Identifying Argumentative Discourse Structures in Persuasive Essays

Christian Stab[†] **and Iryna Gurevych**[†‡]

[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
www.ukp.tu-darmstadt.de

Abstract

In this paper, we present a novel approach for identifying argumentative discourse structures in persuasive essays. The structure of argumentation consists of several components (i.e. claims and premises) that are connected with argumentative relations. We consider this task in two consecutive steps. First, we identify the components of arguments using multiclass classification. Second, we classify a pair of argument components as either support or non-support for identifying the structure of argumentative discourse. For both tasks, we evaluate several classifiers and propose novel feature sets including structural, lexical, syntactic and contextual features. In our experiments, we obtain a macro F1-score of 0.726 for identifying argument components and 0.722 for argumentative relations.

1 Introduction

Argumentation is a crucial aspect of writing skills acquisition. The ability of formulating persuasive arguments is not only the foundation for convincing an audience of novel ideas but also plays a major role in general decision making and analyzing different stances. However, current writing support is limited to feedback about spelling, grammar, or stylistic properties and there is currently no system that provides feedback about written argumentation. By integrating argumentation mining in writing environments, students will be able to inspect their texts for plausibility and to improve the quality of their argumentation.

An *argument* consists of several components. It includes a claim that is supported or attacked by at least one premise. The *claim* is the central component of an argument. It is a controversial statement that should not be accepted by the reader without additional support.[1] The *premise* underpins the validity of the claim. It is a reason given by an author for persuading readers of the claim. *Argumentative relations* model the discourse structure of arguments. They indicate which argument components are related and constitute the structure of argumentative discourse. For example, the argument in the following paragraph contains four argument components: one claim (in bold face) and three premises (underlined).

> "(1) ***Museums and art galleries provide a better understanding about arts than Internet***. (2) *In most museums and art galleries, detailed descriptions in terms of the background, history and author are provided.* (3) *Seeing an artwork online is not the same as watching it with our own eyes, as* (4) *the picture online does not show the texture or three-dimensional structure of the art, which is important to study.*"

In this example, the premises (2) and (3) support the claim (1) whereas premise (4) is a support for premise (3). Thus, this example includes three argumentative support relations holding between the components (2,1), (3,1) and (4,3) signaling that the source component is a justification of the target component. This illustrates two important properties of argumentative discourse structures. First, argumentative relations are often implicit (not indicated by discourse markers; e.g. the relation holding between (2) and (1)). Indeed, Marcu and Echihabi (2002) found that only 26% of the evidence relations in the RST Discourse Treebank (Carlson et al., 2001) include discourse markers.

[1] We use the term claim synonymously to *conclusion*. In our definition the differentiation between claims and premises does not indicate the validity of the statements but signals which components include the gist of an argument and which are given by the author as justification.

Second, in contrast to Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), argumentative relations also hold between non-adjacent sentences/clauses. For instance, in the corpus compiled by Stab and Gurevych (2014) only 37% of the premises appear adjacent to a claim. Therefore, existing approaches of discourse analysis, e.g. based on RST, do not meet the requirements of argumentative discourse structure identification, since they only consider discourse relations between adjacent sentences/clauses (Peldszus and Stede, 2013). In addition, there are no distinct argumentative relations included in common approaches like RST or the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), since they are focused on identifying general discourse structures (cp. section 2.2).

Most of the existing argumentation mining methods focus solely on the identification of argument components. However, identifying argumentative discourse structures is an important task (Sergeant, 2013) in particular for providing feedback about argumentation. First, argumentative discourse structures are essential for evaluating the quality of an argument, since it is not possible to examine how well a claim is justified without knowing which premises belong to it. Second, methods that recognize if a statement supports a given claim enable the collection of additional evidence from other sources. Third, the structure of argumentation is needed for recommending better arrangements of argument components and meaningful usage of discourse markers. Both foster argument comprehension and recall (Britt and Larson, 2003) and thus increase the argumentation quality. To the best of our knowledge, there is currently only one approach that aims at identifying argumentative discourse structures proposed by Mochales-Palau and Moens (2009). However, it relies on a manually created context-free grammar (CFG) and is tailored to the legal domain, which follows a standardized argumentation style. Therefore, it is likely that it will not achieve acceptable accuracy when applied to more general texts in which discourse markers are missing or even misleadingly used (e.g. student texts).

