Parsing Argumentation Structures in Persuasive Essays

Christian Stab*
Technische Universität Darmstadt

Iryna Gurevych**
Technische Universität Darmstadt and
German Institute for Educational
Research

In this article, we present the first end-to-end approach for parsing argumentation structures in persuasive essays. We model the argumentation structure as a tree including several types of argument components connected with argumentative support and attack relations. We consider the identification of argumentation structures in several consecutive steps. First, we segment a persuasive essay in order to identify relevant argument components. Second, we jointly model the classification of argument components and the identification of argumentative relations using Integer Linear Programming. Third, we recognize the stance of each argument component in order to discriminate between argumentative support and attack relations. By evaluating the joint model using two corpora, we show that our approach not only considerably improves the identification of argument component types and argumentative relations but also significantly outperforms a challenging heuristic baseline. In addition, we introduce a novel corpus including 402 *persuasive essays annotated with argumentation structures and show that our new annotation guideline successfully guides annotators to substantial agreement.*

1. Introduction

Argumentative practices are omnipresent in our daily verbal communication and thinking. We engage argumentation in order to infer certainty, to obtain widely accepted conclusions or to persuade a particular audience. In addition, argumentation is crucial to learning itself and thus constitutes a key objective of current education programs (Davies 2009, p. 94). One pedagogical method to teach argumentation skills is *persuasive essay writing*. Although, it is widely-used and considered to be a powerful tool for teaching students to construct well-reasoned arguments (Botley 2014, p. 46), a great many of students are still underprepared in developing good arguments (Butler and Britt 2011, p. 70) (Wolfe and Britt 2009, p. 183). One reason for this deficit is that teachers are not able to provide sufficient writing assignments in view of growing class sizes and the enormous load for providing adequate feedback (Burstein, Chodorow, and Leacock 2004, p. 27).

In order to relieve the load of teachers, *Automated Essay Scoring* (AES) aims at automatically grading writing assignments by exploiting state-of-the-art *Natural Language Processing* (NLP) methods (Shermis and Burstein 2013). Although, current methods

* Ubiquitous Knowledge Processing Lab (UKP-TUDA), Department of Computer Science, Technische Universität Darmstadt
** Ubiquitous Knowledge Processing Lab (UKP-TUDA), Department of Computer Science, Technische Universität Darmstadt and Ubiquitous Knowledge Processing Lab (UKP-DIPF), German Institute for Educational Research

like *e-rater* imitate human grading with a correlation of up to .97 (Attali and Burstein 2006, p. 22), the holistic scores produced by these systems do not adequately guide students to improve specific writing skills. Therefore, *intelligent writing assistance* aims at providing *formative feedback* intended to highlight particular weaknesses and providing instructions for improving underdeveloped writing skills (Shute 2008). For instance, systems like *Criterion Online Essay Evaluation Service* (Burstein, Chodorow, and Leacock 2004) provide feedback about grammar, spelling, style and organization of the text. However, there is no system that provides feedback about written arguments in order to foster students' argumentation abilities and to improve the quality of written arguments respectively.

In contrast to existing work in *Computer-Supported Argumentation* which focuses on visualization, interaction and collaboration tools (Scheuer et al. 2010), our objective is to develop an *Argumentative Writing Support System* which analyzes a given persuasive essay in order to provide tailored feedback about written arguments. Accordingly, methods that automatically recognize argumentation structures in natural language texts represent a key challenge for implementing such a system. Besides the identification of argument components and their argumentative type (e.g. claim or premise), this task also includes the identification of argumentative relations between argument components which is a major prerequisite for analyzing arguments (Henkemans 2000, p. 448) (Govier 2010, p. 22) (Sampson and Clark 2006).

Argumentation Mining is a recent and rapidly growing field in NLP. Originating from early approaches in the legal domain (Mochales-Palau and Moens 2009), there are currently various approaches focusing on particular subtasks like separating argumentative from non-argumentative text units (Moens et al. 2007; Florou et al. 2013; Levy et al. 2014; Lippi and Torroni 2015), classifying argument components (Kwon et al. 2007; Rooney, Wang, and Browne 2012; Mochales-Palau and Moens 2011; Stab and Gurevych 2014b; Lippi and Torroni 2015), or recognizing argumentative relations either between complete arguments (Cabrio and Villata 2012b; Ghosh et al. 2014; Boltužić and Šnajder 2014) or between argument components (Mochales-Palau and Moens 2009; Peldszus 2014; Stab and Gurevych 2014b; Peldszus and Stede 2015). However, an approach covering all subtasks required to identify fine-grained argumentation structures is still missing. Existing approaches focusing on argumentation structure parsing are either limited in granularity and are not capable of recognizing implicit argumentative relations (Mochales-Palau and Moens 2009), or omit required subtasks like segmenting argument components (Peldszus 2014; Peldszus and Stede 2015; Stab and Gurevych 2014b).

Besides the lack of end-to-end approaches for parsing argumentation structures, there are only few corpora available that include annotations of fine-grained argumentation structures. Most of the existing resources for Argumentation Mining focus only on particular subtasks or address information retrieval tasks over multiple documents (Levy et al. 2014) instead of argumentation structures at the document-level (cp. section 3.1.1). Apart from our previously created corpus including 90 persuasive essays (Stab and Gurevych 2014a), the few resources annotated with document-level argumentation structures lack non-argumentative text units and are fairly small (Peldszus 2014), include text genres different from our target domain (Kirschner, Eckle-Kohler, and Gurevych 2015), or the reliability of the annotations is unknown (Reed et al. 2008).

In this article, we present the first end-to-end approach including all required subtasks to identify fine-grained argumentation structures in persuasive essays. In addition, we introduce a novel corpus of persuasive essays which represents to the

best of our knowledge the largest resource annotated with argumentation structures. In particular the contributions of this article are the following:

- *An annotation scheme for modeling argumentation structures* derived from argumentation theory. In contrast to existing work in Argumentation Mining, our scheme enables to model the entire argumentation structure of a document as a connected tree instead of isolated single arguments or certain aspects of argumentation.
- *A novel corpus of persuasive essays* including 402 documents annotated with argumentation structures. Compared to our previous work (Stab and Gurevych 2014a), this corpus is considerably larger and annotated based on a revised version of our previous annotation guideline.[1] We show that the annotation scheme as well as our novel annotation guideline lead to substantial agreement in an annotation study with three annotators.
- *An approach for parsing argumentation structures* based on supervised machine learning and joint modeling. Whereas our previous work (Stab and Gurevych 2014b) addressed only the classification of argument components and the identification of argumentative support relations, our novel approach covers all required steps to extract the entire argumentation structure of persuasive essays. In particular, it includes the segmentation of argument components, a joint modeling approach for argument component types and argumentative relations, and the discrimination of argumentative support and attack relations beyond our previous work.

The remainder of this article is structured as follows. In section 2, we provide a brief theoretical background of argumentation, derive our annotation scheme from argumentation theory and introduce the challenges and requirements for automatically identifying argumentation structures. In section 3, we give an overview of related work, highlight the additional value of our work compared to existing Argumentation Mining approaches and discuss the difference between Argumentation Mining and existing approaches in discourse analysis. Section 4 presents the findings of an annotation study and details of the corpus creation. We show that our proposed annotation scheme as well as the revised guideline lead to substantial agreement among three annotators. In section 5, we present our approach for parsing argumentation structures including the description of our novel feature sets, experimental results, comparison to human upper bounds, and error analyses for each model in our pipeline. We show that jointly modeling argument component types and argumentative relations simultaneously improves both tasks on two independent corpora and that our joint modeling approach significantly outperforms a challenging heuristic baseline. We conclude this article with a discussion in section 6.

2. Argumentation: Overview and Background

Argumentation is a verbal activity of reason which aims at increasing or decreasing the acceptability of a controversial standpoint (van Eemeren, Grootendorst, and

1 The differences will be discussed in section 2.2

Snoeck Henkemans 1996, p. 5). Each *argument* involved in this process consists of several *argument components*. It includes a claim and one or more premises. The *claim* is a controversial statement and the central component of an argument.[2] The *premises* constitute the reasons for believing the claim to be true or false (Damer 2009, p. 14).

Traditionally, there are two different types of arguments: deductive arguments and inductive arguments (Copi and Cohen 1990, p. 44-47). In *deductive arguments* the truth of the claim follows necessarily from its premises and it is impossible that the premises are true and the claim is false. In contrast, in *inductive arguments* the truth of the claim is not a logical consequence of its premises. The premises provide reasons for decreasing or increasing the acceptability of the claim but no absolute logical proof. Accordingly, there are two general perspectives on the study of argumentation.

The main objective of *traditional logic approaches* is to formalize the relation between premises and claims in deductive arguments (Copi and Cohen 1990, p. 46). In particular, traditional logic approaches focus on mathematical formalisms to distinguish between deductively valid and invalid arguments (van Eemeren, Grootendorst, and Snoeck Henkemans 1996, chapter 1.2). In contrast, our aim is to define an argumentation scheme to model argumentation structures in persuasive essays. Since a great many of ordinary arguments are inductive and do not follow deductively valid reasoning (Damer 2009, p. 22) (Copi and Cohen 1990, p. 357), traditional logic approaches are not sufficient for this task (Woods and Walton 2007) and we will not consider them further in this work.

In order to analyze, structure and evaluate inductive arguments, the field of *Informal Logic* evolved in the early 1950^{th}. Pioneered by the influential work of Toulmin (1958), there are various approaches covering different aspects of argumentation. In general, Informal Logic approaches focus on arguments as *product* whereas dialogue logic focuses on the *process* of argumentation (Johnson 2000). Product oriented approaches address the internal structure of arguments (*micro-level*). Their purpose is to define different types of argument components and the type of reasoning. For instance, Toulmin's model includes six argument components (Toulmin 1958, p. 97) whereas *argumentation schemes* proposed by Hastings (1963) and extensively studied by Walton, Reed, and Macagno (2008) capture the reasoning type of single arguments. In contrast, process oriented approaches focus on the formalization of conversations such as discussions, debates or negotiations and thus relations between arguments (*macro-level*). Examples of process oriented approaches are MacKenzie's model of formal dialectics (MacKenzie 1981), the acceptability framework proposed by Dung (1995) or Amgoud's argumentative dialog modeling framework (Amgoud, Maudet, and Parsons 2000). However, both perspectives are closely related since the formulation of a single argument is subject to the process of arguing and dialog situations presuppose arguments as a product (Bentahar, Moulin, and Bélanger 2010; Reed and Walton 2003; Walton and Godden 2007). Accordingly, more recent approaches aim at combining both perspectives and consider argumentation as a hypothetical dialectical exchange between a proponent and an opponent in order to develop a holistic theory of argumentation (van Eemeren and Grootendorst 2004; Freeman 2011; Peldszus and Stede 2013a). Transferred to our objective, this consideration allows to model the entire argumentation structure of natural language documents instead of extracting single arguments and thus it enables to combine the micro- and macro-level perspectives.

2 The claim is also called conclusion in related literature.

2.1 Argumentation Structures

Laying out the structure of arguments is a widely used method in Informal Logic (Copi and Cohen 1990, p. 18-45) (Govier 2010, p. 22-56). This technique, referred to as *argument diagramming*, aims at transferring natural language arguments into a structured representation in order to evaluate everyday arguments in subsequent analysis steps (Henkemans 2000, p. 447). Although argumentation theorists usually define argument diagramming as a manual activity, the diagramming conventions are a good foundation for designing Argumentation Mining systems (Peldszus and Stede 2013a).

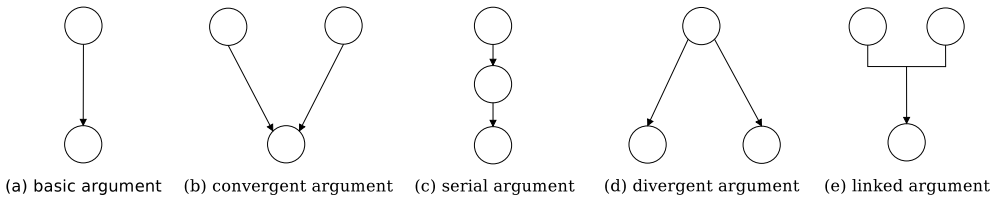


Figure 1
General argument structures at the micro-level proposed by argumentation theorists. Nodes indicate argument components, arrows mark argumentative relations and nodes at the bottom are the claims of individual arguments (Note that it is a common convention in argumentation theory to visualize argument structures upside down).

An argument diagram is a node-link diagram in which each node represents an argument component (a statement represented in natural language) and each link represents a directed argumentative relation (e.g. a support relation indicating that the source component is a premise given for justifying the target component) (Reed, Walton, and Macagno 2007, p. 93). In addition, to (a) *basic arguments* including one claim supported by a single premise, Beardsley (1950) introduced three different types of arguments: (b) *convergent arguments*, (c) *serial arguments*, and (d) *divergent arguments*. In a convergent argument, two independent premises support the claim; an argument is serial if there is a reasoning chain and divergent if a single premise supports several claims. Complementary, Thomas (1973) introduced (e) *linked arguments* including two premises which only support the claim in conjunction. More complex argumentation structures can combine any combination of these basic structures illustrated in figure 1. In addition, Peldszus and Stede (2013a) propose to include argumentative attack relations complementary to support relations to model opposing premises.

However, on closer inspection there are several issues which need to be considered in order to derive a model for computational purposes. First, applying those structures to real texts bears several ambiguities. In particular, the distinction between convergent and linked structures frequently causes problems when analyzing real argumentation structures (Henkemans 2000, p. 448). Second, it is not clear if the argumentation structure is a graph or a tree. Third, there is disagreement about how to label the argument components in more complex argumentation structures. In order to derive a computational model for representing the argumentation structure of persuasive essays, we will discuss each of these questions in the following sections.

2.1.1 Distinguishing between Linked and Convergent Structures. The first decision is wether the argumentation model needs to distinguish between linked and convergent arguments which is still an ongoing debate in argumentation theory (van Eemeren, Grootendorst, and Snoeck Henkemans 1996, p. 176) (Freeman 2011, chapter 4, p. 89)

(Yanal 1991) (Conway 1991). From a traditional logic perspective of argumentation theory, linked structures indicate deductive reasoning and convergent structures represent inductive reasoning (Henkemans 2000, p. 453). Although, this distinction is theoretically appropriate, Freeman (2011, p. 91ff.) illustrates that the traditional definition of linked structures causes ambiguities when analyzing real arguments and suggests a more precise definition taking the relevance of each premise into account. In addition, Yanal (1991) argues that the distinction is equivalent to separating several arguments and Conway (1991) argues that it can be safely omitted when modeling single arguments. From computational perspective, the task of distinguishing between linked and convergent structures is similar to finding groups among premises which belong to the same claim or to classifying the reasoning type of an argument as deductive or inductive. Accordingly, there is no need to consider linked structures in order to recognize arguments in natural language texts, since this task can be considered in subsequent analysis steps which e.g. aim at evaluating argumentation structures.

2.1.2 Argumentation Structures as Tree. Defining the argumentation model as a tree structure is a matter of excluding divergent structures, restricting each premise to support only one particular argument component and omitting circular structures. According to (Freeman 2011, p. 16) divergent structures are equivalent to several arguments (one for each claim). As a result of this treatment, a great many of textbooks neglect divergent structures (Henkemans 2000, p. 447) (Reed and Rowe 2004, p. 972). Although, most Argumentation Mining approaches consider argumentation structures as trees (Mochales-Palau and Moens 2009; Cohen 1987) or allow only one outgoing relation per argument component (Peldszus 2014), we argue that this decision requires a careful investigation of the particular text genre. In particular, modeling argumentation structures as trees might potentially limit the expressiveness of the approach if a particular genre includes large amounts of divergent or circular structures. Usually persuasive essays exhibit a common structure. According to various textbooks about essay writing (Whitaker 2009; Shiach 2009; Perutz 2010; Kemper and Sebranek 2004), the writing process follows a linear procedure. Starting with the formulation of a *thesis statement* included in the opening paragraph and restated in the closing paragraph, each body paragraph includes a single point expressed in a *topic sentence*. The remaining sentences in each body paragraph either increase or decrease the acceptability of the topic sentence. Therefore, it is very unlikely that persuasive essays include divergent or circular structures and we assume that modeling the argumentation structure of persuasive essays as a tree is a reasonable decision. Furthermore, an empirical study of argumentation structures in political speech (which can be generally assumed to exhibit complex argumentation structures) shows that only 5.26% of the arguments are divergent (Indrajani and Angeline 2010).

2.1.3 Argumentation Structures and Argument Component Types. Assigning an argumentative type to argument components is unambiguous if the argumentation structure is shallow. For instance, it is obvious that an argument component a_1 is a premise and argument component a_2 is a claim, if a_1 supports a_2 in a basic structure. However, if the tree structure is deeper, assigning argumentative types becomes ambiguous. Basically there are three different approaches of assigning argumentative types to argument components. First, according to Beardsley (1950) a serial argumentation structure includes an argument component which is both a claim and a premise for a further claim. Therefore, the inner argument component bears two different argumentative types (*multi-label approach*). Second, Govier (2010, p. 24) distinguishes between subarguments

and whole argument, and differentiates between main claim and subclaim. Similarly, Damer (2009, p. 17) distinguishes between premise and subpremise in order to label serial structures. So these approaches define particular labels for each level in the argumentation structure (*level approach*). Third, Cohen (1987) considers only the root node of an argumentation tree as a claim and the following nodes in the structure as evidence (*one-claim approach*). In order to define an argumentation model for persuasive essays, we propose a *hybrid approach* that combines the level and the one-claim approaches.

