respective topic, followed by a more verbose introduction to the topic. Afterwards, points for and against the stance are opposed, both given as a list of arguments. Each argument in turn comes along with points (the argument itself) and counterpoints (counterarguments). Table 1 shows an example.

We downloaded all available discussions from idebate.org. For each discussion, the stance on the topic, the introduction, and the points are extracted from the URL of the web page of the respective discussion. Based on the structure exemplified in Table 1, we stipulate on the following assumptions to automatically map components from the debate portal to annotated argumentativeness instances.

[Component]: *Introduction*

- [Assumption]: The introduction explains the topic and gives important background information in a non-argumentative way.
- [Mapping]: Each sentence in the introduction is an instance of the non-argumentative class.

[Component]: *Points for & Points against*

- [Assumption]: Each point from these lists represents an argument for or against the stance on the topic of discussion.
- [Mapping]: Each point is an instance of the argumentative class.

[Component]: *Point & Counterpoint*

- [Assumption]: The main objective of a point (counterpoint) is to justify (attack) the point in the points-for or points-against list it refers to. We assume that the intention of such a point is to provide reasons for / against an argument.
- [Mapping]: Each sentence in a point / counterpoint is an instance of the argumentative class.

Domain	Documents	Argum. segments	Non-argum. segments
Politics	56	3102	635
Education	40	2057	376
Free speech	31	1435	346
International	113	6190	1324
Religion	8	336	96
Philosophy	10	545	122
Science	3	184	24
Culture	35	1765	307
Environment	11	602	128
Health	35	1985	349
Law	44	2197	440
Society	17	1031	199
Economy	23	1260	288
Sport	19	1191	175
Webis-Debate-16	445	23880	4809

Table 2. Number of documents, argumentative segments, and non-argumentative segments in each domain of our Webis-Debate-16 corpus. Domains correspond to themes from idebate.org.

To optimize the mapping quality, we manually analyzed 50 discussions and then derived three tailored cleansing rules from them: (1) We remove all literature references from the argumentative instances. (2) We delete all special brackets and symbols from the argumentative instances. (3) We delete some keywords from the non-argumentative instances that are used by the community to organize a discussion (e.g., "this house" or "this debate").

3.2 The Webis-Debate-16 Corpus

As a result of applying the defined mapping functions, we obtained a large argumentation mining corpus, called *Webis-Debate-16*. The corpus contains 28,689 text segments from the 14 themes of idebate.org (23,880 argumentative, 4809 non-argumentative). Each theme is assumed to represent one domain. Table 2 lists the distribution of documents over the domains in the corpus. Regarding the number of annotated text segments, Webis-Debate-16 is the largest dataset published so far for argumentation mining. While our review corpus from (Wachsmuth et al., 2014b) is even larger, its annotations are restricted to sentiment-related argumentation. Table 3 compares Webis-Debate-16 to other real argumentation mining corpora, namely, the Essays corpus (Stab and Gurevych, 2014a), the Web discourse corpus (Habernal and Gurevych, 2015), the European Court of Human Rights (ECHR) corpus (Palau and Moens, 2009), and the Araucaria corpus (Reed and Rowe, 2004). The Webis-Debate-16 corpus will be made freely available online.[2]

Corpus	Documents	Argum. segments	Non-argum. segments
Essays	90	1552	327
Web discourse	340	1882	2074
ECHR	47	1067	1449
Araucaria	641	1931	1010
Webis-Debate-16	445	23880	4809

Table 3. Statistics of our Webis-Debate-16 corpus compared to four existing argumentation mining corpora. ECHR is a legal domain corpus that is not publicly available. More details on the others are given in Section 4.

3.3 A Classifier for Argumentativeness

A wide range of statistical and linguistic features has been suggested for argumentation mining and related tasks such as discourse parsing. We employ supervised machine learning to train an argumentativeness classifier based on the features employed by Stab and Gurevych (2014a), Palau and Moens (2009), and Habernal and Gurevych (2015) that cover the following:

Token n-grams: Unigrams, bigrams, and trigrams as Boolean features. In general, n-grams are the most powerful feature type in many related text classification problems (e.g., sentiment analysis).