In this work, we present a novel approach for identifying argumentative discourse structures which includes two consecutive steps. In the first step, we focus on the identification of argument components using a multiclass classification approach. In the second step, we identify argumentative relations by classifying a pair of argument components as either support or non-support. In particular, the contributions of this work are the following: First, we introduce a novel approach for identifying argumentative discourse structures. Contrary to previous approaches, our approach is capable of identifying argumentative discourse structures even if discourse markers are missing or misleadingly used. Second, we present two novel feature sets for identifying argument components as well as argumentative relations. Third, we evaluate several classifiers and feature groups for identifying the best system for both tasks.

2 Related Work

2.1 Argumentation Mining

Previous research on argumentation mining spans several subtasks, including (1) the separation of argumentative from non-argumentative text units (Moens et al., 2007; Florou et al., 2013), (2) the classification of argument components or argumentation schemes (Rooney et al., 2012; Mochales-Palau and Moens, 2009; Teufel, 1999; Feng and Hirst, 2011), and (3) the identification of argumentation structures (Mochales-Palau and Moens, 2009; Wyner et al., 2010).

The separation of argumentative from non-argumentative text units is usually considered as a binary classification task and constitutes one of the first steps in an argumentation mining pipeline. Moens et al. (2007) propose an approach for identifying argumentative sentences in the Araucaria corpus (Reed et al., 2008). The argument annotations in Araucaria are based on a domain-independent argumentation theory proposed by Walton (1996). In their experiments, they obtain the best accuracy (73.75%) using a combination of word pairs, text statistics, verbs, and a list of keywords indicative for argumentative discourse. Florou et al. (2013) report a similar approach. They classify text segments crawled with a focused crawler as either containing an argument or not. Their approach is based on several discourse markers and features extracted from the tense and mood of verbs. They report an F1-score of 0.764 for their best performing system.

One of the first approaches focusing on the identification of argument components is *Argumentative Zoning* proposed by Teufel (1999). The underlying assumption of this work is that argu-

ment components extracted from a scientific article provide a good summary of its content. Each sentence is classified as one of seven rhetorical roles including claim, result or purpose. The approach obtained an F1-score of 0.46[2] using structural, lexical and syntactic features. Rooney et al. (2013) also focus on the identification of argument components but in contrast to the work of Teufel (1999) their scheme is not tailored to a particular genre. In their experiments, they identify claims, premises and non-argumentative text units in the Araucaria corpus and report an overall accuracy of 65%. Feng and Hirst (2011) also use the Araucaria corpus for their experiments but focus on the identification of *argumentation schemes* (Walton, 1996), which are templates for forms of arguments (e.g. argument from example or argument from consequence). Since their approach is based on features extracted from mutual information of claims and premises, it requires that the argument components are reliably identified in advance. In their experiments, they achieve an accuracy between 62.9% and 97.9% depending on the particular scheme and the classification setup.

In contrast to all approaches mentioned above, the work presented in this paper focuses besides the separation of argumentative from non-argumentative text units and the classification of argument components on the extraction of the argumentative discourse structure to identify which components of the argument belong together for achieving a more fine-grained and detailed analysis of argumentation. We are only aware of one approach (Mochales-Palau and Moens, 2009; Wyner et al., 2010) that also focuses on the identification of argumentative discourse structures. However, this approach is based on a manually created CFG that is tailored to documents from the legal domain, which follow a standardized argumentation style. Therefore, it does not accommodate ill-formatted arguments (Wyner et al., 2010), which are likely in argumentative writing support. In addition, the approach relies on discourse markers and is therefore not applicable for identifying implicit argumentative discourse structures.