2.2 Argumentation Structures in Persuasive Essays

Persuasive essays exhibit a common structure and usually include between four and six paragraphs (Botley 2014). The opening paragraph of an essay introduces the controversial topic and usually includes a *major claim* which expresses the stance of the author according to the topic (Whitaker 2009, p. 7). The following example illustrates an opening paragraph of an essay about cloning:[3]

> *Since researchers at the Roslin Institute in Edinburgh cloned an adult sheep, there is an ongoing debate if cloning technology is morally and ethically right or not. Some people argue for and others against and there is still no agreement whether cloning technology should be permitted. However, as far as I'm concerned, [**cloning is an important technology for humankind**]$_{MajorClaim1}$ since [it would be very useful for developing novel cures]$_{Claim1}$.*

The first two sentences introduce the topic and do not include any argumentative content. The third sentence includes the major claim (boldfaced) and a claim which supports the major claim (underlined). The following body paragraphs of the essay include arguments which either support or attack the major claim and the stance of the author respectively. For instance, the following body paragraph includes one argument, supporting the positive stance of the author on cloning:

> *First, [cloning will be beneficial for many people who are in need of organ transplants]$_{Claim2}$. [Cloned organs will match perfectly to the blood group and tissue of patients]$_{Premise1}$ since [they can be raised from cloned stem cells of the patient]$_{Premise2}$. In addition, [it shortens the healing process]$_{Premise3}$. Usually, [it is very rare to find an appropriate organ donor]$_{Premise4}$ and [by using cloning in order to raise required organs the waiting time can be shortened tremendously]$_{Premise5}$.*

The first sentence contains the claim of the argument which is supported by five premises in the remaining three sentences of the paragraph (wavy underlined). The second sentence includes two premises of which premise$_1$ supports claim$_2$ and premises$_2$ supports premise$_1$. The third sentence includes premise$_3$ which supports claim$_2$ and the fourth sentence includes premise$_4$ and premise$_5$ both supporting premise$_3$. The next paragraph illustrates an example including two arguments:

> *Second, [scientists use animals as models in order to learn about human diseases]$_{Premise6}$ and therefore [cloning animals enables novel developments in science]$_{Claim3}$. Furthermore, [parents with no eggs or sperms can have children that are genetically related]$_{Premise7}$. [Even same sex couples can have children without the use of donor*

3 Note that the example essay was written by the authors to illustrate all phenomena of argumentation structures in persuasive essays.

sperm or donor eggs]$_{Premise8}$. *Consequently, [cloning can help human families to get children]*$_{Claim4}$.

The first sentence includes the first argument starting with premise$_6$ followed by claim$_3$. The next two sentences include a premise which supports another claim in the last sentence. Note that both arguments cover different aspects (development in science and cloning humans) which both support the positive stance of the author on cloning. This example also illustrates that knowing argumentative relations is important for separating several arguments in a paragraph. The third body paragraph illustrates a contra argument and argumentative attack relations:

Admittedly, [cloning could be misused for military purposes]$_{Claim5}$. *For example, [it could be used to manipulate human genes in order to create obedient soldiers with extraordinary abilities]*$_{Premise9}$. *However, because [moral and ethical values are internationally shared]*$_{Premise10}$, *[it is very unlikely that cloning will be misused for militant objectives]*$_{Premise11}$.

The paragraph begins with a claim against the stance of the author which is supported by premise$_9$ in the second sentence. The third sentence includes two premises aimed at defending the stance of the author. Premise$_{11}$ is an attack of claim$_5$ which is supported by the preceding premise$_{10}$. Finally, the last paragraph restates the major claim and summarizes the main aspects of the essay:

To sum up, although [permitting cloning might bear some risks like misuse for military purposes]$_{Claim6}$, *I strongly believe that [**this technology is beneficial for humanity**]*$_{MajorClaim2}$. *It is likely that [this technology bears some important cures which will significantly improve life conditions]*$_{Claim7}$.

It starts with an attacking claim followed by the repetition of the major claim in the first sentence. The last sentence includes another claim restating the most important reason of the authors' argumentation. Figure 2 illustrates the entire argumentation of this example essay.

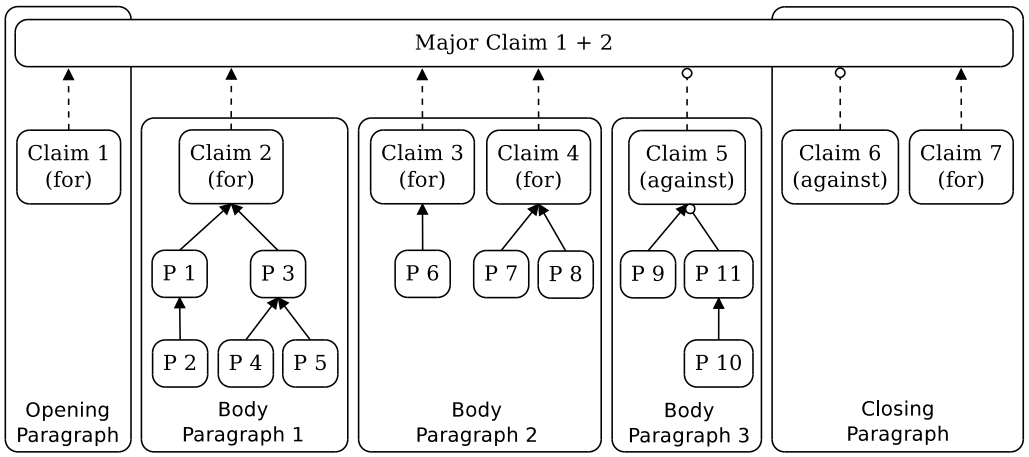


Figure 2
Entire argumentation structure of the example essay. Arrows indicate argumentative relations. Arrowheads denote argumentative support relations and circleheads argumentative attack relations. Dashed arrows are argumentative relations connecting claims with the major claim(s) which are encoded by means of the stance attribute of each claim.

We model the argumentation structure of persuasive essays as a tree. In order to model the first level of the tree, we employ a level approach (cp. section 2.1.3) and define the root node as major claim and all nodes of the second level as claims. Thus argumentative relations of the first level are represented by the argument component types major claim and claim. In order to model supporting and attacking relations between claims and major claims, each claim includes a *stance attribute* which can take the values *for* or *against*. This approach has several advantages. First, it allows to capture several appearances of the major claim (e.g. in the opening and closing paragraph) without explicitly modeling argumentative relations crossing paragraph boundaries. Second, the argumentative relation identification can be limited to individual paragraphs which results in a more balanced class distribution between linked and not linked argument component pairs.[4] Note that by following this approach, we assume that argumentative relations between claims and major claims are the only ones which cross paragraph boundaries and that individual arguments starting with a claim are enclosed in its paragraphs. For modeling individual arguments, we apply a one-claim approach since introducing a new label for each level would result in an oversized set of labels and consequently in a poor class distribution when applied to real data. So our definition of a claim is two-fold: first it is a direct reason given for the major claim and second it represents the main *conclusion* of each individual argument. Finally, the premises are the reasons given for justifying or refuting another argument component. The stance (for or against) of each premise is encoded in the structure constituted by argumentative support and attack relations.

Having introduced our annotation scheme, we first highlight important properties of argumentation structures and the associated challenges for automatically identifying them before discussing the differences to our previous annotation scheme (Stab and Gurevych 2014a). From the example essay, we can infer the following challenges for the automatic identification of argumentation structures:

- *Segmentation of argument components*: A single sentence can include several argument components. Therefore, considering sentences as argument components is not sufficient and there is a need for segmenting the text in order to identify fine-grained reasoning.
- *Non-adjacent argumentative relations*: Argumentative relations can hold between non-adjacent argument components. For instance, in our example above the argumentative relation between premise_3 and claim_2 crosses an entire sentence. In addition, argumentative relations can point forward or backward. Therefore, all possible pairs of argument components need to be considered in order to identify argumentation structures.
- *Implicit vs. explicit argumentative relations*: An argumentative relation can be signaled by a lexical indicator (explicit relation) or not (implicit relation). For instance, the argumentative relation from premise_1 to claim_2 is implicit, whereas the argumentative relation from premise_2 to premise_1 is signaled by the discourse connective "*since*".

4 Note that argumentative relations are directed and thus all possible pairs of argument components need to be considered when identifying them. Concrete numbers illustrating the improvement of the class distribution are provided in section 5.

- *Several arguments in a paragraph*: Although essay writing guidelines recommend to include only a single aspect in a paragraph, the body paragraphs of real essays could include several arguments and thus several claims respectively (cp. third paragraph of the example essay).

In contrast to our previous work (Stab and Gurevych 2014a), the current annotation scheme exhibits the following differences. First, in previous work we restricted the number of major claims to exactly one per essay and asked the annotators to select the most representative major claim either in the opening or closing paragraph. However, in this work, we do not restrict the number of major claims resulting in more consistent argument component types since reformulations of major claims are not confused with claims. Second, our novel annotation scheme does not include argumentative relations crossing paragraph boundaries. In our previous work, we also annotated argumentative relations between claims and major claims which are exclusively represented by the argument component type in this work (level-approach). Third, to comply with the level-approach, we model reasons given for or against the major claim as claims, whereas in our previous work a premise could be also linked to a major claim. Thus, our novel annotation scheme separates the argumentative types of argument components more precisely and consequently results in more consistent annotations of argument components.

3. Related Work

3.1 Argumentation Mining

Existing work on Argumentation Mining focuses on a variety of subtasks. In general, there are two different research directions. *Argument Extraction* focuses on the separation of argumentative from non-argumentative text units (Moens et al. 2007; Florou et al. 2013), the identification of argument components (Rooney, Wang, and Browne 2012; Mochales-Palau and Moens 2009; Teufel 1999) and the recognition of argumentation structures (Mochales-Palau and Moens 2009; Wyner et al. 2010; Carstens and Toni 2015). Complementary, existing work on *Argument Attribution* aims at identifying certain properties of a given argument like the type of reasoning (Feng and Hirst 2011), the argumentation style (Oraby et al. 2015), the stance of the author (Somasundaran and Wiebe 2009; Hasan and Ng 2012), the acceptability of an argument (Cabrio and Villata 2012b, 2012a) or the types of appropriate support (Park and Cardie 2014). Accordingly, existing corpora in the field cover different aspects of argumentation.

3.1.1 Existing Corpora for Argumentation Structure Parsing. Existing corpora annotated with argumentation structures are still very rare. One prominent resource in this field is *AraucariaDB* (Reed et al. 2008) which includes structural annotations of arguments created with a graphical annotation tool (Reed and Rowe 2004). In addition to the structure of arguments, the corpus also contains annotations of argumentation schemes and thus the reasoning type of each argument. The corpus consists of more than 700 argument analyses and includes heterogeneous text genres different from our target domain like newspaper editorials, parliamentary records, judicial summaries and discussions. In addition, the reliability of the annotations is unknown.

Peldszus (2014) created a small corpus including 115 German microtexts of controlled linguistic and rhetoric complexity. The annotation scheme models the argumen-

tation structure including supporting and attacking relations as well as additional information like proponent and opponent. In a first annotation study, 26 naïve annotators applied the scheme to a subset of 23 microtexts in a classroom annotation experiment yielding an agreement of $\kappa = .38$ (Peldszus and Stede 2013b). In subsequent work, they extended their corpus and obtained an inter-annotator agreement of $\kappa = .83$ among three expert annotators using the full label set. Recently, they also translated the corpus to English resulting in the first parallel corpus for Argumentation Mining including 112 arguments (Peldszus and Stede 2015). However, since the microtexts are written according to predefined rules, the corpus lacks non-argumentative text units and thus is not applicable for end-to-end argumentation structure parsing. For the same reason, it exhibits an unusually high proportion of attack relations (Peldszus and Stede 2013a, p. 197). In particular, 86.6% of all arguments include at least one attack relation.

Kirschner, Eckle-Kohler, and Gurevych (2015) annotate argumentation structures in introduction and discussion sections of 24 German scientific articles from the educational domain. Their annotation scheme includes four argumentative relations (support, attack, detail and sequence). Considering argumentative relations and their types between argument components less distant than 6 sentences, they achieve an agreement of $\kappa = .43$ among four annotators. However, in contrast to previously mentioned corpora, the granularity of the argument components is limited to sentences and it does not include annotations of argument component types.

Table 1
Existing corpora annotated with document-level argumentation structures at the micro-level. (#doc = number of documents; #comp = number of argument components; NoArg indicates the presence of non-argumentative text units; IAA = inter-annotator agreement; *Note that non-argumentative text units are not included in recent releases.)

Source	*Genre*	*#doc*	*#comp*	*NoArg*	*Granularity*	*IAA*
(Reed et al. 2008)	various	~700	~2,000	yes*	clause	unknown
(Stab and Gurevych 2014a)	student essays	90	1,552	yes	clause	$\alpha_U = .72$
(Peldszus and Stede 2015)	microtexts	112	576	no	clause	$\kappa = .83$
(Kirschner et al. 2015)	scientific articles	24	~2,700	yes	sentence	$\kappa = .43$

In previous work, we created a corpus of 90 persuasive essays (Stab and Gurevych 2014a). We annotated the corpus in two consecutive steps. First, we identified the boundaries of argument components and their argumentative type yielding an inter-annotator agreement of $\alpha_U = .72$ among three annotators. Second, we annotated argumentative support and attack relations with an agreement of $\alpha = 0.8$. Compared to the microtext corpus from Peldszus, this resource also includes non-argumentative text units and exhibits a more realistic proportion of argumentative attack relations, since the essays are not written in a controlled experiment.

Table 1 provides an overview of existing corpora annotated with argumentation structures. Apart from those resources, there are several corpora which focus on other aspects of Argumentation Mining. For instance, there are corpora including only annotations of argument components like those created by Mochales-Palau and Ieven (2009), Kwon et al. (2007) or Habernal and Gurevych (2015). Other resources focus on the Information Retrieval perspective and include topic-specific claims and evidences across different documents (Aharoni et al. 2014) or macro-level relations between arguments (Cabrio and Villata 2012b; Boltužić and Šnajder 2014) and do not include document-level argumentation structures.

Persuasive essays are extensively studied in computational linguistics. For instance, previous work on persuasive essays focuses on automated essay grading (Shermis and Burstein 2013; Attali, Lewis, and Steier 2013), the identification of shell expressions (Madnani et al. 2012), style criteria (Burstein and Wolska 2003), metaphors (Beigman Klebanov and Flor 2013), factual knowledge (Beigman Klebanov and Higgins 2012) or modeling of thesis clarity (Persing and Ng 2013), prompt adherence (Persing and Ng 2014) or argumentation strength (Persing and Ng 2015). Song et al. (2014) annotated 600 essays with argumentation schemes. In particular, they select three argumentation schemes from Walton's theory (Walton, Reed, and Macagno 2008) related to two different prompts and derive an annotation scheme of 16 critical questions which represent quality flaws in the justifications of an argument. In an annotation study with 4 annotators, they obtain an inter-annotator agreement of $\alpha = .33$ to $.85$ depending on particular categories. Besides our previous work (Stab and Gurevych 2014a), we are only aware of one additional study of argumentation structures in persuasive essays. Botley (2014) analyzes 10 essays using argument diagramming in order to study argumentation strategies of students. However, the data set is too small for computational purposes and the reliability of the annotations is unknown.

3.1.2 Argument Extraction. The first step of an Argumentation Mining system is to separate argumentative from non-argumentative text units. This task is usually considered as a binary classification of sentences or text segments and a subsequent step for identifying the precise boundaries of argument components. For instance, Moens et al. (2007) identify argumentative sentences in AraucariaDB and obtain the best accuracy of .738 with a multinomial naïve Bayes classifier using word pair, text statistics, verb, and keyword features. Florou et al. (2013) present a similar approach. They classify text segments crawled with a focused crawler as either containing an argument or not using several discourse markers and features extracted from the tense and mood of verbs. They report a F1-score of .764 for their best performing system. For enabling a more fine-granular analysis of arguments, Levy et al. (2014) propose a pipeline including three consecutive steps for identifying the precise boundaries of context-dependent claims in Wikipedia articles (Aharoni et al. 2014). Given a debatable topic, their first component detects sentences containing a claim which is related and relevant to the topic. Subsequently, their boundary component generates several candidates of sub-sentences based on a maximum likelihood model of which the most probable claim is selected using a logistic regression classifier. Finally, they rank the identified candidates using another logistic regression classifier in order to identify the most probable claims for the given topic. Goudas et al. (2014) present a two step approach for identifying argument components in social media texts. First, they classify each sentence as argumentative or non-argumentative and achieve the best accuracy of .774 using logistic regression. Subsequently, they segment each argumentative sentence using an IOB-tagset and a *Conditional Random Field* in order to identify the boundaries of argument components and obtain an accuracy of .424. In contrast to existing approaches, we employ a single sequence model to separate argumentative from non-argumentative text units and to identify the boundaries of argument components (cp. section 5.2) in order to prevent potential error propagation in a pipeline approach.

The next step of an Argumentation Mining system is to classify the argumentative type (e.g. claims or premises) of argument components. For example, Kwon et al. (2007) experiment with several classifiers to identify main claims in public comments about

environmental protection of a rule making platform. First, they classify each sentence as including a claim and obtain a F1-score of .55 using a boosting algorithm (Schapire and Singer 2000). Second, they classify each claim as support, oppose or propose and achieve an F1-score of .91 using several subjectivity clues, FrameNet frames, syntactic and lexical features. Rooney, Wang, and Browne (2012) apply kernel methods for classifying argument components in AraucariaDB. They classify each proposition as claim, premise or non-argumentative and report an overall accuracy of .65. Mochales-Palau and Moens (2011) applied two binary classifiers for classifying propositions of the ECHR corpus as either claim or premise. In addition to domain-dependent key phrases, they use token counts, location features, information about verbs, and the tense of the sentence and achieve a F1-score of .741 for claims and .681 for premises using a *Support Vector Machine* (SVM). In our previous work, we employed a multiclass SVM to classify propositions of persuasive essays as non-argumentative, major claim, claim or premise (Stab and Gurevych 2014b). Using several structural, lexical, syntactic, indicator and contextual features, we achieved an accuracy of .773. Recently, Nguyen and Litman (2015) found that argument and domain words extracted from unlabeled persuasive essays increase the accuracy to .79 in the same data set, and Lippi and Torroni (2015) achieved promising results using partial tree kernels for identifying sentences containing a claim.