Discourse markers: Features that represent the existence of words such as "because", which are frequently used in argumentative texts.

Syntax: This feature category contains the number of sub-clauses and production rules.

- Number of sub-clauses: Counter for the number of SBAR tags in the constituency parse tree of a text segment, referring to subordinate clauses in the Penn treebank syntactic tagset.
- Production rules: Boolean features capturing the specific production rules extracted from the constituency parse tree.

Part of speech: Features that capture information related to the parts of speech in a text segment:

[2] http://www.uni-weimar.de/medien/webis/corpora

- Verbs: A boolean feature capturing whether a segment contains a verb. Verbs such as "believe" strongly indicate of argumentative text.
- Adverbs: A boolean feature capturing whether a segment contains an adverb. Many adverbs such as "personally" can play a role in identifying argumentative text.
- Modals: A boolean feature capturing whether a segment contains a modal verb. Modal verbs such as "should" can be important for argumentativeness.
- Verb tense: Boolean features capturing whether a segment contains a past or present tense verb.
- First person pronouns: Pronouns such as "I" and "myself" can be good indicators of claims, a major component of argumentative texts.

Using these features, we train a binary statistical classifier for argumentativeness. Given a set of text segments, the classifier decides for each text segment whether it is argumentative or not.

4 Evaluation

We now report on several in-domain and cross-domain experiments with the classification of argumentativeness. The goals are (1) to demonstrate the effectiveness and robustness of training on the Webis-Debate-16 corpus for cross-domain classification, and (2) to analyze the effectiveness of the proposed features across domains and registers.

4.1 Experimental Setup

To evaluate the effect of using the Webis-Debate-16 corpus for training, appropriate argumentation corpora are needed for comparison. We consider an available corpus as appropriate if (1) the corpus is annotated in a way that allows the distinction of argumentative from non-argumentative text segments, and if (2) the corpus comes with clear annotation guidelines and reported inter-annotator agreement. In addition, we aim at corpora that differ in terms of the covered domains and registers to provide an adequate cross-domain setting. While the Araucaria corpus does not meet the second requirement (Reed and Rowe, 2004), two recently published corpora fulfill both; we refer to them as the *Essays* corpus and the *Web discourse* corpus.

Essays: The Argument Annotated Essays corpus of Stab and Gurevych (2014a) consists of 90 manually annotated persuasive student essays from the education domain. Argumentative text segments are assigned with their type (major claim, claim, or premise). Following Stab and Gurevych (2014b), we consider all sentences that do not have an annotation as being non-argumentative, and the annotated segments as argumentative.

Web discourse: The Argument Annotated User-generated Web Discourse corpus of Habernal and Gurevych (2015) consists of 340 documents from six different topics and four registers. The annotation of arguments is conducted based on the argument model of Toulmin (1958) using five types (claim, premise, backing, rebuttal, and refutation). Again, we consider all annotated text segments as being argumentative and sentences without annotation as being non-argumentative.

Only in case of the Essays corpus, the authors already provide a split into a training and a test set (72 essays for training and 18 for testing). For both the Web discourse corpus and our corpus, we randomly split the document set into 80% for training and 20% for testing. As a result, the training set of the Web discourse corpus consists of 272 documents, and its test set of 68 documents, while the training and test sets of our corpus consist of 356 and 89 documents, respectively.

We train classifiers for each of the above feature types and for the full feature set on the training set of each corpus using the default configuration of the naive Bayes implementation of Weka (Hall et al., 2009). Since all corpora are imbalanced in terms of the number of argumentative and non-argumentative text segments, we perform undersampling for all training sets—an effective technique for largely imbalanced datasets (Japkowicz and Stephen, 2002). All feature values are computed based on the output of the StanfordNLP library (Manning and Klein, 2003). For the different classifiers, we measure the resulting classification performance on all three test sets in terms of accuracy and F_1-score.