2.2 Discourse Relations

Identifying argumentative discourse structures is closely related to discourse analysis. As illustrated in the initial example, the identification of argumentative relations postulates the identification of implicit as well as non-adjacent discourse relations. Marcu and Echihabi (2002) present the first approach focused on identifying implicit discourse relations. They exploit several discourse markers (e.g. *'because'* or *'but'*) for collecting large amounts of training data. For their experiments they remove the discourse markers and discover that word pair features are indicative for implicit discourse relations. Depending on the utilized corpus, they obtain accuracies between 64% and 75% for identifying a cause-explanation-evidence relation (the most similar relation of their work compared to argumentative relations).

With the release of the PDTB, the identification of discourse relations gained a lot of interest in the research community. The PDTB includes implicit as well as explicit discourse relations of different types, and there are multiple approaches aiming at automatically identifying implicit relations. Pitler et al. (2009) experiment with polarity tags, verb classes, length of verb phrases, modality, context and lexical features and found that word pairs with non-zero Information Gain yield best results. Lin et al. (2009) show that beside lexical features, production rules collected from parse trees yield good results, whereas Louis et al. (2010) found that features based on named-entities do not perform as well as lexical features. However, current approaches to discourse analysis like the RST or the PDTB are designed to analyze general discourse structures, and thus include a large set of generic discourse relations, whereas only a subset of those relations is relevant for argumentative discourse analysis. For instance, the argumentation scheme proposed by Peldszus and Stede (2013) includes three argumentative relations (support, attack and counter-attack), whereas Stab and Gurevych (2014) propose a scheme including only two relations (support and attack). The difference between argumentative relations and those included in general tagsets like RST and PDTB is best illustrated by the work of Biran and Rambow (2011), which is to the best of our knowledge the only work that focuses on the identification of argumentative relations. They argue that existing definitions of discourse relations are only relevant as a building block for identifying argumentative discourse and that existing approaches do not contain a single relation that corresponds to

[2]Calculated from the precision and recall scores provided for individual rhetorical roles in (Teufel, 1999, p. 225).

a distinct argumentative relation. Therefore, they consider a set of 12 discourse relations from the RST Discourse Treebank (Carlson et al., 2001) as a single argumentative relation in order to identify justifications for a given claim. They first extract a set of lexical indicators for each relation from the RST Discourse Treebank and create a word pair resource using the English Wikipedia. In their experiments, they use the extracted word pairs as features and obtain an F1-score of up to 0.51 using two different corpora. Although the approach considers non-adjacent relations, it is limited to the identification of relations between premises and claims and requires that claims are known in advance. In addition, the combination of several general relations to a single argumentative relation might lead to consistency problems and to noisy corpora (e.g. not each instance of a contrast relation is relevant for argumentative discourse).

3 Data

For our experiments, we use a corpus of persuasive essays compiled by Stab and Gurevych (2014). This corpus contains annotations of argument components at the clause-level as well as argumentative relations. In particular, it includes annotations of *major claims*, *claims* and *premises*, which are connected with argumentative *support* and *attack* relations. Argumentative relations are directed (there is a specified source and target component of each relation) and can hold between a premise and another premise, a premise and a (major-) claim, or a claim and a major claim. Except for the last one, an argumentative relation does not cross paragraph boundaries.

Three raters annotated the corpus with an inter-annotator agreement of $\alpha_U = 0.72$ (Krippendorff, 2004) for argument components and $\alpha = 0.81$ for argumentative relations. In total, the corpus comprises 90 essays including 1,673 sentences. Since it only contains a low number of attack relations, we focus in this work solely on the identification of argument components and argumentative support relations. However, the proposed approach can also be applied to identify attack relations in future work.

4 Identifying Argument Components

We consider the identification of argument components as a multiclass classification task. Each clause in the corpus is either classified as major claim, claim, premise or non-argumentative. So this task includes besides the classification of argument components also the separation of argumentative and non-argumentative text units. We label each sentence that does not contain an argument component as class *'none'*. Since many argument components cover an entire sentence (30%), this is not an exclusive feature of this class. In total, the corpus contains 1,879 instances.

Table 1 shows the class distribution among the instances. The corpus includes 90 major claims (each essay contains exactly one), 429 claims and 1,033 premises. This proportion between claims and premises is common in argumentation since claims are usually supported by several premises for establishing a stable standpoint.