Complementary to the identification of argument components and their argumentative types, the approach presented in this article also focuses on the identification of argumentative relations between argument components. One of the few approaches is based on a manually created *Context-Free Grammar* (CFG) tailored to legal argumentation (Mochales-Palau and Moens 2009). Although, their corpus does not contain annotations of argumentative relations, the approach allows to identify argumentation structures as trees.[5] However, since their approach relies on the presence of domain-dependent key words and is specially tailored to documents from the legal domain, it does not accommodate ill-formatted arguments (Wyner et al. 2012). In addition, the approach is not capable of identifying implicit argumentative relations (Stab and Gurevych 2014b). In order to also identify implicit argumentative relations, we proposed to consider the task as a binary classification of ordered argument component pairs (Stab and Gurevych 2014b). We classified each pair including a target and source argument component as support or not argumentatively related. Using several structural, lexical, syntactic and indicator features, we obtained a macro F1-score of .722 using an SVM in persuasive essays. Peldszus (2014) presents another approach for identifying argumentation structures in the before mentioned German microtexts (cp. section 3.1.1). In contrast to classifying argument component pairs, he encodes the target of argumentative relations together with additional information like opponent, proponent, support and attack, etc. in a single tagset and considers the task as a multiclass classification task. For instance, there are several tags which denote if an argument component at position n is argumentatively related to preceding argument components $n-1$, $n-2$, etc. or following argument components $n+1$, $n+2$ etc. Although, he achieved promising results (accuracy of .48 for the 16 target tags and .39 for the whole tagset including 48 tags), the approach is only applicable to relatively small texts since for larger documents the tagset will increase tremendously and it poses additional challenges to the size of corpora due to potential class distribution issues. Very recently Peldszus and Stede (2015) presented the first approach for globally modeling

5 Note that Mochales-Palau and Moens (2009) evaluate the approach only by means of the identified claims and premises and don't consider argumentative relations in their evaluation.

several aspects of argumentation structures in a *Minimum Spanning Tree* model (MST) and report a F1-score of .720 for identifying argumentative relations in their english corpus. They jointly modeled several aspects of the argumentation structure and found that the function (support or attack) and the role (opponent and proponent) are the most beneficial dimensions for improving the identification of argumentative relations. However, since their corpus includes a comparatively high proportion of attacking argument components due to the employed writing guideline (cp. section 3.1.1), it is not clear wether the results can be reproduced on real data. In addition, their approach is not capable of segmenting argument components and assumes that a given text includes exactly one tree or argument respectively.

Other approaches aim at identifying argumentative relations at the macro-level and thus focus on argumentative relations between arguments. For instance, Ghosh et al. (2014) identify the targets of arguments in blog comments and classify the relations as agree, disagree or other. Boltužić and Šnajder (2014) link user comments to a set of predefined arguments either with a support or attack relation, and Cabrio and Villata (2012b) identify argumentative support and attack relations between arguments of a debating portal using textual entailment (Dagan et al. 2013). However, all of these approaches are limited to argumentative relations at the macro-level and do not consider argumentative relations between argument components.

3.1.3 Argument Attribution. In contrast to the recognition of arguments, their components and structure, argument attribution focuses on identifying certain properties of arguments. For instance, Feng and Hirst (2011) automatically identify the five most frequent argumentation schemes in AraucariaDB. 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. They experiment with several classification setups and achieve an accuracy between 62.9% and 97.9% using a binary C4.5 Decision Tree for each individual argumentation scheme. Oraby et al. (2015) identify the argumentation style of arguments included in the *Internet Argument Corpus* (IAC) (Walker et al. 2012). They classify each argument as either factual or emotional in order to separate arguments exhibiting an argumentative merit from those which are based on emotional reasons. They employ a bootstrapping approach for extracting linguistic patterns from unlabeled arguments and classify an argument as feeling or fact if it matches at least three of the patterns of a category. Although, this approach increases precision, it exhibits a significantly lower recall compared to a supervised unigram baseline.

Another task related to argument attribution is stance recognition which aims at identifying the stance of an author on a controversial topic. This task is usually considered as labeling an authors' comment in online debates as either for or against. Since a single comment can also include concessions or statements opposing the view of the author, Somasundaran and Wiebe (2009) suggest an approach for maximizing the overall side-score of a comment by using *Integer Linear Programming* (ILP). They identify the probability that a particular term is associated either positively or negatively with the topic of the debate by extracting subjectivity clues and the associated targets from topic-relevant documents. In addition, they consider concessions recognized with a list of discourse constructs. In their experiments, they achieve accuracies between .611 and 1.0 depending on the four topics of 117 comments in their test set. In addition, to this unsupervised approach, Somasundaran and Wiebe (2010) experiment with arguing and sentiment features using a supervised classifier. They extract 3,094 positive and 668 negative arguing cues from the annotations of the MPQA corpus (Wilson, Wiebe, and

Hoffmann 2005) and show that combining the arguing polarity of a sentence with its content words yields promising results. They achieve the best results of .639 accuracy by using an SVM and a combination of sentiment and arguing features. In addition, there are several other approaches of stance recognition. For instance, Anand et al. (2011) experiment with lexical, structural, dependency and context features, and Hasan and Ng (2012) show that jointly modeling contextual information and authors' stances on particular subtopics or combining reason classification and stance classification yield promising results (Hasan and Ng 2014). Qiu and Jiang (2013) propose a novel generative latent variable model to capture the viewpoint, user identity and user interactions.

In contrast to existing approaches, we consider stance recognition as the task of classifying individual argument components as supporting or attacking rather than identifying the overall stance of a document since persuasive essays can include several arguments with either supporting or attacking components (cp. section 2.2).

3.2 Discourse Analysis

The identification of argumentation structures is closely related to discourse analysis since both share similar subtasks. Like argumentation structure parsing, discourse analysis aims at identifying elementary discourse units (EDU) and discourse relations between them. Existing approaches on discourse analysis mainly differ in the employed discourse theory. The two most common approaches are *Rhetorical Structure Theory* (RST) (Mann and Thompson 1987) and the *Penn Discourse Tree Bank* (PDTB) (Prasad et al. 2008). Whereas existing approaches based on RST focus on representing the discourse structure of a document as a tree by iteratively linking adjacent discourse units (Feng and Hirst 2014; Hernault et al. 2010), PDTB approaches identify more shallow structures linking two adjacent sentences or clauses (Lin, Ng, and Kan 2014). However, similar to argumentation structure parsing the main challenge of both approaches is to identify frequently occurring *implicit discourse relations* since those are not signaled by discourse connectives (Braud and Denis 2014, p. 1694).

Marcu and Echihabi (2002) propose one of the first approaches for identifying implicit discourse relations. In order to collect large amounts of training data, they exploit several discourse connectives like *'because'* or *'but'*. After removing the discourse connectives, they found that word pair features are indicative for implicit discourse relations and achieve accuracies between 64% and 75% depending on the utilized corpus for identifying cause-explanation-evidence relations (their most similar relation compared to argumentative relations). Pitler, Louis, and Nenkova (2009) identify four different types of implicit discourse relations in the PDTB and achieve F1-scores between .22 and .76 depending on the particular relation type. They also found that using a tailored feature sets for each individual relation leads to the best results. Lin, Kan, and Ng (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.

Although, argumentation structure parsing and discourse analysis share similar subtasks and challenges, there are some major differences between both approaches. First, existing approaches on discourse parsing are limited to the identification of discourse relations between adjacent text units (Peldszus and Stede 2013a), whereas argumentative relations also hold between non-adjacent text units (Stab and Gurevych 2014b). For instance, in our previous corpus only 37% of the premises appear adjacent to their target component. Second, the set of employed discourse relations differs. Whereas approaches on discourse analysis usually include a large set of discourse relations to

capture general discourse structures, only a subset of these relations is relevant for argumentation structure parsing. For example, Peldszus and Stede (2013a) introduce support, attack and counter-attack relations for modeling argumentative discourse, whereas our work includes only support and attack relations. This difference is also illustrated by the work of Biran and Rambow (2011). They select a subset of 12 relations from the RST Discourse Treebank (Carlson, Marcu, and Okurowski 2001) and argue that only a subset of the RST relations are relevant for identifying justifications for a given claim. Therefore, there is a need to employ novel methods for identifying argumentation structures, which consider relations between non-adjacent text units and solely focus on argumentative-relevant discourse units and relations

4. Corpus Creation

The motivation for creating a novel corpus complementary to our previous work is threefold. First, our previous corpus includes only 90 persuasive essays which is comparatively few to derive strong conclusions. We also believe that a larger corpus will result in more accurate identification methods. Second, by creating another corpus and conducting an additional annotation study, we ensure the reproducibility of the approach and validate our previous results. Third, we aim to create a more consistent language resource for Argumentation Mining by modifying our annotation scheme (cp. section 2.2). Accordingly, we revised our annotation guideline and included details about the common structure of persuasive essays in order to facilitate the annotation process. In addition, we added more precise rules for identifying argument component boundaries.[6]

Since we found in our preliminary investigation that knowing the topic of the essay and the stance of the author results in better agreement among the annotators (Stab and Gurevych 2014a, p. 1505), we define a top-down annotation process starting with the identification of the major claim before focusing on the identification of claims and premises. In addition, we ask the annotators to read the entire essay before starting with the annotation to ensure that they are aware of the topic of the essay. The entire annotation process includes the following steps:

1. *Identification of topic and stance*: The annotators identify the topic of the essay and the authors' stance by reading the entire essay before starting with the actual annotation task.
2. *Annotation of argument components*: The second step includes the annotation of major claims in the opening and closing paragraph followed by the annotation of claims, their stance attributes and premises.
3. *Linking claims and premises with argumentative relations*: In the last step, the structure of each argument is annotated by linking argument components with argumentative support and attack relations.

In contrast to our previous annotation study, we do not conduct any collaborative training session in order to receive a better estimation of the reliability and the quality of our annotation guideline. Three non-native speakers participate in our study, of

6 The revised guidelines as well as the created corpus will be available at www.ukp.tu-darmstadt.de/data/argumentation-mining

whom one annotator already participated in our previous study (expert annotator). The remaining two annotators trained the task solely by independently reading the guidelines. We used the *brat rapid annotation tool* (Stenetorp et al. 2012) which allows to annotate the boundaries of argument components as well as linking them with argumentative relations.

4.1 Data

Our novel corpus contains 402 persuasive essays in English which we selected from essayforum.com. This online portal is an active community that provides feedback for different writing tasks like scientific writing, poetry or speeches. The writing feedback section of this forum allows students to retrieve feedback about their essays while preparing themselves for standardized tests. Since most of the essay prompts anticipate an argumentative writing style, we extended our previous corpus with randomly selected essays from this section. However, since some forum posts lack the prompt of the essay, we manually reviewed each essay and selected only those essays which include a sufficiently detailed description of the writing task since this information can be valuable for future research. The final corpus includes $7{,}116$ sentences with $147{,}271$ tokens.

4.2 Inter-Annotator Agreement

In order to evaluate the inter-annotator agreement, all three annotators independently annotate a subset of 80 randomly selected persuasive essays. The remaining 322 essays are annotated by the expert annotator only. We evaluate the reliability of the argument component annotations by using two different strategies. First, we evaluate if a sentence contains an argument component of a particular type using percentage agreement and the two chance-corrected measures Fleiss' *multi-*π (Fleiss 1971)[7] and Krippendorff's α (Krippendorff 1980). Although, the boundaries of argument components differ from sentence boundaries, evaluating the reliability of argument components at the sentence level provides a good approximation of the inter-annotator agreement, since only 4.3% of the sentences in our evaluation set contain several argument components of different type. Additionally, it enables the comparison with previous and future annotation studies conducted at the sentence level. Second, we evaluate the agreement between the annotators by using Krippendorff's α_U (Krippendorff 2004) which also considers the differences in the argument component boundaries and thus allows for assessing the reliability of our annotation study more accurately. For determining the scores, we use *DKPro Agreement* (Meyer et al. 2014) which provides well-tested implementations of many different agreement measures and a unified interface for evaluating coding and unitizing studies.

The annotators agree best on the annotation of major claims (table 2). The percentage agreement of 97.9% and $\alpha = .879$ indicates that annotators can reliably annotate major claims in persuasive essays. In addition, the unitized alpha measure of $\alpha_U = .810$ shows that there are only few disagreements regarding the boundaries of major claims. The agreement scores of $\alpha = .833$ and $\alpha_U = .824$ also indicate good agreement for premises

7 Although Fleiss introduced his coefficient as a generalization of Cohen's κ (Cohen 1960), it is actually a generalization of Scott's π (Scott 1955), since it assumes a cumulative distribution of annotations by all annotators (Artstein and Poesio 2008). We follow the naming proposed by Artstein and Poesio and refer to the measure as *multi-*π.

Table 2
Inter-annotator agreement of argument component annotations estimated on a subset of 80 persuasive essays with three annotators.

	%	*multi-π*	α	α_U
MajorClaim	.979	.877	.879	.810
Claim	.889	.635	.635	.524
Premise	.916	.833	.833	.824

(Carletta 1996). We obtain the lowest agreement of $\alpha = .635$ for claims which shows that the identification of claims is a more complex task compared to the identification of other argument components. Compared to our previous study, the agreement of major claims is slightly better (Stab and Gurevych 2014a). The α-score increases by .043 and α_U by .037 which can be attributed to the relaxed constraints on major claim annotations because the annotators didn't need to select a particular instance of a major claim (cp. section 2.2). The α-score for premises also increases by .120 which might be a consequence of our more precise guidelines and the stringent compliance to the level-approach described in section 2.2. The α-score for claims decreases by .031. One reason for this slight decrease might be the fact that we did not conduct any collaborative training session as in our previous study. However, the joint unitized measure for all argument components is $\alpha_U = .767$, and thus we obtain an overall improvement of .043 compared to our previous study. Therefore, we conclude that our guideline as well as our annotation process guide annotators to substantial agreement.

In order to determine the agreement of the stance attribute of each claim, we follow the same methodology as for the computation of the argument component agreement, but treat each sentence containing a claim as either for or against. Note, that following this strategy, the agreement of claim annotations constitutes the upper bound for the agreement of the stance attribute. Using this strategy, we obtain an agreement of 88.5% and $\alpha = .623$ for the stance attribute which is only slightly below the agreement of claims. Therefore, the distinction between supporting or attacking claims seems to be an easy task for human annotators and is feasible with high agreement.

Table 3
Inter-annotator agreement of argumentative relation annotations estimated on a subset of 80 persuasive essays with three annotators.

	%	*multi-π*	α
support	.923	.708	.708
attack	.996	.737	.737

Table 3 shows the inter-annotator agreement of argumentative relations. We determine the scores by considering all pairs of argument components of each paragraph and thus the set of markables corresponds to all argumentative relations that were possible according to our annotation guideline. We obtain for both argumentative relations an α-score above .70 which allows tentative conclusions (Carletta 1996). Although, only 0.9% of the $4{,}922$ markables in our evaluation set are annotated as argumentative attack relations and the agreement is usually much lower if a category occurs rarely, the annotators agree better on argumentative attack relations which indicates that the identification of argumentative attack relations is a simpler task than identifying argumentative support relations. In comparison to our previous study, the α-scores are approximately .10 lower.

This can be attributed to the fact that we did not consider argumentative relations between claims and major claims which are easy to annotate since the labels of the components and the stance attribute of claims are known when annotating relations.

4.3 Error Analysis

In order to analyze the disagreements among annotators and to identify the most frequently confused categories, we determine *Confusion Probability Matrices* (CPM) (Cinková, Holub, and Kríž 2012) for argument components and argumentative relations. Compared to traditional confusion matrices, a CPM also allows to analyze confusions if more than two annotators are involved in the annotation study. In particular, a CPM includes conditional probabilities that an annotator assigns a category in the column given that another annotator selected the category in the row for a specific item.

Table 4
Confusion probability matrix of argument component annotations ('NoArg' indicates sentences without argumentative content).

	Major Claim	*Claim*	*Premise*	*NoArg*
Major Claim	.771	.077	.010	.142
Claim	.036	.517	.307	.141
Premise	.002	.131	.841	.026
NoArg	.059	.126	.054	.761

Table 4 shows the CPM for argument component annotations. It shows that the highest confusion is between claims and premises. The two main reasons for this confusion are context dependence and ambiguity of argumentation structures (Stab et al. 2014). In particular, if an argument includes a serial argumentation structure, the identification of the correct claim requires that the annotators are aware of the context of the argument which is particularly difficult if the argumentation structure is more complex. In addition, we observe that one annotator frequently did not split sentences including a claim. For instance, the annotator labeled the whole sentence as a claim although it includes another premise. These errors also explain the lower unitized alpha score compared to the sentence-level agreements in table 2. We also observe that concessions before claims were frequently not annotated as an attacking premise as defined in our guidelines. For instance, annotators frequently did not split sentences similar to the following example

> *Although, [in some cases technology makes peoples' life more complicated]$_{premise}$, [the convenience of technology outweighs its drawbacks]$_{claim}$.*

which also explains the high confusion between claims and premises. However, the distinction between major claims and claims exhibits less confusion. Apparently, the identification of major claims is easier compared to general claims which cover a certain aspect of the overall topic of the essay.

Table 5
Confusion probability matrix of argumentative relation annotations ('Not-Linked' indicates argument component pairs which are not argumentative related).

	Support	*Attack*	*Not-Linked*
Support	.605	.006	.389
Attack	.107	.587	.307
Not-Linked	.086	.004	.910

Table 5 shows the CPM of argumentative relation annotations which reveals that annotators well distinguished between argumentative support and attack relations. The highest confusion is between argumentative relations (support and attack) and not linked argument component pairs which can be attributed to the identification of the correct targets of a relation. In particular, in the presence of multiple claims or serial argumentation structures, the identification of the targets is a challenging task.