4.2 In-Domain Results

Table 4 shows the results of the in-domain experiments. For the full feature set, the achieved F_1-score

	Essays		Web discourse		Webis-Debate-16	
Feature type	**Accuracy**	**F_1-score**	**Accuracy**	**F_1-score**	**Accuracy**	**F_1-score**
N-grams	0.640	0.698	0.815	0.816	0.905	0.908
Syntax	0.599	0.664	0.874	0.874	0.855	0.664
Discourse markers	0.390	0.438	0.584	0.444	0.236	0.180
Part of speech	0.625	0.684	0.541	0.543	0.659	0.702
Full feature set	**0.668**	**0.722**	**0.877**	**0.878**	**0.918**	**0.922**

Table 4. The results of all in-domain experiments on the three corpora for each feature type and the full feature set.

of 0.922 and the accuracy of 0.918 on the Webis-Debate-16 corpus are high compared to those on the Essays and Web discourse corpus. This might be a result of guidelines suggested by the debate portal community, which make the corpus quite homogeneous in terms of style.

Using the full feature set leads to the best results on all three corpora. N-grams denote the most effective single feature type on the Essays copus and on the Webis-Debate-16 corpus, while the syntax features outperform the n-grams on the Web discourse corpus. On the Essays and on the Webis-Debate-16 corpus, the syntax features are sometimes better and sometimes worse than the part of speech features. The discourse markers are the least effective single feature type, largely failing on all test sets, especially in terms of F_1-score.

Note that a comparison to the exact values reported by Stab and Gurevych (2014b) for the Essays corpus and by Habernal and Gurevych (2015) for the Web discourse corpus is not be meaningful due to their experimental set-ups with different class sets. However, their reported results for the non-argumentative class are comparable to the performance we achieved: Stab and Gurevych (2014b) achieve an F_1-score of 0.275 with lexical features and 0.426 with syntax features on the Essays corpus, while Habernal and Gurevych (2015) obtain an F_1-score of 0.718 with lexical features and 0.671 with syntax features on the Web discourse corpus.

4.3 Cross-Domain Results

Table 5 shows the results of the cross-domain experiments. For comparison, we again show the in-domain results in grey color.

As usual, the obtained cross-domain effectiveness values are lower than the in-domain values in most cases and the full feature set usually outperforms feature subsets. One notable exception are the results for the part of speech features on the Essays corpus. The cross-domain effectiveness trained on the Webis-Debate-16 corpus is about six points higher than the in-domain effectiveness in terms of F_1-score and four points in terms of accuracy. For testing on the Web discourse corpus, training on the Webis-Debate-16 corpus using the full features gives the best cross-domain performance. For testing on the Webis-Debate-16 corpus, training on the Web discourse corpus using the n-gram feature type achieves the best cross-domain performance.

Overall, the best corpus for cross-domain classification in our evaluation is clearly the Webis-Debate-16 corpus. Training on Webis-Debate-16 leads to the best cross-domain results for the full feature set and three out of four of the single feature types (n-grams, syntax, and part of speech). Only for the discourse markers, the Web discourse corpus performs better in the cross-domain scenario.

Finally, we observe that the n-grams feature type turns out to be the most domain-dependent in our evaluation. In contrast, both the syntax and the part of speech features appear quite robust across domains. The performance of the discourse markers greatly depends on how frequent they are used in the target domain and register.

Although combining the Webis-Debate-16 corpus to the training datasets of the Essays or the Web discourse corpus increased the performance compared to training only on Webis-Debate-16, it did not outperform the in-domain performance for both corpora. For conciseness, we therefore omit to report the results of our respective experiments here.