MajorClaim	Claim	Premise	None
90 (4.8%)	429 (22.8%)	1,033 (55%)	327 (17.4%)

Table 1: Class distribution among the instances. The corpus contains 1552 argument components and 327 non-argumentative instances.

For our experiments, we randomly split the data into a 80% training set and a 20% test set with the same class distribution and determine the best performing system using 10-fold cross-validation on the training set only. In our experiments, we use several classifiers (see section 4.2) from the Weka data mining software (Hall et al., 2009). For preprocessing the corpus, we use the Stanford POS-Tagger (Toutanova et al., 2003) and Parser (Klein and Manning, 2003) included in the DKPro Framework (Gurevych et al., 2007). After these steps, we use the DKPro-TC text classification framework (Daxenberger et al., 2014) for extracting the features described in the following section.

4.1 Features

Structural features: We define structural features based on token statistics, the location and punctuations of the argument component and its covering sentence. Since Biran and Rambow (2011) found that premises are longer on the average than other sentences, we add the number of tokens of the argument component and its covering sentence to our feature set. In addition, we define the number of tokens preceding and following an argument component in the covering sentence, the token ratio between covering sentence and argument component, and a Boolean feature that indicates if the

argument component covers all tokens of its covering sentence as token statistics features.

For exploiting the structural properties of persuasive essays, we define a set of location-based features. First, we define four Boolean features that indicate if the argument component is present in the introduction or conclusion of an essay and if it is present in the first or the last sentence of a paragraph. Second, we add the position of the covering sentence in the essay as a numeric feature. Since major claims are always present in the introduction or conclusion of an essay and paragraphs frequently begin or conclude with a claim, we expect that these features are good indicators for classifying (major-) claims.

Further, we define structural features based on the punctuation: the number of punctuation marks of the covering sentence and the argument component, the punctuation marks preceding and following an argument component in its covering sentence and a Boolean feature that indicates if the sentence closes with a question mark.

Lexical features: We define n-grams, verbs, adverbs and modals as lexical features. We consider all n-grams of length 1-3 as a Boolean feature and extract them from the argument component including preceding tokens in the sentence that are not covered by another argument component. So, the n-gram features include discourse markers that indicate certain argument components but which are not included in the actual annotation of argument components.

Verbs and adverbs play an important role for identifying argument components. For instance, certain verbs like *'believe'*, *'think'* or *'agree'* often signal stance expressions which indicate the presence of a major claim and adverbs like *'also'*, *'often'* or *'really'* emphasize the importance of a premise. We model both verbs and adverbs as Boolean features.

Modal verbs like *'should'* and *'could'* are frequently used in argumentative discourse to signal the degree of certainty when expressing a claim. We use the POS tags generated during preprocessing to identify modals and define a Boolean feature which indicates if an argument component contains a modal verb.

Syntactic features: To capture syntactic properties of argument components, we define features extracted from parse trees. We adopt two features proposed by (Mochales-Palau and Moens, 2009): the number of sub-clauses included in the covering sentence and the depth of the parse tree. In addition, we extract *production rules* from the parse tree as proposed by Lin et al. (2009) to capture syntactic characteristics of an argument component. The production rules are collected for each function tag (e.g. VP, NN, S, etc.) in the subtree of an argument component. The feature set includes e.g. rules like $VP \rightarrow VBG, NP$ or $PP \rightarrow IN, NP$. We model each production rule as a Boolean feature and set it to true if it appears in the subtree of an argument component.

Since premises often refer to previous events and claims are usually in present tense, we capture the tense of the main verb of an argument component as proposed by Mochales-Palau and Moens (2009) and define a feature that indicates if an argument component is in the past or present tense.

Indicators: Discourse markers often indicate the components of an argument. For example, claims are frequently introduced with *'therefore'*, *'thus'* or *'consequently'*, whereas premises contain markers like *'because'*, *'reason'* or *'furthermore'*. We collected a list of discourse markers from the Penn Discourse Treebank 2.0 Annotation Manual (Prasad et al., 2007) and removed markers that do not indicate argumentative discourse (e.g. markers which indicate temporal discourse). In total, we collected 55 discourse markers and model each as a Boolean feature set to true if the particular marker precedes the argumentative component.