4.4 Creation of the Final Corpus and Corpus Statistics

Since not all annotators labeled the whole corpus (cp. section 4.2), we create a partial gold standard including only essays annotated by all annotators. In particular, we use this partial gold standard including 80 essays as our test set (20%) and the remaining 322 essay which are annotated by the expert annotator as our training set (80%). The creation of our gold test set consists of the following steps. First, we consolidate the argument components before the annotation of argumentative relations. So each annotator uses the same argument components when annotating argumentative relations. Second, we merge the argumentative relations to compile our final gold test set. Since the argument component types are strongly related - for instance, the selection of the premises depends on the selected claim(s) in a paragraph - we didn't merge the annotations by majority voting as in our previous study. Instead, we discussed the disagreements in several meetings with all annotators in order to resolve disagreements and to create a more consistent gold test set.

Table 6
Statistics of the final corpus including 402 argument structure annotated essays.

		all	*avg. per essay*	*standard deviation*
size	Sentences	7,116	18	4.2
	Tokens	147,271	366	62.9
	Paragraphs	1,833	5	0.6
arg. comp.	Arg. components	6,089	15	3.9
	Major Claims	751	2	0.5
	Claims	1,506	4	1.2
	Premises	3,832	10	3.4
	Claims (for)	1,228	3	1.3
	Claims (against)	278	1	0.8
rel.	Support	3,613	9	3.3
	Attack	219	1	0.9

Table 6 shows an overview of the entire corpus. It includes 6,089 argument components of which 751 are major claims, 1,506 are claims and 3,832 are premises. The proportion between claims and premises is common in argumentative writing since writers usually provide several reasons for ensuring a robust standpoint (Mochales-Palau and Moens 2011, p. 10). In addition, the less frequent attacking claims and argumentative attack relations confirm that students tend to support their own standpoint instead of including opposing views (Wolfe and Britt 2009).

5. Approach

Our approach for automatically identifying argumentation structures includes five subtasks depicted in figure 3. We will first provide a brief overview of each task before describing the individual models in detail.

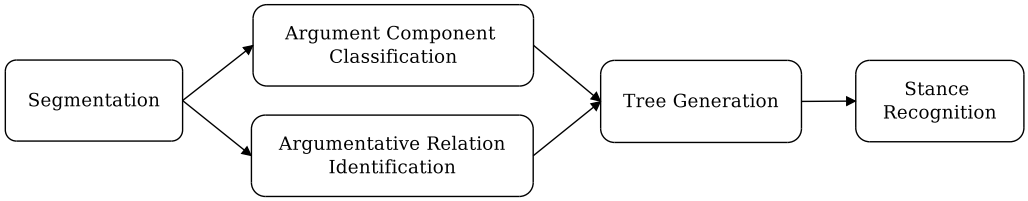


Figure 3
Architecture of the argumentation structure parser

The first model segments a persuasive essay in order to identify argument components and to separate argumentative from non-argumentative text units. Since, argument component boundaries differ from sentence boundaries and a single sentences can include several argument components (Sergeant 2013), we consider this task as a sequence labeling task at the token-level (*segmentation*). In the next step, we jointly model argumentative relations and argument component types. First, we learn two base classifiers, of which the first one classifies each argument component as either major claim, claim or premise (*argument component classification*). The second base classifier recognizes if two argument components are argumentatively linked or not (*argumentative relation identification*). Second, we define a joint model that merges the results of the two base classifiers in order to find a tree (or several ones) in each paragraph which optimizes the results of the two base classifiers (*tree generation*). Since, argumentative relations between claims and major claims are encoded using a level-approach (cp. section 2.1.3), we only consider subtrees starting with a claim and thus we don't consider major claims in our joint model. Following this procedure has not only positive effects on the class distribution but also reduces the total amount of instances for the argumentative relation identification model. In particular, the number of argument component pairs reduces from 92,382 to 22,172 and the proportion of linked argument component pairs improves from 4.15% to 17.3%.[8] Finally, the stance recognition model discriminates between supporting and attacking relations (*stance recognition*). We model this task as a binary classification of the source components of each argumentative relation and also consider the stance attribute of claims in order to identify the stance of each individual argument on the major claim.

In the following sections, we present each model and its evaluation in detail. In particular, we conduct 5-fold cross-validation on our training set to prevent overfitting and in order to find the best performing models before testing them on our gold test set. In section 5.2, we introduce our approach for segmenting persuasive essays including the experimental results on the gold test set since the model is independent of the following steps. In sections 5.3, 5.4 and 5.5, we present the two base classifiers and our joint model for identifying argumentation structures. We analyze each model, the influence of different feature sets and several configurations before evaluating the joint model on our gold test set and comparing the results to human performance (section 5.5.1). Finally, we introduce our stance classification model in section 5.6. Note that we evaluate both the joint model and the stance recognition model using the gold segments in our corpus. In addition, we show in section 5.7 that the joint modeling approach also

8 Note that the numbers are determined on our entire corpus.

simultaneously improves the results of argument component classification and relation identification using the microtext corpus from Peldszus and Stede (2015).

Throughout this article, we determine the evaluation scores of cross-validation experiments 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 order to determine the overall evaluation scores, we employ macro-averaging as described by Sokolova and Lapalme (2009, p. 430) and report macro precision, macro recall and macro F1-scores.

5.1 Preprocessing

For preprocessing our data we use several models included in *DKPro* (Eckart de Castilho and Gurevych 2014). In order to tokenize the data and to identify sentence boundaries we employ the *Language Tool Segmenter*[9]. Next, we identify the paragraphs of persuasive essays by checking for line breaks. We lemmatize each token using the *Mate Tools Lemmatizer* (Bohnet et al. 2013) and apply the Stanford part-of-speech tagger (Toutanova et al. 2003), constituent and dependency parsers (Klein and Manning 2003), and sentiment analyzer (Socher et al. 2013). Finally, we use a PDTB-Parser (Lin, Ng, and Kan 2014) for recognizing general discourse relations. After these steps, we use the *DKPro-TC* text classification framework (Daxenberger et al. 2014) for feature extraction and experimentation.

5.2 Segmenting Argument Components

We consider the task of identifying argument components and their boundaries as a sequence labeling task at the token-level and encode the argument components using an IOB-tagset (Ramshaw and Marcus 1995). Accordingly, we label the first token of an argument component as *Arg-B*, the remaining tokens of an argument component as *Arg-I* and tokens which are not covered by an argument component as *O*. Table 7 depicts the resulting class distribution in our train and test set. It indicates that $67{,}8\%$ of the tokens belong to argument components and $32{,}2\%$ are not argumentative. In total, 583 (9.6%) sentences in our entire corpus include several argument components of which 302 (4.9%) include argument components of different argumentative types.

Table 7
Class distribution of the train and test set for argument component segmentation.

	train	*test*
Arg-B	4,823 (4.1%)	1,266 (4.3%)
Arg-I	75,053 (63.6%)	18,655 (63.6%)
O	38,071 (32.3%)	9,403 (32.1%)

Since persuasive essays exhibit a common structure, we employ a heuristic baseline based on sentence boundaries. Usually, the first sentences of an essay introduce the controversial topic and are not argumentatively relevant. Similarly, the last sentence of an essay frequently includes a summarization or a recommendation which does not contribute to the argumentation. Therefore, our heuristic baseline selects all sentences

9 www.languagetool.org

as argument components except the first two and the last sentence. In addition, we use a majority baseline which classifies all tokens of an essay as Arg-I.

For our system, we employ a *Conditional Random Field* (CRF) (Lafferty, McCallum, and Pereira 2001) implemented in *CRFSuite* (Okazaki 2007) and embedded in DKPro-TC using the averaged perceptron training method (Collins 2002). Since a CRF also considers contextual information, the model is particularly suited for sequence modeling tasks and thus for the segmentation of argument components (Goudas et al. 2014, p. 292).

5.2.1 Features. We extract the following features for argument component segmentation:

Structural Features: For each token we define two binary features indicating if the token is present in the opening or closing paragraph. Since both include less argumentatively relevant information, we expect that these features are effective for filtering non-argumentative content in persuasive essays. In addition, we use six features indicating the absolute and relative position of a token in its sentence, its paragraph and the whole essay and two binary features set to true if the token is the first or last word of a sentence. For capturing the position of the covering sentence of each token, we define four features representing the absolute and relative position of the sentence in its paragraph and the entire essay. In addition, we define eight binary features which indicate if the token directly follows or precedes any punctuation, a full stop, a comma or a semicolon, since it is more likely that an argument component begins or ends after or before a punctuation. Two additional binary features signal if the token is any punctuation or a full stop.

Syntactic Features: To capture the syntactic characteristic of each token, we extract several features based on part-of-speech annotations and constituent parse trees. First, we extract for each token its POS-tag since it is less likely that for instance a verb indicates the beginning or end of an argument component. Second, we define two features from the constituent parse tree of the covering sentence of the token. In particular, we measure the length of the path to the *Longest Common Ancestor* (LCA) of the current token and the preceding and following tokens, and normalize the length according to the total depth of the tree. So we define the first feature considering the preceding token as $LCA_{preceding}(t_i) = \frac{|lcaPath(t_i,t_{i-1})|}{depth}$, where t_i is the current token, $|lcaPath(u,v)|$ the length of the path from u to the LCA of u and v, and $depth$ the depth of the constituent parse tree. Similarly, we define $LCA_{following}(t_i) = \frac{|lcaPath(t_i,t_{i+1})|}{depth}$ as the second feature considering the current token t_i and its following token t_{i+1}.[10] In addition, we define the constituent type of the LCA between the current and the following and preceding tokens as two additional features since it is less likely that e.g. a noun or verb phrase is split by argument component boundaries.

Lexico-syntactic Features: We adopt the lexico-syntactic features introduced by Soricut and Marcu (2003) which have been shown to be effective for segmenting elementary discourse units (Hernault et al. 2010). We use lexical head projection rules (Collins 2003) implemented in the Stanford tool suite to lexicalize the syntactic constituent parse tree. For each token t, we extract its uppermost node n having the lexical head t and define the first lexico-syntactic feature as the combination of t and the constituent type of n. In

10 Note that we set $LCA_{preceding} = -1$ and $LCA_{following} = -1$ if t is the first or last token of its covering sentence.

addition, we consider the child node of n in the path to t and its right sibling, and also combine their lexical heads and constituent types analog to the approach described by Soricut and Marcu (2003).

Probability Features: Argument components are frequently embedded in content-independent elements (also referred to as *shell language*) which indicate how the argument components are related to each other (Madnani et al. 2012). For instance, argument components frequently exhibit preceding discourse connectives like "*therefore*", "*thus*", "*because*" or phrases like "*to sum up*", "*another reason*" or "*in addition*" which signal the beginning of an argument component. Therefore, we define the conditional probability $P(c_i = \textit{Arg-B}|\textit{precedingTokens})$ where c_i is the label of token t_i and $\textit{precedingTokens} = t_{i-n}, ..., t_{i-1}$. We determine these probabilities for preceding tokens of length up to $n = 3$ by counting its occurrences preceding an argument component divided by the total number of its occurrences in our training data. During feature extraction, we determine for each token in our dataset the probability of its preceding unigram, bigram and trigram and select the highest probability as a feature.

5.2.2 Results. In order to identify the best performing system and to investigate the effectiveness of individual feature groups, we conduct several 5-fold cross-validation experiments on our training data before comparing the best system to the gold test set, baselines and human performance.

Table 8
Feature analysis of argument component segmentation using 5-fold cross-validation on the training set (Acc = Accuracy; F1 = Macro F1; P = Macro Precision; R = Macro Recall)

	Acc	*F1*	*P*	*R*	*F1 Arg-B*	*F1 Arg-I*	*F1 O*
CRF only lexSyn	.803	.762	.780	.744	.714	.873	.620
CRF only probability	.686	.605	.698	.534	.520	.806	.217
CRF only structural	.856	.748	.757	.740	.542	.906	.789
CRF only syntactic	.792	.730	.752	.710	.638	.868	.601
CRF all w/o structural	.827	.793	.809	.777	.753	.887	.677
CRF all features	**.895**	**.849**	**.853**	**.846**	**.777**	**.927**	**.842**

By conducting feature ablation tests, we found that a combination of all features yields the best results on our training data. The system achieves an accuracy of $.895$ and a macro F1-score of $.849$ (table 8). Compared to all other feature groups, lexico-syntactic features perform best for identifying the beginning of argument components (Arg-B). Syntactic features are also effective for identifying the beginning of argument components and also contribute to the separation of argumentative from non-argumentative text units. Although, our probability feature exhibits the lowest scores, it contributes to the performance of our system when combining it with other features. In particular, we observe a significant decrease of $.006$ accuracy and $.028$ F1-score of Arg-B when evaluating the system without the probability feature (McNemar test (McNemar 1947) with $p = 0.05$). Using only structural features performs particularly well for separating argumentative from non-argumentative text units. In addition, structural features also contribute to the identification of the beginning of argument components since information about punctuations and sentence positions successfully guides the learner to exclude shell language preceding an argument component. However, since some of the structural features are exclusive to persuasive essays and do not capture the semantics of argumentative discourse, we also report the results of running the system without

structural features which yields a macro F1-score of .793. Although, this score is .056 lower compared to the system which uses all features, it still significantly outperforms the heuristic baseline ($p = 0.05$). This result indicates that our system still performs reasonably well when removing genre-dependent features.

Table 9 shows the evaluation results on our gold test set. Our heuristic baseline performs reasonably well. It achieves a macro F1-score of .642 and thus outperforms the majority baseline by .383. In addition, it achieves an F1-score of .677 for tokens not included in argument components (O) and .867 for tokens covered by argument components (Arg-I). This indicates that it effectively separates argumentative from non-argumentative text units. However, since it considers only sentence boundaries, it achieves a low F1-score of .364 for the identification of the beginning tokens of argument components (Arg-B). For the same reason, it identifies 11.5% less argument components compared to the total amount in our gold test set.

Table 9
Results of argument component segmentation on the gold test set. († = significant improvement over the heuristic baseline (McNemar test (McNemar 1947) with p=0.05); Acc = Accuracy; F1 = Macro F1; P = Macro Precision; R = Macro Recall)

	Acc	*F1*	*P*	*R*	*F1 Arg-B*	*F1 Arg-I*	*F1 O*
Human Upper Bound	.919	.886	.887	.885	.821	.941	.892
Baseline Majority	.636	.259	.212	.333	0	.778	0
Baseline Heuristic	.794	.642	.664	.621	.364	.867	.677
CRF all features †	**.905**	**.867**	**.873**	**.861**	**.809**	**.934**	**.857**

Our best performing system using all features significantly outperforms the heuristic baseline on our test set and achieves an accuracy of .905 and a macro F1-score of .867. Compared to the heuristic baseline, it performs considerably better in identifying the beginning of argument components and also exhibits a better performance for separating argumentative from non-argumentative text units. In addition, the number of identified argument components differs only slightly from the total number in our gold test set. In particular, it identifies 1,272 argument components, whereas our gold test set includes 1,266 argument components.

We determine the upper bound for this task by means of human performance. In particular, we compare pairs of annotators and average the three resulting scores to obtain the human upper bound in table 9. Compared to our system, the human upper bound is only slightly better. The accuracy of our system is only .014 less than the human upper bound and thus our model achieves 98.5% of human performance.

5.2.3 Error Analysis. To identify frequent errors of our system, we analyze the most frequent confusions and manually investigate the identified argument components. We observe the most frequent errors in the false positives of Arg-I. The system classifies 16.5% of non-argumentative tokens as Arg-I and thus tends to identify more argumentatively relevant tokens. The reasons for this high confusion are threefold. First, we observe that the system frequently annotates argument components in the closing paragraphs which are not labeled as argumentative in our gold test set. In most cases, these include recommendations or summarizations which do not contribute to the argumentation. Second, compared to our gold test set, the model identifies more argument components in body paragraphs. Although, body paragraphs exhibit only few non-argumentative content apart from shell language, the model wrongly identifies argument components in 13 out of the 15 non-argumentative body paragraph sentences

in our test set. The reason for these classification errors can be attributed to the high class imbalance in our training data which exhibits only 3.5% non-argumentative body paragraph sentences. Third, we observe that the model tends to annotate longer shell expressions as argument components. For instance, in sentences similar to "*Although it is true in some cases that [Actual Arg-B] ...*" it identifies the first underlined text unit as argument component, since it is syntactically very similar to argument components though it does not include argumentatively relevant content.

Beside the confusion of argumentative and non-argumentative tokens, the second most frequent error is due to the identification of the correct beginning of argument components. Although, the model identifies 81.1% of the beginning tokens correctly, it classifies 10.8% as argumentative (Arg-I). By investigating the identified argument components, we observe that these errors are also due to longer shell expressions. For example, the model fails to identify the correct beginning in sentences like "*Hence, from this case we are capable of stating that [Actual Arg-B] ...*" or "*Apart from the reason I mentioned above, another equally important aspect is that [Actual Arg-B] ...*" (underlined text units represent the annotation of our model). These examples also explain the false negatives of non-argumentative tokens which are wrongly classified as Arg-B.

5.3 Argument Component Classification

The next model in our pipeline identifies the argumentative type of each argument component. We consider this task as a multiclass classification task and classify each argument component as either major claim, claim or premise. Table 10 shows the class distribution of the train and test set indicating that major claims are the minority class with only 12.3%, and premises are the majority class with 63.3% in our entire corpus.

Table 10
Class distribution of the train and test set for argument component classification.

	train	*test*
MajorClaim	598 (12.4%)	153 (12.1%)
Claim	1,202 (24.9%)	304 (24.0%)
Premise	3,023 (62.7%)	809 (63.9%)

As the baseline, we employ a majority baseline which classifies each argument component as premise. In addition, we exploit the structure of persuasive essays to define a heuristic baseline. Essay writing guidelines recommend that each body paragraph should start with a *topic sentence* which presents the main idea of the paragraph, relates the idea to the topic of the essay and expresses the authors' viewpoint about the idea (Whitaker 2009, p. 15) (Perutz 2010, p. 4). The remaining sentences in a body paragraph should provide evidence and reasons for supporting the topic sentence. Therefore, our heuristic baseline classifies the first argument component in body paragraphs as claim and the remaining sentences as premises. Furthermore, the heuristic baseline classifies the last argument component in the opening paragraph as major claim and all remaining argument components in the opening paragraph as claims since those are likely to support the major claim. Similarly, it classifies the first component in the closing paragraph as major claim and the remaining argument components in closing paragraphs as claims. Since 62% of the body paragraphs in our entire corpus start with a claim and all major claims appear either in the opening or closing paragraph, we expect that this baseline will yield good results for classifying argument components.