4.4 Discussion of our Approach to Robustness

As expected, our experiments reveal the domain dependence of feature distributions in classifying argu-

		Test on Essays		Test on Web discourse		Test on Webis-Debate-16	
Feature type	**Training corpus**	**Accuracy**	**F_1-score**	**Accuracy**	**F_1-score**	**Accuracy**	**F_1-score**
Majority baseline	–	0.867	0.806	0.572	0.417	0.832	0.757
N-grams	Essays	0.640	0.698	0.485	0.488	0.512	0.571
	Web discourse	0.196	0.159	0.815	0.816	**0.854**	**0.867**
	Webis-Debate-16	**0.528**	**0.601**	**0.719**	**0.718**	0.902	0.908
Syntax	Essays	0.599	0.664	0.494	0.497	0.481	0.541
	Web discourse	0.163	0.095	0.874	0.874	**0.767**	**0.795**
	Webis-Debate-16	**0.573**	**0.642**	**0.719**	**0.717**	0.855	0.867
Discourse markers	Essays	0.390	0.438	**0.584**	**0.444**	0.236	0.179
	Web discourse	**0.415**	**0.468**	0.584	0.444	**0.237**	**0.181**
	Webis-Debate-16	0.387	0.434	**0.584**	**0.444**	0.236	0.180
Part of speech	Essays	0.625	0.684	0.490	0.484	**0.560**	**0.616**
	Web discourse	0.446	0.521	0.541	0.543	0.445	0.507
	Webis-Debate-16	**0.686**	**0.724**	**0.538**	**0.533**	0.659	0.702
Full feature set	Essays	0.668	0.722	0.524	0.524	0.483	0.541
	Web discourse	0.181	0.128	0.877	0.878	**0.844**	**0.859**
	Webis-Debate-16	**0.617**	**0.678**	**0.726**	**0.725**	0.918	0.922

Table 5. The results of all cross-domain experiments on the three corpora for each feature type and the full feature set.

mentativeness. This finding emphasizes the importance of explicitly dealing with domain robustness in argumentation mining whenever more than one domain (in terms of a topic, register, or similar) is of interest. To achieve robustness, we have proposed a simple but effective approach that applies distant supervision to create a corpus for classifying argumentativeness. Our results are promising: Classification clearly improves across domains when being trained on our Webis-Debate-16 corpus instead of other available argumentation mining corpora.

The obtained results suggest that our approach can be effectively leveraged to achieve domain robustness. One reason is probably the larger size and domain coverage of our Webis-Debate-16 corpus compared to the other tested corpora. This makes our corpus and the underlying distant supervision idea a valuable resource for research on argumentation. More noise reduction might even further increase the performance of training on the corpus.

In its current form, our corpus contains annotations for distinguishing argumentative from non argumentative text only. While more fine-grained annotations of argumentative texts, such as premise vs. claim, are important for argumentation mining, they cannot be obtained directly from the metadata of idebate.org. Still, the positions of segments in some parts of the debate portal (e.g, point and counterpoint) often indicate whether they are claims or premises. We plan to investigate the exploitation of such information for future versions of the corpus.

So far, we have shown how to create an annotated corpus classifying argumentativeness exploiting one specific debate portal via distant supervision. In principle, our approach is rather general and, thus, could also be applied to other argumentation resources and tasks. Indeed, idebate.org is only one of many web resources with lots of argumentative texts and argumentation-relevant metadata. Aside from debate portals, one such resource is given by Wikipedia talk pages. Very recently, Wikipedia introduced markups within these article discussions, such as *support* or *oppose*. While still being in an early stage, this metadata seems promising to derive argumentative relations from it. We plan to use our distant supervision approach for classifying argumentative relations on such resources. This can then be an important next step to enable the assessment of *argument relevance*—a core building block of an argument retrieval system.

4.5 From Argumentativeness to Relevance

As motivated in the introduction, a retrieval system for arguments not only requires the identification and classification of argumentative text segments. A successful future search engine taking argument

features into account additionally needs a way of ranking arguments according to their relevance. In this regard, we propose a "PageRank for arguments" based on the link network of support and attack relations between arguments.

In particular, given robust algorithms to identify arguments and their relations across web pages (e.g., via distant supervision), we could build an *argument graph* for the web. Related research has already used the argumentation framework of Dung (1995) to find accepted arguments based on such a graph on a much smaller scale (Cabrio and Villata, 2012a). However, the size of the web would allow for recursive analysis of the graph with statistical approaches like the famous PageRank algorithm (Page et al., 1999), enabling an assessment of argument relevance. Several research questions arise from this idea (e.g., how to balance support and attack within the analysis) but argument relevance forms a very important future research direction.