In addition, we define five Boolean features which denote a reference to the first person in the covering sentence of an argument component: *'I'*, *'me'*, *'my'*, *'mine'*, and *'myself'*. An additional Boolean feature indicates if one of them is present in the covering sentence. We expect that those features are good indicators of the major claim, since it is often introduced with expressions referring to the personal stance of the author.

Contextual features: The context plays a major role for identifying argument components. For instance, a premise can only be classified as such, if there is a corresponding claim. Therefore, we define the following features each extracted from the sentence preceding and following the covering sentence of an argument component: the number of punctuations, the number of tokens, the number of sub-clauses and a Boolean feature indicating the presence of modal verbs.

4.2 Results and Analysis

For identifying the best performing system, we conducted several experiments on the training set using stratified 10-fold cross-validation. We determine the evaluation scores by accumulating the confusion matrices of each fold into one confusion matrix, since it is the less biased method for evaluating cross-validation studies (Forman and Scholz, 2010). In a comparison of several classifiers (Support Vector Machine, Naïve Bayes, C4.5 Decision Tree and Random Forest), we found that each of the classifiers significantly outperforms a majority baseline (McNemar Test (McNemar, 1947) with $p = 0.05$) and that a Support Vector Machine (SVM) achieves the best results using 100 top features ranked by Information Gain.[3] It achieves an accuracy of 77.3% on the test set and outperforms the majority baseline with respect to overall accuracy as well as F1-score (table 2).

	Baseline	**Human**	**SVM**
Accuracy	0.55	0.877	0.773
Macro F1	0.177	0.871	0.726
Macro Precision	0.137	0.864	0.773
Macro Recall	0.25	0.879	0.684
F1 MajorClaim	0	0.916	0.625
F1 Claim	0	0.841	0.538
F1 Premise	0.709	0.911	0.826
F1 None	0	0.812	0.884

Table 2: Results of an SVM for argument component classification on the test set compared to a majority baseline and human performance.

The upper bound for this task constitutes the human performance which we determine by comparing each annotator to the gold standard. Since the boundaries of an argument component in the gold standard can differ from the boundaries identified by a human annotator (the annotation task included the identification of argument component boundaries), we label each argument component of the gold standard with the class of the maximum overlapping annotation of a human annotator for determining the human performance. We obtain a challenging upper bound of 87.7% (accuracy) by averaging the scores of all three annotators on the test set (table 2). So, our system achieves 88.1% of human performance (accuracy).

Feature influence: In subsequent experiments, we evaluate each of the defined feature groups on the entire data set using 10-fold cross-validation to find out which features perform best for identifying argument components. As assumed, structural features perform well for distinguishing claims and premises in persuasive essays. They also yield high results for separating argumentative from non-argumentative text units (table 3).

Feature group	**MajorClaim**	**Claim**	**Premise**	**None**
Structural	0.477	0.419	0.781	0.897
Lexical	0.317	0.401	0.753	0.275
Syntactic	0.094	0.292	0.654	0.427
Indicators	0.286	0.265	0.730	0
Contextual	0	0	0.709	0

Table 3: F1-scores for individual feature groups and classes (SVM with 10-fold cross-validation on the entire data set)

Interestingly, the defined indicators are not useful for separating argumentative from non-argumentative text units though they are helpful for classifying argument components. A reason for this could be that not each occurrence of an indicator distinctly signals argument components, since their sense is often ambiguous (Prasad et al., 2008). For example *'since'* indicates temporal properties as well as justifications, whereas *'because'* also indicates causal links. Syntactic features also contribute to the identification of argument components. They achieve an F1-score of 0.292 for claims and 0.654 for premises and also contribute to the separation of argumentative from non-argumentative text units. Contextual features do not perform well. However, they increase the accuracy by 0.7% in combination with other features. Nevertheless, this difference is not significant ($p = 0.05$).