For our system, we use a *Support Vector Machine* (SVM) (Cortes and Vapnik 1995) with polynomial kernel implemented in the Weka machine learning framework (Hall et al. 2009). The motivation for selecting this particular learner stems from the results of our previous work in which we found that SVMs outperform other common classifiers in argument component classification tasks (Stab and Gurevych 2014b, p. 51). In addition, the Weka implementation of SVMs is ranked among the best performing learners in an extensive study conducted by Fernández-Delgado et al. (2014).

5.3.1 Features. We employ the following features for classifying argument components:

Lexical Features: We use lemmatized unigram features extracted from each argument component and its preceding tokens in the sentence which are not covered by another argument component. So we ensure that discourse connectives and argumentative type indicating shell language is included in the set of lexical features. We consider all unigrams as binary features. Instead of using ngrams of higher order, we add the 2k most frequent lemmatized word pairs extracted from the dependencies generated by the Stanford dependency parser since, these enable word dependencies between non-adjacent words.

Structural Features: We add two binary features indicating if the current argument component is the first or last argument component in its paragraph. Two binary features represent if the argument component is present in the opening or closing paragraph of the essay. In addition, we add the relative position of the argument component in its paragraph, the number of argument components in the paragraph, the number of tokens of the argument component and the number of tokens of its covering sentence. Four additional features encode the number of tokens before and after the argument component in its covering sentence, the ratio between tokens of the argument component and its covering sentence, and if the argument component boundaries match the boundaries of the covering sentence. In addition, we add two features representing the number of argument components preceding and following the current argument component in its covering paragraph.

Indicator Features: We manually select a list of indicators which signal the argumentative type of argument components from 30 additional essays not included in our corpus. In particular, we select four different types of indicators which signal the direction of reasoning as well as the type of certain components. First, we select 24 indicators which signal *forward reasoning*. These indicators represent that the argument component following the indicator is a result of preceding components. For instance, the list includes indicators like "*therefore*", "*thus*" or "*consequently*". Second, we select 33 indicators which signal *backward reasoning*. For example, indicators like "*for instance*", "*one of the main reasons*" or "*furthermore*" signal that the argument component following the indicator refers to preceding argument components and might indicate the presence of premises. Third, we select *rebuttal indicators* which indicate opposing reasons or contra arguments. Since, authors of persuasive essays tend to support their own viewpoint instead of including opposing views (Wolfe and Britt 2009), we only found 10 of these indicators in our 30 essays. Examples are, "*although*", "*admittedly*" or "*but*". Fourth, we found 48 *thesis indicators* which signal the presence of the authors' stance and the major claim respectively. These include for instance, "*I think*", "*I totally agree*", or "*in my opinion*". For each argument component, we extract four binary features signaling if one indicator of the four categories is present in the component or its

preceding tokens. An additional binary feature indicates if the argument component or its preceding tokens includes a reference to the first person. In particular, we check the presence of the five words "*I*", "*me*", "*my*", "*mine*" and "*myself*". We expect that this feature can help to identify major claims, since authors frequently refer to their own opinion.

Contextual Features: Contextual information plays a major role for identifying the type of argument components (Mochales-Palau and Moens 2007). For instance, it is likely that argument components close to a known claim serve as evidence and are presumably premises. Therefore, we define several contextual features based on our defined indicators. We add for each argument component eight binary features representing the presence of a forward, backward, rebuttal or thesis indicator preceding or following the argument component in its paragraph. We assume that these features are useful for modeling the local context of an argument component if indicators are present in a paragraph. For instance, if it is known that a forward indicator follows a particular argument component, it is less likely that the current component is a claim. In addition, we add several features representing content overlap with the opening and closing paragraph, since claims frequently restate entities or phrases from the major claim or the general topic of the essay. So, we determine the number of noun and verb phrases of the current component shared with the opening and closing paragraph. Additionally, we add four binary features indicating if the argument component shares any noun or verb phrase with the opening or closing paragraph.

Syntactic Features: To capture the syntactic characteristics of argument components, 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. Premises often refer to previous events and claims are usually in the present tense. So we capture the tense of the main verb of an argument component as binary feature (past or present). Since claims frequently exhibit modals like "*should*", "*can*" or "*could*" to express uncertainty, we use the POS-tags generated during preprocessing to identify modals and define a binary feature which indicates if an argument component contains a modal verb. Finally, we add the number of each POS-tag present in an argument component, hypothesizing that argument components of different types exhibit varying syntactic distributions.

Probability Features: We determine the probability of each argument component type given its preceding tokens as $P(type|precedingTokens)$, where *precedingTokens* are the tokens preceding an argument component in its covering sentence and $type \in \{majorClaim, claim, premise\}$. We determine these conditional probabilities by counting the occurrence of the tokens preceding a particular argument component type and dividing it by the total number of occurrences in our training data. For each argument component we add the probability for major claim, claim and premise to our feature set.

Discourse Features: Since Cabrio, Tonelli, and Villata (2013) showed by analyzing several example arguments that general discourse relations could be useful for identifying argument components, we add features based on the output of the PDTB-style discourse parser (Lin, Ng, and Kan 2014). In particular, we add a set of binary features combining the type of the relation, if the current argument component overlaps with the first or second elementary discourse unit of the discourse relation

and if the discourse relation is implicit or explicit. For instance, we add the feature "*Contrast_imp_Arg1*" if the argument component overlaps with the first elementary discourse unit of an implicit contrast relation, or "*Cause_exp_Arg2*" if the argument component overlaps with the second elementary discourse unit of an explicit cause relation.

Word Embedding Features: We employ the word embeddings trained on a part of the Google News data set (Mikolov et al. 2013) and sum the vectors for each word of an argument component and its preceding tokens.

5.3.2 Results. To analyze the performance of our features and to identify the best performing system, we investigate each feature group individually and experiment with different feature combinations on our training set using 5-fold cross-validation. The heuristic baseline sets a challenging accuracy of .776 and performs well for identifying major claims and premises (table 11). It significantly outperforms the majority baseline and achieves a F1-score of .740 for identifying major claims, .560 for claims and .870 for premises.

Table 11
Results of argument component classification using 5-fold cross-validation on our training set († significant improvement over the heuristic baseline (McNemar test with $p = 0.05$); Acc = Accuracy; F1 = Macro F1; P = Macro Precision; R = Macro Recall; MC = MajorClaim; Cl = Claim; Pr = Premise)

	Acc	*F1*	*P*	*R*	*F1 MC*	*F1 Cl*	*F1 Pr*
Baseline Majority	.627	.257	.209	.333	0	0	.771
Baseline Heuristic	.776	.724	.724	.723	.740	.560	**.870**
SVM only lexical	.655	.591	.603	.580	.591	.405	.772
SVM only structural †	.786	.746	.726	.767	.803	.551	**.870**
SVM only contextual	.709	.601	.603	.600	.656	.248	.836
SVM only indicators	.665	.508	.596	.443	.415	.098	.799
SVM only syntactic	.642	.387	.371	.405	.313	0	.783
SVM only probability	.652	.561	.715	.462	.448	.002	.792
SVM only discourse	.663	.521	.563	.484	.016	.538	.786
SVM only embeddings	.695	.588	.620	.560	.560	.355	.815
SVM all w/o prob & emb †	**.796**	.771	.771	**.772**	.855	**.596**	.863
SVM all w/o genre-dependent	.771	.742	.745	.739	.819	.560	.847
SVM all features †	.794	**.773**	**.774**	.771	**.865**	.592	.861

To analyze the influence of our features, we investigate each feature group individually. Structural features perform well for classifying argument components and are the most effective features for identifying major claims, since they encode if an argument component is present in the opening or closing paragraph of an essay. In addition, using only structural features significantly outperforms the heuristic baseline and yields a macro F1-score of .746 and thus an improvement of .022 over the heuristic baseline. Furthermore, we found that none of the remaining features significantly outperforms our heuristic baseline when employed without other features. Discourse features are the second best features for identifying claims (F1-score of .538). Therefore, we can confirm the assumption that general discourse relations are useful for identifying argument component types (Cabrio, Tonelli, and Villata 2013). Lexical features also perform reasonably well for identifying the type of argument components. They yield a macro F1-score of .591 and contribute to the identification of major claims (F1-score of .591). Although, word embeddings contribute to the identification of argument

component types, they do not perform as well as common lexical features. They yield lower F1-scores for major claims and claims, though they achieve better results for classifying premises. Since, our contextual features implicitly capture the occurrence of an argument component in the opening and closing paragraph by determining the shared noun and verb phrases (cp. section 5.3.1), they are effective for identifying major claims and also contribute to the identification of claims. Using only indicator features yields a F1-score of .508. They are effective for identifying major claims and slightly contribute to the identification of claims. Syntactic features contribute to the identification of major claims and premises only. The probability features yield a F1-score of .448 for identifying major claims and .799 for premises, but contribute only slightly to the identification of claims since forward indicators also signal the presence of premises if serial argumentation structures appear in a paragraph.

We also evaluate our system without genre-dependent features. In particular, we remove the structural features which indicate if an argument component is present in the opening or closing paragraph, since the major claim might be present somewhere else in other text genres. We also remove the shared noun and verb phrases, since those implicitly encode the same information. The resulting model yields a macro F1-score of .742 and no significant difference to the heuristic baseline on our training set which shows that our feature set successfully captures the characteristic of argument component types without using genre-dependent features.

By experimenting with various feature combinations, we found that omitting the probability and embedding features yields the best accuracy and the best F1-scores for claims and premises. However, we select the best performing system by means of the macro F1-score since it assigns equal weights to classes and not to individual instances which is a more appropriate measure for imbalanced data sets. Accordingly, we select the system which uses all features as our best performing system (table 11).

5.3.3 Error Analysis. In order to analyze frequent errors, we manually investigate the classification results and the confusion matrix of our best performing system (*SVM all features*). The confusion matrix (table 12) reveals that the most frequent confusion is between claims and premises. In particular, the system classifies 410 actual premises as claims and 422 claims as premises.

Table 12
Confusion matrix of the argument component classification determined with *SVM all features* on our training set using 5-fold cross-validation

		predictions		
		Major Claim	Claim	Premise
actual	Major Claim	**514**	80	4
	Claim	68	**712**	422
	Premise	8	410	**2,605**

By investigating the confusions of claims, we found that the errors are most often due to reasoning chains and co-occurring premises in the same sentence. First, the system tends to label premises which are part of a reasoning chain as claims since those are frequently signaled by claim indicating discourse connectives. For instance, the system wrongly classifies the second premise in the following paragraph as claim.

First of all, [students who study outside their countries can gain a lot of experience]$_{Claim}$. For example, [students might face many challenges in the host country]$_{Premise1}$. There-

fore, [they will learn to better overcome obstacles during their semester abroad]$_{Premise2}$. [Overcoming these problems teaches the students how to be more mature and confident]$_{Premise3}$.

Second, we observe several cases in which the classifier wrongly identifies claims in sentences including two premises. For instance, the classifier fails to resolve the required contextual information from surrounding argument components if a sentence includes two premises connected with discourse connectors like "*because*" or "*since*", and wrongly classifies the first premise as a claim.

The confusion matrix also shows that our model confuses major claims most frequently with claims. In particular, it wrongly classifies 68 claims as major claim and 80 actual major claims as claim. One cause of these confusions is that the classifier learns predominant patterns which frequently appear in persuasive essays. For instance, it is a common pattern to start the closing paragraph with an attacking claim before restating the major claim in the same sentence (cp. example essay in section 2.2). Therefore, the model tends to classify a claim followed by a major claim if the first sentence of the closing paragraph includes two argument components and no common indicators which signal the presence of other patterns. Similarly, the model tends to wrongly classify an argument component in the opening or closing paragraph as major claims if it includes references to the first person.

5.4 Argumentative Relation Identification

The relation identification model aims at recognizing argumentative relations between argument components. We model this task as a pair classification task and label each ordered pair of argument components in the same paragraph as argumentatively linked or not. Note that we consider both argumentative support and attack relations as argumentatively linked and that the distinction between support and attack relations will be described in section 5.6. The class distribution is skewed towards not argumentatively linked pairs (table 13). In our entire data set, 82.7% of the argument component pairs are not argumentatively linked and only 17.3% are argumentatively linked.

Table 13
Class distribution of the train and test set for argumentative relation identification.

	train	*test*
Not-Linked	14,227 (82.5%)	4,113 (83.5%)
Linked	3,023 (17.5%)	809 (16.5%)

Analogue to previous tasks, we define a heuristic baseline which exploits the structure of persuasive essays. Since essay writing guidelines recommend to state the claim before providing evidence in each body paragraph, our baseline classifies argument component pairs as argumentatively linked if both components appear in the same body paragraph and the target component is the first argument component of the paragraph. So this heuristic baseline correctly identifies convergent argumentation structures if the claim is the first argument component of a paragraph. However, it does not recognize serial arguments and fails if several arguments appear in the same paragraph. Note that in our data set, 62% of all body paragraphs start with a claim. We also employ a majority baseline which classifies each argument component pair as not argumentatively linked.

As the learner for our model, we employ an SVM implemented in the Weka machine learning framework (Hall et al. 2009). We found in our previous work that SVMs outperform other classifiers for argumentative relation identification (Stab and Gurevych 2014b).

5.4.1 Features. We use the following features for argumentative relation classification:

Lexical Features: We employ binary lemmatized unigrams of the source and target component to capture the lexical information of each argument component pair. Since the preceding tokens of each argument component can include discourse connectives or reasoning indicating shell language, we add all preceding tokens of the source and target component. In addition, we limit the number of the unigrams to the 500 most frequent words in our training data to prevent a too sparse feature set.

Syntactic Features: To capture the syntactic properties of the source and target components, we add binary features encoding the presence of all POS-tags generated by the Stanford pos-tagger. In addition, we employ the 500 most common production rules extracted from the constituent parse tree as described in our previous work (Stab and Gurevych 2014b, p. 50)

Structural Features: We add the number of tokens of the source and target component and a binary feature encoding if both argument components appear in the same sentence, since it is likely that those exhibit an argumentative relation. In addition, we add the number of argument components between the source and target component, a binary feature encoding if the target component appears before the source component and the number of argument components in the covering paragraph of the current pair. Since claims appear frequently as the first or last component in a paragraph, we add four binary features encoding if the source or target component is the first or last argument component of the covering paragraph. Two additional binary features encode if the current pair is present in the opening or closing paragraph of the essay.

Indicator Features: We employ the same set of indicators as used for argument component classification. In particular, we assume that those indicators are helpful for modeling the direction of argumentative relations and the local context of the current pair. We define eight binary features indicating if the current source or target component and their preceding tokens exhibit a forward, backward, thesis or rebuttal indicator. In order to model the local context of the current pair, we define four binary features encoding the presence of an indicator between the target and the source component and 16 features indicating if an indicator type is present preceding or following the source and target component.

Discourse Features: Although, the PDTB parser (Lin, Ng, and Kan 2014) is limited to adjacent discourse relations, we expect that the types of general discourse relations can be helpful for identifying argumentative relations. So we extract for each source and target component the type of the general discourse relation, if the component is the first or second text unit of the discourse relation and if the relation is implicit or explicit. Since the boundaries of the elementary discourse elements recognized by the PDTB-parser can differ from the boundaries of our argument component, we consider only those relations which overlap with the argument components of the current pair. Note that we also experimented with features capturing PDTB relations between

the target and source component. However, those were not effective for capturing argumentative relations.

PMI Features: As discussed above, argument components frequently exhibit preceding shell language which indicates the argumentative type of a component. We expect that these preceding tokens can also signal if an argument component has incoming or outgoing relations. For instance, argument components which exhibit preceding discourse connectors like "*therefore*", "*thus*" or "*hence*" are likely to exhibit incoming discourse relations, whereas discourse connectors like "*because*", "*since*" or "*furthermore*" can signal outgoing relations of the following argument component. To exploit this information from our training data complementary to our indicator features, we determine the *Pointwise Mutual Information* (PMI) between each word preceding an argument in its covering sentence and the direction and type of the relation. In particular, we determined the PMI that a word signals an incoming or outgoing relation. Consequently, we determine two different scores and add for each argument component 4 binary features which indicate if the component exhibits a positive or negative association with one of the two categories and add two additional features including the average PMI-score of the preceding tokens for the incoming and outgoing category.

Shared Noun Features (shNo): We expect that argument components are more likely connected if they share the same noun phrases. For instance in classical syllogisms, the noun phrases are shared between both premises and the claim (Govier 2010, p. 199). So we add a binary feature set to true, if the source and target component share any nouns, and the number of shared nouns to our feature set.

5.4.2 Results. Our heuristic baseline performs comparatively well for identifying argumentative relations. It yields a macro F1-score of $.660$ and thus outperforms the majority baseline by $.205$. However, since the heuristic baseline identifies only convergent arguments if the claim appears as the first argument component in a body paragraph, it fails to correctly identify serial arguments. Note that in our entire corpus 28.4% of all arguments include serial structures.

In order to analyze the effectiveness of each individual feature group, we report the results of a feature ablation test conducted with a 5-fold cross-validation on our training set (table 14), since none of the feature groups yields remarkable results when used individually. Removing structural features from our feature set yields the highest decrease of the macro F1-score. Therefore, we assume that these features are the most effective ones for identifying argumentative relations in persuasive essays. However, even without structural features, our system significantly outperforms the heuristic baseline by $.055$ macro F1-score.