5 Conclusion

Most existing approaches tackle argumentation mining in a supervised manner trained on manually annotated documents from a specific domain. Such approaches neither tend to be effective on documents from other domains, nor do they scale to applications that deal with huge document collections, such as search engines. In this paper, we investigate how to achieve robust performance for argumentation mining across domains, focusing on the classification of the argumentativeness of text segments. In particular, we approach the data side of this problem, namely, we apply distant supervision to automatically create a large annotated corpus with argumentative and non-argumentative text segments from several domains, exploiting metadata from the online debate portal idebate.org.

Based on the created corpus and on common manually annotated corpora, we conduct several in-domain and cross-domain argumentativeness experiments. Our results clearly indicate that training on the created Webis-Debate-16 corpus yield the most robust cross-domain classifier. Thereby, our approach serves as a starting point for bringing argumentation mining to practical applications like search engines. The corpus as well as an implementation of the approach will be made freely available. Besides a robust identification of argumentative segments, search engines will also need to decide which arguments are the most relevant to a given query— a very promising future research direction in the field of argumentation mining.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68.

Trevor Bench-Capon, Katie Atkinson, and Peter McBurney. 2009. Altruism and Agents: An Argumentation Based Approach to Designing Agent Decision Mechanisms. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2, AAMAS 2009*, pages 1073–1080.

Filip Boltužić and Jan Šnajder. 2014. Back up your Stance: Recognizing Arguments in Online Discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Elena Cabrio and Serena Villata. 2012a. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 208–212.

Elena Cabrio and Serena Villata. 2012b. Generating Abstract Arguments: A Natural Language Approach. In *Proceedings of the 2012 Conference on Computational Models of Argument, COMMA 2012*, pages 454–461.

Elena Cabrio and Serena Villata. 2012c. Natural Language Arguments: A Combined Approach. In *20th European Conference on Artificial Intelligence, ECAI 2012*, pages 205–210.

Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77(2):321–357.

Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. 2013. Learning Topics and Positions from Debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1858–1868.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting Debate Portals for Semi-Supervised Argumentation

Mining in User-Generated Web Discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining on the Web from Information Seeking Perspective. In *Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 26–39.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT 2011*, pages 541–550.

Nathalie Japkowicz and Shaju Stephen. 2002. The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.*, 6(5):429–449.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context Dependent Claim Detection. In *Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014*, pages 1489–1500.

Christopher Manning and Dan Klein. 2003. Optimization, Maxent Models, and Conditional Estimation Without Magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials - Volume 5*, pages 8–8.

Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for Microblogs with Distant Supervision: Political Forecasting with Twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2012*, pages 603–612.

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 of the AFNLP: Volume 2, ACL 2009*, pages 1003–1011.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL 2009*, pages 98–107.

Andreas Peldszus. 2014. Towards Segment-based Recognition of Argumentation Structure in Short Texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97.

Matthew Purver and Stuart Battersby. 2012. Experimenting with Distant Supervision for Emotion Classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2012*, pages 482–491.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for Argument Analysis, Diagramming and Representation. *International Journal on Artificial Intelligence Tools*, 13.

Kevin Reschke, Martin Jankowiak, Mihai Surdeanu, Christopher D. Manning, and Dan Jurafsky. 2014. Event extraction using distant supervision. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference, LREC 2014*.

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

Christian Stab and Iryna Gurevych. 2014b. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 46–56.

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

Maria Paz Garcia Villalba and Patrick Saint-Dizier. 2012. Some Facets of Argument Mining for Opinion Analysis. In *Proceedings of the 2012 Conference on Computational Models of Argument, COMMA 2012*, pages 23–34.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014a. Modeling Review Argumentation for Robust Sentiment Analysis. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 553–564.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014b. A Review Corpus for Argumentation Analysis. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 115–127.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment Flow – A General Model of Web Review Argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 601–611, Lisbon, Portugal.