Error analysis: The system performs well for separating argumentative and non-argumentative text units as well as for identifying premises. However, the identification of claims and major claims yields lower performance. The confusion matrix (table 4) reveals that the most common error is between claims and premises. In total, 193 claims are incorrectly classified as premise. In a manual assessment, we observed that many of these errors occur if the claim is present in the first paragraph sentence and exhibits preceding indicators like *'first(ly)'* or *'second(ly)'* which are also frequently used to enumerate premises. In these cases, the author introduces the claim of the argument as support for the major claim and thus its characteristic is similar to a premise. To prevent

[3] Although the Naïve Bayes classifier achieves lowest accuracy, it exhibits a slightly higher recall compared to SVM.

this type of error, it might help to define features representing the location of indicators or to disambiguate the function of indicators.

Actual \ *Predicted*	MC	Cl	Pr	No
MC	**38**	34	18	0
Cl	19	**210**	193	7
Pr	6	104	**904**	19
No	0	12	23	**292**

Table 4: Confusion matrix (SVM) for argument component classification (MC = Major Claim; Cl = Claim; Pr = Premise; No = None)

We also observed, that some of the misclassified claims cover an entire sentence and don't include indicators. For example, it is even difficult for humans to classify the sentences *'Competition helps in improvement and evolution'* as a claim without knowing the intention of the author. For preventing these errors, it might help to include more sophisticated contextual features.

5 Identifying Argumentative Relations

We consider the identification of argumentative relations as a binary classification task of argument component pairs and classify each pair as either support or non-support. For identifying argumentative relations, all possible combinations of argument components have to be tested. Since this results in a heavily skewed class distribution, we extract all possible combinations of argument components from each paragraph of an essay.[4] So, we omit argumentative relations between claims and major claims which are the only relations in the corpus that cross paragraph boundaries, but obtain a better distribution between true (support) and false (non-support) instances. In total, we obtain 6,330 pairs, of which 15.6% are support and 84.4% are non-support relations (table 5).

Support	Non-support
989 (15.6%)	5341 (84.4%)

Table 5: Class distribution of argument component pairs

Equivalent to the identification of argument components, we randomly split the data in a 80% training and a 20% test set and determine the best performing system using 10-fold cross-validation on the training set. We use the same preprocessing pipeline as described in section 4 and DKPro-TC for extracting the features described below.

5.1 Features

Structural features: We define structural features for each pair based on the source and target components, and on the mutual information of both. Three numeric features are based on token statistics. Two features represent the number of tokens of the source and target components and the third one represents the absolute difference in the number of tokens. Three additional numeric features count the number of punctuation marks of the source and target components as well as the absolute difference between both. We extract both types of features solely from the clause annotated as argument component and do not consider the covering sentence. In addition, we define nine structural features based on the position of both argument components: two of them represent the position of the covering sentences in the essay, four Boolean features indicate if the argument components are present in the first or last sentence of a paragraph, one Boolean feature for representing if the target component occurs before the source component, the sentence distance between the covering sentences, and a Boolean feature which indicates if both argument components are in the same sentence.

Lexical features: We define lexical features based on word pairs, first words and modals. It has been shown in previous work that word pairs are effective for identifying implicit discourse relations (Marcu and Echihabi, 2002). We define each pair of words between the source and target components as a Boolean feature and investigate word pairs containing stop words as well as stop word filtered word pairs.

In addition, we adopt the first word features proposed by Pitler et al. (2009). We extract the first word either from the argument component or from non-annotated tokens preceding the argument component in the covering sentence if present. So, the first word of an argument component is either the first word of the sentence containing the argument component, the first word following a preceding argument component in the same sentence or the first word of the actual argument component if it commences the sentence or directly follows another argument component.

[4] Only 4.6% of 28,434 possible pairs are true instances (support), if all combinations are considered.

So, we ensure that the first word of an argument component includes important discourse markers which are not included in the annotation. We define each first word of the source and target components as a Boolean feature and also add the pairs of first words to our feature set.

Further, we define a Boolean feature for the source as well as for the target component that indicates if they contain a modal verb and a numerical feature that counts the number of common terms of the two argument components.

Syntactic features: For capturing syntactic properties, we extract production rules from the source and target components. Equivalent to the features extracted for the argument component classification (section 4.1), we model each rule as a Boolean feature which is true if the corresponding argument component includes the rule.