The second and third most effective feature groups are indicator and PMI features. Both improve the macro F1-score when combining them with other features by $.014$ and $.013$ respectively. This result shows that lexical cues and preceding shell language play an important role for identifying argumentative relations and thus the structure of arguments. Syntactic and discourse features are not as effective as our indicator and pmi features. However, both contribute to the identification of argumentative relations and yield a slight improvement when combining them with other features. Removing the shared noun features does not yield a difference in accuracy or macro F1-score but results in a slightly improved F1-score of linked argument component pairs (F1 Linked).

Table 14
Results of argumentative relation identification using 5-fold cross-validation on the train set († = significant improvement over *Baseline Heuristic* (McNemar test with $p = 0.05$); Acc = Accuracy; F1 = Macro F1; P = Macro Precision; R = Macro Recall)

	Acc	*F1*	*P*	*R*	*F1 Not-linked*	*F1 Linked*
Baseline Majority	.835	.455	.418	.500	.910	0
Baseline Heuristic	.809	.660	.657	.664	.885	.436
SVM all w/o lexical †	**.860**	**.736**	.762	.711	**.917**	**.547**
SVM all w/o syntactic †	.859	.729	**.764**	.697	**.917**	.526
SVM all w/o structural †	.849	.715	.740	.692	.911	.511
SVM all w/o indicators †	.851	.719	.743	.697	.912	.520
SVM all w/o discourse †	.856	.732	.755	.709	.915	.540
SVM all w/o pmi †	.851	.720	.745	.697	.912	.521
SVM all w/o shNo †	.857	.733	.756	**.712**	.915	.545
SVM all w/o lexical & shNo †	.858	.734	.760	.710	.916	.543
SVM all features †	.857	.733	.756	.711	.915	.544

However, when removing the feature from our best performing system, we observe a decline of $.002$ macro F1 score (table 14). Therefore, we keep the shared noun feature in the feature set of our best performing system.

We achieve the best results by removing lexical features from our feature set. The system yields the best accuracy of $.860$ and a macro F1-score of $.736$. It also exhibits the highest score for linked and not linked argument component pairs. Note that increasing the number of lemma unigrams improved the accuracy only when using lexical features without other features. However, employing more lexical features did not improve the overall results when using a combined feature set. So we assume that using more lexical features introduces too much noise in the form of content-relevant words, which do not contribute to the task since our corpus exhibits a high variety of different topics.

5.4.3 Error Analysis. By analyzing the identified relations of our best performing system, we observe that our model identifies fewer linked argument component pairs than present in our data. In particular, the model identifies only $2{,}319$ linked pairs, although our training set includes $3{,}023$ linked argument component pairs. Consequently, the model does not recognize any relation in 15.7% of all paragraphs including at least one premise (Note that only paragraphs including a premise should include argumentative relations, since only premises are linked to other argument components; cp. section 2.2). On the other hand, the model does only recognize argumentative relations in 3.7% of the paragraphs not including any premise which shows that the model successfully identifies not linked argument component pairs.

Further, we observe that the results of the relation identification model strongly deviate from the targeted tree structures. First, as a consequence of the fewer identified argumentative relations, the model does not recognize an outgoing relation for 37.1% of all premises. In addition, the model identifies only for 55.6% of the premises exactly one outgoing relation and for 7.3% several outgoing relations. Second, the model recognizes only 20.9% valid trees in our training set and thus fails to identify the correct tree of 79.1% of all arguments. Although, these results are substantially different from our targeted argumentation structures, we show in the next section that the results of the relation identification model are valuable for identifying the targeted argumentation trees.

5.5 Jointly Modeling Argumentative Relations and Argument Component Types

Previous models identified particular properties of the argumentation structure independently. However, the classification of argument components and the identification of argumentative relations are closely related. For instance, knowing the type of argument components is a strong indicator for identifying argumentative relations and information about the argumentative structure facilitates the classification of argument components (Stab and Gurevych 2014b, p. 54). In particular, if an argument component is classified as a claim, it is less likely that it exhibits outgoing relations and more likely that it has incoming relations. On the other hand, the predicted argumentative relations can be exploited to infer information about the argument component types. For example, an argument component with several incoming and few outgoing relations is more likely to be a claim, whereas an argument component with few incoming relations is likely to be a premise. Therefore, it is reasonable to combine both types of information in order to find the tree structure which optimizes the results of the previous analysis steps.

We formalize this task as an Integer Linear Programming (ILP) problem. Given a paragraph including n argument components, we define the following objective function

$$\arg\max_{x} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_{ij} \tag{1}$$

with variables $x_{ij} \in \{0, 1\}$ representing an argumentative relation from argument component i to argument component j.[11] Each coefficient $w_{ij} \in \mathbb{R}$ is a weight for a relation and is determined by incorporating the results of previous analysis steps. In order to ensure that the resulting structure is a tree, we define the following constraints:

$$\forall i : \sum_{j=1}^{n} x_{ij} \leq 1 \tag{2}$$

$$\forall i : x_{ii} = 0 \tag{3}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \leq n - 1 \tag{4}$$

Since each premise should exhibit exactly one outgoing argumentative relation and claims do not have outgoing relations, equation 2 ensures that each argument component i has one or zero outgoing relations. Equation 3 prevents that an argumentative relation has the same source and target component. Since each paragraph with argumentative content needs to include a claim and there might also be cases where several arguments and therefore several claims occur in the same paragraph, equation 4 ensures that a paragraph includes at least one root node without outgoing relation. Finally, for

11 We use the lpsolve framework (http://lpsolve.sourceforge.net) and set each variable in our objective function to "*binary mode*" for ensuring the upper bound of 1.

preventing cycles, we follow the approach described by Kübler et al. (2008, p. 92) and include a set of auxiliary variables $b_{ij} \in \{0, 1\}$ in our objective function (1) where $b_{ij} = 1$ if there is a directed path from argument component i to argument component j and add the following three constraints:

$$\forall i \forall j : x_{ij} - b_{ij} \leq 1 \tag{5}$$

$$\forall i \forall j \forall k : b_{ik} - b_{ij} - b_{jk} \leq -1 \tag{6}$$

$$\forall i : b_{ii} = 0 \tag{7}$$

The first of these constraints ties the variables x_{ij} to the auxiliary variables b_{ij} and states that if there is a direct relation between the argument components i and j then there is also a path from i to j represented in variable b_{ij}. The second constraint covers all paths of length greater than 1 in a transitive way. It states that if there is a path from argument component i to j ($b_{ij} = 1$) and another path from argument component j to k ($b_{jk} = 1$) then there is also a path from argument component i to k. So, it iteratively covers paths of length $l + 1$ by having covered paths of length l. Finally, the third constraint restricts any cycle in the graph by preventing all directed paths starting and ending with the same argument component.

Having defined the ILP model, we consolidate the results of the argumentative relation identification and argument component classification models. We consider this task as determining the *weight matrix* $W \in \mathbb{R}^{n \times n}$ which includes the coefficients $w_{ij} \in W$ of our objective function for each argumentative relation. Note that this matrix can be considered as an adjacency matrix and that a greater weight for a particular relation denotes a higher likelihood that the relation is included in the optimal tree found by the ILP-solver.

We start by incorporating the results of the argumentative relation identification model whose result can be considered as an adjacency matrix $R \in \{0, 1\}^{n \times n}$. For each pair of argument components (i, j) with $0 < i \leq n$ and $0 < j \leq n$, each $r_{ij} \in R$ is 1 if the relation identification model predicts an argumentative relation from argument component i (source) to argument component j (target), or 0 if the model does not predict an argumentative relation. So, the first approach of determining the relation weights is using matrix R as weight matrix W without further adaptation. We refer to this approach as "*ILP-naïve*" and set $w_{ij}^{\text{(ILP-naïve)}} = r_{ij}$.

However, as mentioned in the beginning of this section, the results of the argumentative relation identification model bear more valuable information which can be exploited for determining more elaborated weights. For incorporating this information into the weight matrix W, we first determine for each argument component i the following *claim score* (cs) by means of the predicted structure represented in R:

$$cs_i = \frac{relin_i - relout_i + n - 1}{rel + n - 1} \tag{8}$$

where $relin_i = \sum_{k=1}^{n} r_{ki}[i \neq k]$ is the number of predicted incoming argumentative relations of argument component i, $relout_i = \sum_{l=1}^{n} r_{il}[i \neq l]$ is the number of predicted outgoing argumentative relations of argument component i and $rel = \sum_{k=1}^{n} \sum_{l=1}^{n} r_{kl}[k \neq l]$ is the total number of predicted relations in the given paragraph. Note that cs_i is

bigger for argument components with many incoming argumentative relations and fewer outgoing argumentative relations. It becomes smaller for argument components which exhibit less incoming and more outgoing argumentative relations. In addition, by normalizing the score with the total number of predicted relations and argument components, it also accounts for context information in the current paragraph and prevents over optimistic scores. For instance, if all the predicted argumentative relations point to an argument component i which has no outgoing relations, cs_i is exactly 1. On the other hand, if there is an argument component j with no incoming and one outgoing argumentative relation in a paragraph with 4 argument components and 3 predicted relations in R, cs_j is $\frac{1}{3}$. So, cs_i represents if an argument component i is a claim based on the predicted argumentative relations represented in the adjacency matrix R. Since, it is more likely that an argumentative relation links an argument component which has a lower claim score to an argument component with a higher claim score, we determine the weight for each potential argumentative relation as:

$$cr_{ij} = cs_j - cs_i \tag{9}$$

By adding the claim score cs_j of the target component j, we assign a higher weight to argumentative relations pointing to argument components which are likely to be a claim. Additionally, by subtracting the claim score cs_i of the source component i, we assign smaller weights to relations outgoing argument components with larger claim score. Accordingly, we define our second model as $w_{ij}^{(ILP\text{-}relation)} = \frac{1}{2}r_{ij} + \frac{1}{2}cr_{ij}$ and refer to it as "*ILP-relation*" since it uses only information from our argumentative relation identification model.

Next, we incorporate the predicted types of argument components. Since it is more likely that an argumentative relations points to a claim, we assign a higher score to the weight w_{ij} if the target component j is predicted as claim. Accordingly, we define our third model as $w_{ij}^{(ILP\text{-}claim)} = c_{ij}$ where $c_{ij} = 1$ if argument component j is predicted as claim and $c_{ij} = 0$ if argument component j is not predicted as claim. Note that we also experimented with subtracting the type information of the source component which didn't yield an improvement of the final model.

Finally, we combine the information of the argumentative relation identification and component classification model as

$$w_{ij} = \phi_r r_{ij} + \phi_{cr} cr_{ij} + \phi_c c_{ij} \tag{10}$$

and experiment with several proportions for each score. In particular, the "*ILP-equal*" model assigns $\phi_r = \phi_{cr} = \phi_c = \frac{1}{3}$ and thus uses an equal proportion of all scores. The "*ILP-same*" model uses the same amount of information from our two base classifiers which is realized by setting the coefficients to $\phi_r = \phi_{cr} = \frac{1}{4}$ and $\phi_c = \frac{1}{2}$. The "*ILP-balanced*" model balances the information for argument component types and relations by using $\phi_r = \frac{1}{2}$ and $\phi_{cr} = \phi_c = \frac{1}{4}$. Note that we incorporate our heuristic baselines in the weight calculation of all models if the base classifiers do neither recognize claims nor relations in a paragraph. In these cases we set $w_{i1} = 1$ for $1 < i \leq n$ and the remaining $w_{ij} = 0$. In order to evaluate the effect of the incorporated baseline, we investigate the results of the base classifiers incorporating the baseline and refer to it as "*IncBaseline*".[12]

12 Note that *IncBaseline* does not use the ILP-model but only incorporates the baseline in the results of our base classifiers if a paragraph does not exhibit predicted relations or claims but premises.

Finally, we adapt the argumentative relations and argument component types according to the results of the ILP-solver. In particular, we revise each relation according to the determined x_{ij} scores defined in our objective function, set the types of each root node of the identified trees to claim and the types of all remaining components in the tree to premise.

5.5.1 Results. We determine the best configuration of the joint model by conducting experiments on our training data before testing the best model on our test set. Note again, that our joint model does not incorporate major claims to prevent a too skewed class distribution in the argumentative relation identification task (cp. beginning of 5). However, in order to enable a realistic evaluation scenario, we rely on the major claim predictions of our argument component classification model and evaluate the argument component types including predicted major claims. Consequently, the upper bound of the argument component classification is $.934$ macro F1-score and $.996$ for the argumentative relation identification model since the false positives of major claims are excluded in the joint modeling approach and false negatives are included.

Applying the heuristic baseline to paragraphs in which the base classifiers neither identified argumentative relations nor claims (IncBaseline) yields a slight improvement of the argument component types and argumentative relations (table 15). In total, the heuristic baseline is triggered in 31 paragraphs and thus 31 premises are converted to claims. As a consequence, IncBaseline identifies 3.3% more correct trees than the base classifier. However, the difference between IncBaseline and the base classifiers is not statistically significant ($p = 0.05$). Therefore, we can attribute any further improvements to the joint modeling approach instead of the integration of the heuristic baseline.

Table 15
Results of the joint modeling approach. The table shows the results of the argumentative component classification and argumentative relation identification determined on the training set using 5-fold cross-validation († significant improvement over *baseline heuristic*; ‡ significant improvement over *base classifier*; F1 = Macro F1; MC = F1 Major Claim; Cl = F1 Claim; Pr = F1 Premise; NoLi = F1 Not-Linked; Link = F1 Linked; Cl $\rightarrow$ Pr = number of claims converted to premises; Pr $\rightarrow$ Cl = number of premises converted to claims; Trees = Percentage of correctly identified trees)

	component classification					*relation identification*			*statistics*		
		F1	MC	Cl	Pr		F1	NoLi	Link	Cl $\rightarrow$ Pr	Pr $\rightarrow$ Cl	Trees
Baseline heuristic		.724	.740	.560	.870		.660	.885	.436	-	-	100%
Base classifier	†	.773	**.865**	.592	.861	†	.736	.917	.547	-	-	20.9%
IncBaseline	†	.776	**.865**	.601	.861	†	.739	.917	.555	0	31	24.2%
ILP-naïve	†	.765	**.865**	.591	.761	†	.732	.918	.530	206	1,144	100%
ILP-relation	†‡	.809	**.865**	.677	.875	†‡	**.759**	**.919**	**.598**	299	571	100%
ILP-claim	†	.740	**.865**	.549	.777		.666	.894	.434	229	818	100%
ILP-equal	†‡	.822	**.865**	.699	.903	†	.751	.913	.590	294	280	100%
ILP-same	†‡	.817	**.865**	.687	.898	†	.738	.908	.569	264	250	100%
ILP-balanced	†‡	**.823**	**.865**	**.701**	**.904**	†	.752	.913	.591	297	283	100%

Using only predicted relations in the ILP-naïve model does not yield an improvement over the base classifiers. Since a great many of argument components are not linked by the base classifier, the model also converts $1,144$ premises to claims and

thus identifies 78% more claims than present in our training data.[13] Combining the claim score with the predicted relations in the ILP-relation model leads to a significant improvement of both tasks over their base classifiers ($p = 0.05$). The argument component classification improves by .036 macro F1-score and the argumentative relation identification by .023 macro F1-score. In particular, this model yields a considerable improvement of .085 F1-score for the claim identification and yields the highest score for identifying argumentative relations. This result shows that the predicted argumentative relations and our defined claim score are valuable for the component classification task. However, the model still converts a high number of premises to claims and thus identifies 22.46% more claims than present in our training data. Incorporating only predicted claims in the ILP-claim model is detrimental for both tasks. Without the predicted relations a great many of relation weights in W are 0 which induces a conversion of 818 premises to claims and thus the identification of 49.1% more claims. In addition, the model identifies 19.22% less argumentative relations than present in our training set caused by the fewer number of premises, and the relation identification exhibits no significant difference compared to the heuristic baseline.

Combining the results of the component classification model and the argumentative relation identification model yields a considerably more balanced proportion of argument component type conversions compared to the other models. On average, the ILP-equal, ILP-same and ILP-balanced model identify only 1.16% fewer claims and consequently 0.73% more argumentative relations compared to the total amount in our training data. Therefore, the combination of all scores is more accurate compared to the integration of individual base classifier results. Although, all three models lead to a significant improvement of the component classification task over the base classifier, none of the three models significantly outperforms the base classifier for argumentative relation identification though ILP-balanced improves the macro F1-score by .016.

We identify the best performing system by the average macro F1-score of the argument component classification and argumentative relation identification task. Accordingly, we select ILP-balanced as our best performing system. It achieves a macro F1 score of .823 and .752 for the classification of argument components and the identification of argumentative relations. In particular, it improves the F1-score for identifying claims by .109 and the F1-score for related argument components by .043 over the base classifiers. This indicates that jointly modeling argument components and argumentative relations considerably improves the performance and additionally leads to the correct identification of the targeted tree structures.

Table 16 shows the results of the ILP-balanced model on our test set. Similar to previous results, the model significantly outperforms the base classifier of the argument component classification. The argumentative relation identification does not yield a significant improvement over the base classifier though the model yields an improvement of .036 macro F1-score. However, the joint modeling approach significantly outperforms the heuristic baselines of both tasks though the baseline yields better results on our gold test set than on our training set since the training data exhibits 7.5% more body paragraphs starting with a claim. The results also show that the identification of claims and linked argument component pairs benefit most by the joint modeling approach. In particular, the ILP-balanced model yields .071 better F1 score for claims and .077 better F1 score for linked argument component pairs.

13 Note that the model anyhow identifies 100% correct trees, since we consider also claims without linked premises as valid trees (forests).

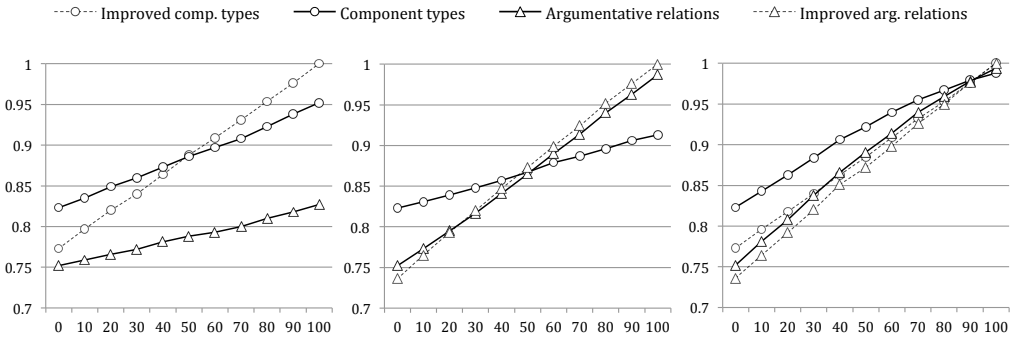
Table 16
Results of the joint model on the test set including the human upper bound for argument component classification and argumentative relation identification († significant improvement over *baseline heuristic*; ‡ significant improvement over *base classifier*; F1 = Macro F1; P = Macro Precision; R = Macro Recall; MC = F1 Major Claim; Cl = F1 Claim; Pr = F1 Premise; NoLi = F1 Not-Linked; Link = F1 Linked)

	argument component classification						*argumentative relation identification*					
		F1	P	R	MC	Cl	Pr		F1	P	R	NoLi	Link
Human Upper Bound		.868	.870	.866	.926	.754	.924		.854	.853	.856	.954	.755
Baseline Heuristic		.759	.761	.758	.759	.620	.899		.700	.700	.700	.901	.499
Base classifier		.794	.793	.796	**.891**	.611	.879	†	.717	.745	.692	.917	.508
ILP-balanced	†‡	**.826**	**.824**	**.827**	**.891**	**.682**	**.903**	†	**.751**	**.750**	**.752**	**.918**	**.585**

We determine the human upper bound for both tasks by comparing all pairs of annotators and averaging the resulting scores. Since, the boundaries of argument components can differ between the annotators, we consider only maximum overlapping annotations for determining the human upper bound. Compared to the joint modeling approach, the macro F1-score of the human performance is $.042$ higher and exhibits a considerably better score for claims (table 16). In contrast to the argument component classification task, the human performance of the argumentative relation identification is considerably better than the joint model. Human annotators achieve a macro F1-score of $.854$ which is 0.103 higher compared to our joint model. In total, our system achieves 95.2% and 87.9% of the human performance for argument component classification and argumentative relation identification respectively (macro F1-score).

5.5.2 Error Analysis and Influence of Base Classifiers. We observe that the model tends to identify more shallow trees compared to our gold test set. In particular, the model correctly identifies only 34.7% of the 98 serial arguments in our gold test set which can be attributed to the claim-centered weight calculation in our objective function. In particular, only the predicted relations in the adjacency matrix R include information about serial arguments if the argumentative model correctly classifies serial structures whereas the other two scores (c and cr) primarily assign higher weights to relations pointing to claims. We also observe that the model identifies 42.5% fewer paragraphs including several claims and thus several arguments.

In order to further analyze the approach, we simulate the effects of improving the base classifiers analogue to the approach presented by Peldszus and Stede (2015). The dashed lines in figure 4 show the performance of the artificially improved base classifiers and continuous lines show the resulting performance of the argumentative relation identification and argument component classification after applying the joint modeling approach (ILP-balanced).
Figure 4a+b depicts the effect of improving the argument component type base classifier and the argumentative relation base classifier respectively. It shows that correct results of one base classifier are not maintained after applying the ILP-model if the other base classifier exhibits less accurate predictions. In particular, a less accurate prediction of argumentative relations has a more detrimental effect on the argument component types (figure 4a) than less accurate argument component types on the outcomes of the argumentative relation identification (figure 4b). Figure 4c depicts the effect of improving both base classifiers which illustrates that the joint modeling approach benefits



(a) Improve Types (b) Improve Relations (c) Improve Both

Figure 4
Influence of improving the base classifiers (ILP-balanced model; 5-fold cross-validation on the training set). The x-axis shows the percentage of improved predictions and the y-axis the macro F1-score. Dashed lines show the artificially improved base classifiers; (a) illustrates the effect of improving the argument component types; (b) the improvement of argumentative relations and (c) the improvement of both base classifiers.

the component type classification more than the argumentative relation identification. However, since accurate predictions of both argument component types and argumentative relations positively influence the outcomes of the other task, we conclude that the joint modeling approach successfully models the dependency between both types of information.

5.6 Stance Recognition

The last step includes the discrimination of supporting and attacking argumentative relations. We model this task as a binary classification of the source components of each argumentative relation, since we expect that the source component exhibits more substantial information for differentiating between supporting and attacking relations than the target component. So we classify each premise and claim as either support or attack. Note that the stance of each premise is encoded in the type of its outgoing argumentative relation whereas the stance of each claim is encoded in its stance attribute. We evaluate the stance recognition using the gold argument component types. Thus the number of all instances slightly differs from the real application scenario since we assume correctly identified major claims which are not included in this task.

Table 17 shows the class distribution of the train and test set. Since authors tend to support their own view instead of providing opposing arguments (Wolfe and Britt 2009), the class distribution is heavily skewed towards attack relations. In total, our data set includes 90.7% support relations and only 9.3% attack relations.

Table 17
Class distribution of the train and test set for stance recognition task.

	train	*test*
Support	3,820 (90.4%)	1,021 (91.7%)
Attack	405 (9.6%)	92 (8.3%)

Since essay writing guidelines recommend to include opposing arguments in the second last paragraph and it is likely that authors also include opposing reasons to defend their standpoint close to them, we define a heuristic baseline which classifies each argument component in the second last paragraph as attack. In addition, we employ a majority baseline which classifies each argument component as support.

5.6.1 Features. For the stance recognition task, we employ the following features:

Lexical Features: We use binary and lemmatized unigram features and also consider the preceding tokens of each argument component since it is likely that those include clue words or phrases which are valuable for detecting the stance.

Sentiment Features: In order to identify the polarity of each argument component including its preceding tokens, we employ the subjectivity lexicon provided by Wilson, Wiebe, and Hoffmann (2005). In particular, we define one binary feature which indicates the presence of negative words and the number of words having a negative, positive and neutral polarity. In addition, we determine the overall polarity of each argument component by adding the number of positive words and subtracting the number of negative words in an additional numeric feature. Complementary, we add five sentiment scores of the covering sentence of each argument component determined with the Stanford Sentiment Analyzer (Socher et al. 2013). These include scores for very negative, negative, neutral, positive and very positive.

Syntactic Features: In order to capture the syntactic characteristic of argument components, we add the distribution of POS-tags of each argument component and the production rule features described previously.

Structural Features: We use the number of argument components in the covering paragraph, the number of tokens of the covering sentence, the ratio between the sentence and component tokens, and the number of preceding and following tokens of the component in its covering sentence as structural features. In addition, we employ the relative position of the argument component and the number of preceding and following argument components in the covering paragraph in order to capture common linearization strategies of arguments.

Discourse Features: We employ the same discourse features as for the component classification task. Since PDTB also includes contrast and concession relations, we expect that these features will be helpful for identifying attacking components.

Word Embedding Features: We employ the same word embeddings trained on a part of the Google News data set (Mikolov et al. 2013) as for the argument component classification task.

5.6.2 Results. Table 18 shows the results of the heuristic and majority baselines determined with 5-fold cross-validation on the training set. The heuristic baseline achieves a macro F1-score of .521 and thus outperforms the majority baseline by .046. In order to find the best learner, we compared Naïve Bayes (John and Langley 1995), Random Forests (Breiman 2001), Multinomial Logistic Regression (le Cessie and van Houwelingen 1992), C4.5 Decision Trees (Quinlan 1993) and SVMs (Cortes and Vapnik 1995). We

found that the latter considerably outperforms all other classifiers. Therefore, we report the results of the feature analysis using the SVMs only.

Table 18
Feature analysis of the stance recognition model conducted with 5-fold cross-validation on the train set († = significant improvement over *Baseline Heuristic*; ‡ significant difference compared to *SVM all features*; Acc = Accuracy; F1 = Macro F1; P = Macro Precision; R = Macro Recall)

	F1	*P*	*R*	*F1 Support*	*F1 Attack*
Baseline Majority	.475	.452	.500	.950	0
Baseline Heuristic	.521	.511	.530	.767	.173
SVM only lexical †	.663	.677	.650	.941	.383
SVM only syntactic †	.649	.725	.587	.950	.283
SVM only discourse †	.630	**.746**	.546	**.951**	.169
SVM all w/o lexical †	.696	.719	.657	.948	.439
SVM all w/o syntactic †‡	.687	.691	.684	.941	.433
SVM all w/o sentiment †	.699	.710	.688	.945	.451
SVM all w/o structural †	.698	.710	.686	.946	.449
SVM all w/o discourse †‡	.675	.685	.666	.941	.408
SVM all w/o embeddings †	.692	.703	.682	.944	.439
SVM all features †	**.702**	.714	**.690**	.946	**.456**

Using sentiment, structural and embedding features individually does not yield any improvement over the majority baseline. The best individual feature group are lexical features which significantly outperform the heuristic baseline when used individually ($p = 0.05$). By ranking the lexical features using information gain, we found that the best ranked words primarily include terms like "*although*", "*however*", "*though*", "*admittedly*" and "*oppose*" which all signal attacking argument components. Syntactic features also perform well when employed without other features. They significantly outperform the heuristic baseline by $.128$ macro F1-score. The results also confirm our assumption that discourse features contribute to the identification of the argument components' stances.

We identified the best performing system by conducting feature ablation tests and found that omitting any of the feature groups yields worse results than a combination of all features. Interestingly, omitting lexical features does not exhibit a significant difference although they perform best when used individually (table 18). We speculate that this effect is due to the multiplicity of content words which are uninformative given the high topic variety of our corpus. The best performing system using all features achieves a macro F1-scores of $.702$ and a F1-score of $.946$ and $.456$ for supporting and attacking components respectively. Thus the system successfully identifies supporting components. The lower score for attacking components can be primarily attributed to the skewed class distribution in our data set (table 17).

Table 19 shows the results of our best performing system on the test set. It achieves comparable results as on our training data and also significantly outperforms the heuristic baseline ($p = 0.05$). We also evaluated the results of claims and premises independently and found that the system yields better results for claims. In particular, the macro F1-score for identifying the stance of claims is $.694$ and for premises $.655$. The human upper bound exhibits a considerably higher F1-score for attacking components and a macro F1-score of $.844$. However, our system achieves 80.6% of human performance with respect to the macro F1-score.

5.6.3 Error Analysis. By manually analyzing the classification results, we observe that the false positives of attacking components are frequently due to a misleading use of clue phrases. For instance, the system tends to classify argument components as

Table 19
Results of the stance recognition model on the test set including the comparison to human upper bound († = significant improvement over *Baseline Heuristic*; Acc = Accuracy; F1 = Macro F1; P = Macro Precision; R = Macro Recall)

	F1	*P*	*R*	*F1 Support*	*F1 Attack*
Human Upper Bound	.844	.874	.816	.975	.703
Baseline Majority	.478	.459	.500	**.957**	0
Baseline Heuristic	.562	.532	.596	.776	.201
SVM all features †	**.680**	**.680**	**.680**	.947	**.413**

attack if they include phrases like "*although*", "*however*" or "*nevertheless*" though they are actually meant to support another argument component. In addition, the system wrongly classifies supporting components as attacking if the essay exhibits a strong opposition against the topic. In particular, if the author strongly disagrees with the prompt of an essay, the reasons given frequently include negative terms and also attack indicating clue phrases. So for future work, it might be helpful to incorporate the overall polarity of an essay and to consider contextual information of argument components to improve the accuracy of the classifier.

Among the false negatives, we observe several cases which include references to the first person like "*I*", "*my*" or "*myself*". Usually, these cases represent rebuttals of contra arguments meant to support the stance of the author on the topic. For instance, these cases include phrases like "*In my opinion*", "*I believe*" or "*from my viewpoint*" without including attacking clue words. Since, these phrases usually occur in supporting argument component, it is hard to correctly recognize them as attack. Additionally, some false negatives are due to missing clue phrases for attacking components. In these cases, the coupling of content words with subjectivity clues as proposed by Somasundaran and Wiebe (2009) could be useful. However, it would require a corpus with less heterogeneous topics in order to identify content words which are generally negatively or positively related to a particular topic.

5.7 Evaluation on argumentative microtexts

To verify the effectiveness of our joint modeling approach and to compare its performance to previous work, we also evaluate our system on the English version of the microtext corpus created by Peldszus and Stede (2015).[14]

The first three rows in table 20 show the results of the base classifiers (simple), the best system for argument component classification and stance recognition (Best EG), and the best system for argumentative relation identification (MP+p) reported by Peldszus and Stede (2015). Although, none of these models achieve the best results for all tasks, both of their joint models improve the results of the component and relation base classifiers. However, their best model for argumentative relation identification (MP+p) yields considerably lower results for stance classification compared to their base classifier.

14 Note that we adapted all features of our two base classifiers which require non-argumentative tokens preceding argument components, since the corpus does not include non-argumentative text units. For the same reason, we did not evaluate our segmentation model on this corpus. In addition, we removed the major claim from the component classification model since the corpus includes only claims and premises. Also note, that we employed the same evaluation setup as described by Peldszus and Stede (2015).

Table 20
Results of the ILP joint model on the microtext corpus compared to the results of the MST-style joint model proposed by Peldszus and Stede (2015). The first three rows show the results of the base classifiers (simple), the best performing system for argument component classification (Best EG) and the best performing system for argumentative relation identification (Mp+p) reported by Peldszus and Stede (2015). The following two rows show the results of our approach. (F1 = Macro F1; C = Claim; P = Premise; N = Not-Linked; L=Linked; S = Support; A = Attack)

	Components			*Relations*			*Stance*			
	F1	*F1 C*	*F1 P*	*F1*	*F1 N*	*F1 L*	*F1*	*F1 S*	*F1 A*	*avg*
Simple	.817	?	?	.663	?	.478	.671	?	?	.717
Best EG	**.869**	?	?	.693	?	.502	.710	?	?	.757
MP+p	.831	?	?	**.720**	?	.546	.514	?	?	.688
Base classifier	.830	.712	.937	.650	.841	.446	**.745**	.855	.628	.742
ILP Joint Model	.857	.770	.943	.683	.881	.486	**.745**	.855	.628	**.762**

Compared to the base classifiers of Peldszus and Stede (2015), our component classification model yields better results. Furthermore, our stance recognition model considerably outperforms all approaches reported by Peldszus. Although, we can confirm that our joint model successfully improves the results of our two base classifiers, neither the relation identification nor the component classification outperforms the results of Peldszus' Best EG Model. This difference can be mainly attributed to the additional information about the stance and the role attribute (cp. section 3.1.2) incorporated in Peldszus' MST-model. They showed that both have a beneficial effect on the component classification and relation identification in their corpus (Peldszus and Stede 2015, figure 3). However, the role attribute is a unique feature of their data set and as discussed in section 3.1.1, the arguments in their corpus include an unusually high proportion of attacking components. In particular, 86.6% of their arguments include attack relations whereas real arguments usually exhibit few attacking components which is confirmed by only 12.4% of arguments including attack relations in our corpus. We assume that this proportion might be even lower in other text genres, since essay writing guidelines also encourage students to include opposing arguments in their writing. Therefore, it is unlikely that incorporating stance and role attributes will have the same effect using real data.

On average, our approach slightly outperforms the best model presented by Peldszus and Stede (2015) due to the better performance of our stance recognition model. Since our model also simultaneously improves the results of the component classification and relation identification base classifiers, we conclude that our approach successfully models the natural dependency between argument component types and argumentative relations and represents a robust model for identifying argumentation structures.

6. Discussion

Our argumentation structure parser includes several consecutive steps. Consequently, potential errors of the upstream models can negatively influence the results of the downstream models. In particular, errors of the segmentation model might result in flawed overall results if argumentatively relevant text units are not recognized or non-argumentative text units are identified as relevant. However, our segmentation model yields a good accuracy and an α_U of $.958$ for identifying argument components.

Therefore, it is unlikely that segmentation errors will significantly influence the outcome of the upstream components when applied to persuasive essays. However, as demonstrated by e.g. Levy et al. (2014) and Goudas et al. (2014) the identification of argument components is significantly harder in other text genres which do not exhibit a common structure. Another potential issue of the presented pipeline architecture is the identification of major claims which we didn't incorporate in the joint modeling approach. Similarly to wrongly identified argument components, misclassified major claims will decrease the accuracy of the joint model which could lead to worse overall results. So in future work, it would be worthwhile to experiment with structured machine learning methods to incorporate several tasks in one model as suggested by Moens (2013).

In this work, we have demonstrated that our annotation scheme can be reliably applied to persuasive essays. However, since persuasive essays exhibit a common structure, applying the scheme to other text genres like news articles, user generated web data or scientific articles might be more challenging. In particular, we expect that other text genres exhibit many more non-argumentative text units. Nevertheless, we believe that our annotation scheme can be applied with minor adaptations to other text genres. Although, other text genres might not include major claims, previous work has already demonstrated that claims and premises can also be reliably annotated in legal cases (Mochales-Palau and Moens 2011), written dialogs (Biran and Rambow 2011) and even over multiple wikipedia articles (Aharoni et al. 2014). In addition, it has to be studied if our tree assumption generalizes to other text genres. Although, most previous work considered argumentation structures as trees, other text genres might include significantly more divergent arguments and even cyclic argumentation structures which would pose additional requirements to joint modeling approaches.