Indicators: We use the same list of discourse markers introduced above (section 4.1) as indicator features. For each indicator we define a Boolean feature for the source as well as for the target component of the pair and set it to true if it is present in the argument component or in its preceding tokens.

Predicted type: The *argumentative type* (major claim, claim or premise) of the source and target components is a strong indicator for identifying argumentative relations. For example, there are no argumentative relations from claims to premises. Thus, if the type of the argument component is reliably identified many potential pairs can be excluded. Therefore, we define two features that represent the argumentative type of the source and target components identified in the first experiment.

5.2 Results and Analysis

The comparison of several classifiers reveals that an SVM achieves the best results. In our experiments, all classifiers except the C4.5 Decision Tree significantly outperform a majority baseline which classifies all pairs as non-support ($p = 0.05$). We also conducted several experiments using word pair features only and found in contrast to Pitler et al. (2009) that limiting the number of word pairs decreases the performance. In particular, we compared the top 100, 250, 500, 1000, 2500, 5000 word pairs ranked by Information Gain, non-zero Information Gain word pairs and non-filtered word pairs. The results show that non-filtered word pairs perform best (macro F1-score of 0.68). Our experiments also reveal that filtering stop words containing word pairs decreases the macro F1-score to 0.60. We obtain the best results using an SVM without any feature selection method. Due to the class imbalance, the SVM only slightly outperforms the accuracy of a majority baseline on the test set (table 6). However, the macro F1-score is more appropriate for evaluating the performance if the data is imbalanced since it assigns equal weight to the classes and not to the instances. The SVM achieves a macro F1-score of 0.722 and also outperforms the baseline with respect to the majority class.

	Baseline	**Human**	**SVM**
Accuracy	0.843	0.954	0.863
Macro F1	0.458	0.908	0.722
Macro Precision	0.422	0.937	0.739
Macro Recall	0.5	0.881	0.705
F1 Support	0	0.838	0.519
F1 Non-Support	0.915	0.973	0.92

Table 6: Results of an SVM for classifying argumentative relations on the test set compared to a majority baseline and human performance.

We determined the upper bound constituted by the human performance by comparing the annotations of all three annotators to the gold standard. The scores in table 6 are the average scores of all three annotators. Our system achieves 90.5% of human performance (accuracy).

Feature influence: A comparison of the defined feature groups using 10-fold cross-validation on the entire data set shows that lexical features perform best. They achieve an F1-score of 0.427 for support and 0.911 for non-support pairs (table 7). The syntactic features also perform well followed by the indicators. It turned out that structural features are not effective for identifying argumentative relations though they are the most effective features for identifying argument components (cp. section 4.2). However, when omitted from the entire feature set the performance significantly decreases by 0.018 macro F1-score ($p = 0.05$).

Interestingly, the predicted types from our first experiment are not effective at all. Although the argumentative type of the target component exhibits the highest Information Gain in each fold compared to all other features, the predicted type does not yield a significant difference when combined with all other features ($p = 0.05$). It only improves the macro F1-score by 0.001 when in-

cluded in the entire feature set.

Feature group	Support	Non-Support
Structural	0	0.915
Lexical	0.427	0.911
Syntactic	0.305	0.911
Indicators	0.159	0.916
Predicted types	0	0.915

Table 7: F1-scores for individual feature groups using an SVM and the entire data set

Error analysis: For identifying frequent error patterns, we manually investigated the mistakes of the classifier. Although our system identifies 97.5% of the non-support pairs from claim to premise correctly, there are still some false positives that could be prevented if the argument components had been classified more accurately. For instance, there are 18 non-support relations from claim to another claim, 32 from claim to premise, 5 from major claim to premise and 4 from major claim to claim among the false positives. However, the larger amount of errors is due to not identified support relations (false negatives). We found that some errors might be related to missing contextual information and unresolved coreferences. For instance, it might help to replace *'It'* with *'Exercising'* for classifying the pair *'It helps relieve tension and stress'* $\rightarrow$ *'Exercising improves self-esteem and confidence'* as support relation or to include contextual information for the premise *'This can have detrimental effects on health'* supporting the claim *'There are some serious problems springing from modern technology'*.