Although, our approach shows promising results, it is still an open question if the identified argumentation structures can be successfully exploited to provide adequate feedback about argumentation and to improve the argumentation skills of students. In general, the identified argumentation structures enable various kinds of feedback about the argumentation. For instance, it enables to recommend more meaningful and comprehensible argumentation structure. In particular, the extracted structure can be exploited to prevent several reasoning directions (forward and backward reasoning) in a single argument which might lead to a more comprehensible reasoning structure. It could be also used to highlight unsupported claims and to ask the author for additional reasons. Additionally, the identified argumentation structure enables the recommendation of additional discourse markers to make the arguments more coherent or to encourage authors to discuss opposing views. Finally, the visualization of the identified argumentation structure could stimulate self reflection and plausibility checking of the written arguments. However, finding adequate feedback types and investigating their effect on the argumentation skills of students requires the integration of the models in writing environments and extensive long term user studies in future work.

7. Conclusion

We presented the first end-to-end approach for parsing argumentation structures in persuasive essays. Previous approaches suffer from several limitations: existing approaches either focus only on particular subtasks of argument structure parsing and are consequently not capable of extracting argumentation structures in real application scenarios or rely on manually created rules and thus are not capable to identify implicit argumentation structures. To the best of our knowledge and to the date of writing, the presented work is the first approach which covers all required subtasks for identifying

the global and fine-grained argumentation structure of entire documents. We showed that jointly modeling argument component types and argumentative relations not only optimizes the results of base classifiers to identify the targeted argumentation structures but also simultaneously improves the results of argument component classification and argumentative relation identification. In addition, we introduced a novel annotation scheme and a new corpus of persuasive essays annotated with argumentation structures which represents to the best of our knowledge the largest resource of its kind. For encouraging future research, we will make the corpus as well as the annotation guideline freely available.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806 and by the German Federal Ministry of Education and Research (BMBF) as a part of the Software Campus project AWS under grant No. 01|S12054. We thank Can Diehl, Ilya Kuznetsov and Anshul Tak for their valuable contributions as well as Andreas Peldszus for providing details about his corpus.

References

[Aharoni et al.2014]Aharoni, Ehud, 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, Baltimore, MD, USA.

[Amgoud, Maudet, and Parsons2000]Amgoud, Leila, Nicolas Maudet, and Simon Parsons. 2000. Modelling dialogues using argumentation. In *Proceedings of the 4th International Conference on Multi-Agent Systems*, pages 31–38, Boston, MA, USA.

[Anand et al.2011]Anand, Pranav, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 1–9, Portland, OR, USA.

[Artstein and Poesio2008]Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

[Attali and Burstein2006]Attali, Yigal and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment (JTLA)*, 4(3):1–30.

[Attali, Lewis, and Steier2013]Attali, Yigal, Will Lewis, and Michael Steier. 2013. Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, 30(1):125–141.

[Beardsley1950]Beardsley, Monroe C. 1950. *Practical Logic*. Prentice-Hall.

[Beigman Klebanov and Flor2013]Beigman Klebanov, Beata and Michael Flor. 2013. Argumentation-Relevant Metaphors in Test-Taker Essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, GA, USA.

[Beigman Klebanov and Higgins2012]Beigman Klebanov, Beata and Derrick Higgins. 2012. Measuring the use of factual information in test-taker essays. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 63–72, Montreal, Quebec, Canada.

[Bentahar, Moulin, and Bélanger2010]Bentahar, Jamal, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

[Biran and Rambow2011]Biran, Or and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 05(04):363–381.

[Bohnet et al.2013]Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

[Boltužić and Šnajder2014]Boltužić, Filip 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, Baltimore, MA, USA.

[Botley2014]Botley, Simon Philip. 2014. Argument structure in learner writing: a corpus-based analysis using argument mapping. *Kajian Malaysia*, 32(1):45–77.
[Braud and Denis2014]Braud, Chloé and Pascal Denis. 2014. Combining natural and artificial examples to improve implicit discourse relation identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1694–1705, Dublin, Ireland.
[Breiman2001]Breiman, Leo. 2001. Random forests. *Machine Learning*, 45(1):5–32.
[Burstein, Chodorow, and Leacock2004]Burstein, Jill, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25(3):27–36.
[Burstein and Wolska2003]Burstein, Jill and Magdalena Wolska. 2003. Toward evaluation of writing style: finding overly repetitive word use in student essays. In *Proceedings of the tenth conference of European chapter of the Association for Computational Linguistics*, EACL '03, pages 35–42, Budapest, Hungary.
[Butler and Britt2011]Butler, Jodie A. and M. Anne Britt Britt. 2011. Investigating instruction for improving revision of argumentative essays. *Written Communication*, 28(1):70–96.
[Cabrio, Tonelli, and Villata2013]Cabrio, Elena, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In *Computational Logic in Multi-Agent Systems*, volume 8143 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 1–17.
[Cabrio and Villata2012a]Cabrio, Elena 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 (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea.
[Cabrio and Villata2012b]Cabrio, Elena and Serena Villata. 2012b. Natural language arguments: A combined approach. In *Proceedings of the 20th European Conference on Artificial Intelligence*, ECAI '12, pages 205–210, Montpellier, France.
[Carletta1996]Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.
[Carlson, Marcu, and Okurowski2001]Carlson, Lynn, 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.
[Carstens and Toni2015]Carstens, Lucas and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO, USA.
[Cinková, Holub, and Kríž2012]Cinková, Silvie, Martin Holub, and Vincent Kríž. 2012. Managing uncertainty in semantic tagging. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 840–850, Avignon, France.
[Cohen1960]Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.
[Cohen1987]Cohen, Robin. 1987. Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13(1-2):11–24.
[Collins2002]Collins, Michael. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Pennsylvania, PA, USA.
[Collins2003]Collins, Michael. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.
[Conway1991]Conway, David A. 1991. On the distinction between convergent and linked arguments. *Informal Logic*, 13:145–158.
[Copi and Cohen1990]Copi, Irving M. and Carl Cohen. 1990. *Introduction To Logic*. Macmillan Publishing Company, 8th edition.
[Cortes and Vapnik1995]Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.
[Dagan et al.2013]Dagan, Ido, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

[Damer2009]Damer, T. Edward. 2009. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Reasoning*. Wadsworth Cengage Learning, 6th edition.
[Davies2009]Davies, Peter. 2009. Improving the quality of students' arguments through 'assessment for learning'. *Journal of Social Science Education (JSSE)*, 8(2):94–104.
[Daxenberger et al.2014]Daxenberger, Johannes, 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.
[Dung1995]Dung, Phan Minh. 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.
[Eckart de Castilho and Gurevych2014]Eckart de Castilho, Richard and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In Nancy Ide and Jens Grivolla, editors, *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland.
[Feng and Hirst2011]Feng, Vanessa Wei 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.
[Feng and Hirst2014]Feng, Vanessa Wei and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, MA, USA.
[Fernández-Delgado et al.2014]Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181.
[Fleiss1971]Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
[Florou et al.2013]Florou, Eirini, 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.
[Forman and Scholz2010]Forman, George and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, 12(1):49–57.
[Freeman2011]Freeman, James B. 2011. *Argument Structure: Representation and Theory*, volume 18 of *Argumentation Library*. Springer.
[Ghosh et al.2014]Ghosh, Debanjan, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, MA, USA.
[Goudas et al.2014]Goudas, Theodosis, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, volume 8445 of *Lecture Notes in Computer Science*. Springer International Publishing, pages 287–299.
[Govier2010]Govier, Trudy. 2010. *A Practical Study of Argument*. Wadsworth, Cengage Learning, 7th edition.
[Habernal and Gurevych2015]Habernal, Ivan 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, Lisbon, Portugal.
[Hall et al.2009]Hall, Mark, 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.
[Hasan and Ng2012]Hasan, Kazi Saidul and Vincent Ng. 2012. Predicting stance in ideological debate with rich linguistic knowledge. In *Proceedings of COLING 2012: Posters*, pages 451–460, Mumbai, India.
[Hasan and Ng2014]Hasan, Kazi Saidul and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar.

[Hastings1963]Hastings, Arthur C. 1963. *A Reformulation of the Modes of Reasoning in Argumentation*. Ph.D. thesis, Evanston, Illinois.
[Henkemans2000]Henkemans, A. Francisca Snoeck. 2000. State-of-the-art: The structure of argumentation. *Argumentation*, 14(4):447–473.
[Hernault et al.2010]Hernault, Hugo, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.
[Indrajani and Angeline2010]Indrajani, Nani T. and Anggie Angeline. 2010. The types of argument structure used by hillary clinton in the cnn democratic presidential debate. *k@ta*, 11(2):184–200.
[John and Langley1995]John, George H. and Pat Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, Montreal, Quebec, Canada.
[Johnson2000]Johnson, Ralph H. 2000. *Manifest rationality*. Lawrence Erlbaum.
[Kemper and Sebranek2004]Kemper, Dave and Pat Sebranek. 2004. *Inside Writing: Persuasive Essays*. Great Source Education Group.
[Kirschner, Eckle-Kohler, and Gurevych2015]Kirschner, Christian, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO, USA.
[Klein and Manning2003]Klein, Dan 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.
[Krippendorff1980]Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. Sage.
[Krippendorff2004]Krippendorff, Klaus. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Quality & Quantity*, 38(6):787–800.
[Kübler et al.2008]Kübler, Sandra, Ryan McDonald, Joakim Nivre, and Graeme Hirst. 2008. *Dependency Parsing*. Morgan and Claypool Publishers.
[Kwon et al.2007]Kwon, Namhee, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, pages 76–81, Philadelphia, PA, USA.
[Lafferty, McCallum, and Pereira2001]Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA.
[le Cessie and van Houwelingen1992]le Cessie, S. and J.C. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.
[Levy et al.2014]Levy, Ran, 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, Dublin, Ireland.
[Lin, Kan, and Ng2009]Lin, Ziheng, 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, Suntec, Singapore.
[Lin, Ng, and Kan2014]Lin, Ziheng, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.
[Lippi and Torroni2015]Lippi, Marco and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 185–191, Buenos Aires, Argentina.
[Louis et al.2010]Louis, Annie, 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.
[MacKenzie1981]MacKenzie, Jim D. 1981. The dialectics of logic. *Logique et Analyse*, 94:159–177.
[Madnani et al.2012]Madnani, Nitin, Michael Heilman, Joel Tetrault, and Martin Chodorow. 2012. Identifying High-Level Organizational Elements in Argumentative Discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, NAACL HLT '12, pages 20–28, Montreal, Quebec, Canada.

[Mann and Thompson1987]Mann, William C. and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, Information Sciences Institute.

[Marcu and Echihabi2002]Marcu, Daniel 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.

[McNemar1947]McNemar, Quinn. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

[Meyer et al.2014]Meyer, Christian M., Margot Mieskes, Christian Stab, and Iryna Gurevych. 2014. Dkpro agreement: An open-source java library for measuring inter-rater agreement. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations (COLING)*, pages 105–109, Dublin, Ireland.

[Mikolov et al.2013]Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., pages 3111–3119.

[Mochales-Palau and Ieven2009]Mochales-Palau, Raquel and Aagje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments?: A case study on the legal argumentation of the ECHR. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL '09)*, pages 21–30, Barcelona, Spain.

[Mochales-Palau and Moens2007]Mochales-Palau, Raquel and Marie-Francine Moens. 2007. Study on sentence relations in the automatic detection of argumentation in legal cases. In *Proceedings of the 2007 Conference on Legal Knowledge and Information Systems: JURIX 2007: The Twentieth Annual Conference*, pages 89–98, Leiden, Netherlands.

[Mochales-Palau and Moens2009]Mochales-Palau, Raquel 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, Barcelona, Spain.

[Mochales-Palau and Moens2011]Mochales-Palau, Raquel and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

[Moens2013]Moens, Marie-Francine. 2013. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *Post-proceedings of the Forum for Information Retrieval Evaluation (FIRE 2013)*, pages 4–6, New Delhi, India.

[Moens et al.2007]Moens, Marie-Francine, 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, CA, USA.

[Nguyen and Litman2015]Nguyen, Huy and Diane Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO, USA.

[Okazaki2007]Okazaki, Naoaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/.

[Oraby et al.2015]Oraby, Shereen, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. 2015. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 116–126, Denver, CO, USA.

[Park and Cardie2014]Park, Joonsuk and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, MA, USA.

[Peldszus2014]Peldszus, Andreas. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97, Baltimore, MA, USA.

[Peldszus and Stede2013a]Peldszus, Andreas and Manfred Stede. 2013a. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

[Peldszus and Stede2013b]Peldszus, Andreas and Manfred Stede. 2013b. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria.

[Peldszus and Stede2015]Peldszus, Andreas and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 938–948, Lisbon, Portugal.
[Persing and Ng2013]Persing, Isaac and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria.
[Persing and Ng2014]Persing, Isaac and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, MA, USA.
[Persing and Ng2015]Persing, Isaac and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China.
[Perutz2010]Perutz, Vivien. 2010. *A Helpful Guide to Essay Writing!* Student Services, Anglia Ruskin University.
[Pitler, Louis, and Nenkova2009]Pitler, Emily, 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*, pages 683–691, Suntec, Singapore.
[Prasad et al.2008]Prasad, Rashmi, 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.
[Qiu and Jiang2013]Qiu, Minghui and Jing Jiang. 2013. A latent variable model for viewpoint discovery from threaded forum posts. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040, Atlanta, Georgia.
[Quinlan1993]Quinlan, Ross. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
[Ramshaw and Marcus1995]Ramshaw, Lance A. and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the 3rd ACL Workshop on Very Large Corpora*, pages 82–94, Cambridge, MA, USA.
[Reed et al.2008]Reed, Chris, 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.
[Reed and Rowe2004]Reed, Chris and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 14(4):961–980.
[Reed and Walton2003]Reed, Chris and Douglas Walton. 2003. Argumentation schemes in argument-as-process and argument-as-product. In *Proceedings of the Conference Celebrating Informal Logic @25*, Windsor, ON, USA.
[Reed, Walton, and Macagno2007]Reed, Chris, Douglas Walton, and Fabrizio Macagno. 2007. Argument diagramming in logic, law and artificial intelligence. *Knowledge Engineering Review*, 22(1):87–109.
[Rooney, Wang, and Browne2012]Rooney, Niall, 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.
[Sampson and Clark2006]Sampson, Victor D. and Douglas B. Clark. 2006. Assessment of argument in science education: A critical review of the literature. In *Proceedings of the 7th International Conference on Learning Sciences*, ICLS '06, pages 655–661, Bloomington, IN, USA.
[Schapire and Singer2000]Schapire, Robert E. and Yoram Singer. 2000. Boostexter: A boosting-based systemfor text categorization. *Machine Learning*, 39(2-3):135–168.
[Scheuer et al.2010]Scheuer, Oliver, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5(1):43–102.
[Scott1955]Scott, William A. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3):321–325.

[Sergeant2013]Sergeant, Alan. 2013. Automatic argumentation extraction. In *Proceedings of the 10th European Semantic Web Conference*, ESWC '13, pages 656–660, Montpellier, France.
[Shermis and Burstein2013]Shermis, Mark D. and Jill Burstein. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge Chapman & Hall.
[Shiach2009]Shiach, Don. 2009. *How to write essays*. How To Books Ltd, 2nd edition.
[Shute2008]Shute, Valerie J. 2008. Focus on formative feedback. *Review of Education Research*, 78(1):153–189.
[Socher et al.2013]Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, WA, USA.
[Sokolova and Lapalme2009]Sokolova, Marina and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.
[Somasundaran and Wiebe2009]Somasundaran, Swapna and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, ACL '09, pages 226–234, Suntec, Singapore.
[Somasundaran and Wiebe2010]Somasundaran, Swapna and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 116–124, Los Angeles, CA, USA.
[Song et al.2014]Song, Yi, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, Baltimore, MA, USA.
[Soricut and Marcu2003]Soricut, Radu and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 149–156, Edmonton, Canada.
[Stab and Gurevych2014a]Stab, Christian and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510, Dublin, Ireland, August.
[Stab and Gurevych2014b]Stab, Christian 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, Doha, Qatar.
[Stab et al.2014]Stab, Christian, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 40–49, Bertinoro, Italy.
[Stenetorp et al.2012]Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Avignon, France.
[Teufel1999]Teufel, Simone. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.
[Thomas1973]Thomas, Stephen N. 1973. *Practical reasoning in natural language*. Prentice-Hall.
[Toulmin1958]Toulmin, Stephen E. 1958. *The uses of Argument*. Cambridge University Press.
[Toutanova et al.2003]Toutanova, Kristina, 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.
[van Eemeren and Grootendorst2004]van Eemeren, Frans H. and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The pragma-dialectical approach*. Cambridge University Press.
[van Eemeren, Grootendorst, and Snoeck Henkemans1996]van Eemeren, Frans H., Rob Grootendorst, and Francisca Snoeck Henkemans. 1996. *Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments*. Routledge, Taylor & Francis Group.
[Walker et al.2012]Walker, Marilyn, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eight International*

Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.
[Walton and Godden2007]Walton, Douglas and David M. Godden, 2007. *Reason Reclaimed*, chapter Informal Logic and the Dialectical Approach to Argument, pages 3–17. Newport News, Vale Press, Virginia, USA.
[Walton, Reed, and Macagno2008]Walton, Douglas, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.
[Whitaker2009]Whitaker, Anne. 2009. *Academic Writing Guide 2010: A Step-by-Step Guide to Writing Academic Papers*. City University of Seattle.
[Wilson, Wiebe, and Hoffmann2005]Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Vancouver, British Columbia, Canada.
[Wolfe and Britt2009]Wolfe, Christopher R. and M. Anne Britt. 2009. Argumentation schema and the myside bias in written argumentation. *Written Communication*, 26(2):183–209.
[Woods and Walton2007]Woods, John and Douglas Walton. 2007. *Fallacies: Selected Papers 1972-1982*. College Publications.
[Wyner et al.2010]Wyner, Adam, 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. Springer.
[Wyner et al.2012]Wyner, Adam, Jodi Schneider, Katie Atkinson, and Trevor J. M. Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. In *COMMA*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 43–50. IOS Press.
[Yanal1991]Yanal, Robert J. 1991. Dependent and independent reasons. *Informal Logic*, 13(3):137–144.