6 Discussion

In our experiments, we have investigated the classification of argument components as well as the identification of argumentative relations for recognizing argumentative discourse structures in persuasive essays. Both tasks are closely related and we assume that sharing mutual information between both tasks might be a promising direction for future research. On the one hand, knowing the type of argument components is a strong indicator for identifying argumentative relations and on the other hand, it is likely that information about the argumentative structure facilitates the identification of argument components. However, our experiments revealed that the current accuracy for identifying argument components is not sufficient for increasing the performance of argumentative relation identification. Nevertheless, we obtain almost human performance when including the types of argument components of the gold standard (macro F1-score >0.85) in our argument relation identification experiment and when including the number of incoming and outgoing support relations for each argument component in our first experiment (macro F1-score >0.9). Therefore, it can be assumed, that if the identification of argument components can be improved, the identification of argumentative relations will achieve better results and vice versa.

The results also show that the distinction between claims and premises is the major challenge for identifying argument components. It turned out that structural features are the most effective ones for this task. However, some of those features are unique to persuasive essays, and it is an open question if there are general structural properties of arguments which can be exploited for separating claims from premises.

Our experiments show that discourse markers yield only low accuracies. Using only our defined indicator features, we obtain an F1-score of 0.265 for identifying claims, whereas Mochales-Palau and Moens (2009) achieve 0.673 for the same task in legal documents using a CFG. This confirms our initial assumption that approaches relying on discourse markers are not applicable for identifying argumentative discourse structures in documents which do not follow a standardized form. In addition, it shows that discourse markers are either frequently missing or misleadingly used in student texts and that there is a need for argumentative writing support systems that assist students in employing discourse markers correctly.

7 Conclusion and Future Work

We presented a novel approach for identifying argumentative discourse structures in persuasive essays. Previous approaches on argument recognition suffer from several limitations: Existing approaches focus either solely on the identification of argument components or rely on manually created rules which are not able to identify implicit argumentative discourse structures. Our approach is the first step towards computational argument analysis in the educational domain and enables the identification of implicit argumentative discourse structures. The presented approach achieves 88.1% of human performance for identi-

fying argument components and 90.5% for identifying argumentative relations.

For future work, we plan to extend our studies to larger corpora, to integrate our classifiers in writing environments, and to investigate their effectiveness for supporting students.

Acknowledgements

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

References

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 05(04):363–381.

M. Anne Britt and Aaron A. Larson. 2003. Constructing representations of arguments. *Journal of Memory and Language*, 48(4):794 – 810.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pages 1–10, Aalborg, Denmark.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A Java-based framework for supervised learning experiments on textual data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 61–66, Baltimore, MD, USA.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 987–996, Portland, OR, USA.

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, Sofia, Bulgaria. ok.

George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, 12(1):49–57.

Iryna Gurevych, Max Mühlhäuser, Christof Mueller, Juergen Steimle, Markus Weimer, and Torsten Zesch. 2007. Darmstadt Knowledge Processing Repository based on UIMA. In *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology, Tuebingen, Germany*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Sapporo, Japan.

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

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, EMNLP '09, pages 343–351, Stroudsburg, PA, USA.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 59–62, Stroudsburg, PA, USA.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, Information Sciences Institute.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 368–375.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

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 '09, pages 98–107, New York, NY, USA. ACM.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, Stanford, California.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. 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*, ACL '09, pages 683–691, Suntec, Singapore. Association for Computational Linguistics.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L. Webber. 2007. The Penn Discourse Treebank 2.0 annotation manual. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

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

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

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

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, page (to appear), Dublin, Ireland, August.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pages 173–180, Edmonton, Canada.

Douglas N Walton. 1996. *Argumentation schemes for presumptive reasoning*. Routledge.

Adam Wyner, Raquel Mochales Palau, Marie-Francine Moens, and David Milward. 2010. Approaches to text mining arguments from legal cases. In *Semantic Processing of Legal Texts*, volume 6036 of *Lecture Notes in Computer Science*, pages 60–79.