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 a novel approach for parsing argumentation structures. We identify argument components using sequence labeling at the token level and apply a new joint model for detecting argumentation structures. The proposed model globally optimizes argument component types and argumentative relations using Integer Linear Programming. We show that our model significantly outperforms challenging heuristic baselines on two different types of discourse. Moreover, we introduce a novel corpus of persuasive essays annotated with argumentation structures. We show that our annotation scheme and annotation guidelines successfully guide human annotators to substantial agreement.

1. Introduction

Argumentation aims at increasing or decreasing the acceptability of a controversial standpoint (van Eemeren, Grootendorst, and Snoeck Henkemans 1996, page 5). It is a routine that is omnipresent in our daily verbal communication and thinking. Well-reasoned arguments are not only important for decision making and learning but also play a crucial role in drawing widely accepted conclusions.

Computational argumentation is a recent research field in computational linguistics that focuses on the analysis of arguments in natural language texts. Novel methods have broad application potential in various areas such as legal decision support (Mochales-Palau and Moens 2009), information retrieval (Carstens and Toni 2015), policy making (Sardianos et al. 2015), and debating technologies (Levy et al. 2014; Rinott et al. 2015). Recently, computational argumentation has been receiving increased attention in computer-assisted writing (Song et al. 2014; Stab et al. 2014) because it allows the creation of writing support systems that provide feedback about written arguments.

* Technische Universität Darmstadt, Ubiquitous Knowledge Processing (UKP) Lab, Hochschulstrasse 10, D-64289 Darmstadt, Germany. E-mail: stab@ukp.informatik.tu-darmstadt.de.

** Technische Universität Darmstadt, Ubiquitous Knowledge Processing (UKP) Lab, Hochschulstrasse 10, D-64289 Darmstadt, Germany and Ubiquitous Knowledge Processing Lab (UKP-DIPF), German Institute for Educational Research, Schloßstraße 29, D-60486 Frankfurt am Main, Germany.

Submission received: 26 October 2015; revised version received: 27 February 2017; accepted for publication: 10 April 2017.

doi:10.1162/COLI_a_00295

Argumentation structures are closely related to discourse structures such as those defined by Rhetorical Structure Theory (RST) (Mann and Thompson 1987), the Penn Discourse Treebank (PDTB) (Prasad et al. 2008), or Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides 2003). The internal structure of an **argument** consists of several **argument components**. It includes a claim and one or more premises (Govier 2010). The **claim** is a controversial statement and the central component of an argument, and **premises** are reasons for justifying (or refuting) the claim. Moreover, arguments have directed **argumentative relations**, describing the relationships one component has with another. Each such relation indicates that the source component is either a justification for or a refutation of the target component.

The identification of argumentation structures involves several subtasks like separating argumentative from non-argumentative text units (Moens et al. 2007; Florou et al. 2013), classifying argument components into claims and premises (Mochales-Palau and Moens 2011; Rooney, Wang, and Browne 2012; Stab and Gurevych 2014b), and identifying argumentative relations (Mochales-Palau and Moens 2009; Peldszus 2014; Stab and Gurevych 2014b). However, an approach that covers all subtasks is still missing. Furthermore, most approaches operate locally and do not optimize the global argumentation structure. Recently, Peldszus and Stede (2015) proposed an approach based on Minimum Spanning Trees, which jointly models argumentation structures. However, it links all argument components in a single tree structure. Consequently, it is not capable of splitting a text containing more than one argument. In addition to the lack of end-to-end approaches for parsing argumentation structures, there are relatively few corpora annotated with argumentation structures at the discourse-level. Apart from our previous corpus (Stab and Gurevych 2014a), the few existing corpora lack non-argumentative text units (Peldszus 2014), are not annotated with claims and premises (Kirschner, Eckle-Kohler, and Gurevych 2015), or the reliability is unknown (Reed et al. 2008).

Our primary motivation for this work is to create argument analysis methods for argumentative writing support systems and to achieve a better understanding of argumentation structures. Therefore, our first research question is whether human annotators can reliably identify argumentation structures in persuasive essays and whether it is possible to create annotated data of high quality. The second research question addresses the automatic recognition of argumentation structure. We investigate if, and how accurately, argumentation structures can be identified by computational techniques. The contributions of this article are the following:

- An annotation scheme for modeling argumentation structures derived from argumentation theory. Our annotation scheme models the argumentation structure of a document as a connected tree.
- A novel corpus of 402 persuasive essays annotated with discourse-level argumentation structures. We show that human annotators can apply our annotation scheme to persuasive essays with substantial agreement. This corpus and the annotation guidelines are freely available.[1]
- An end-to-end argumentation structure parser that identifies argument components at the token level and globally optimizes component types and argumentative relations.

1 www.ukp.tu-darmstadt.de/data/argumentation-mining.

The remainder of this article is structured as follows: In Section 2, we review related work in computational argumentation and discuss the difference to traditional discourse analysis. In Section 3, we derive our annotation scheme from argumentation theory. Section 4 presents the results of an annotation study and the corpus creation. In Section 5, we introduce the argumentation structure parser. We show that our model significantly outperforms challenging heuristic baselines on two different types of discourse. We discuss our results in Section 6, and provide our conclusions in Section 7.

2. Related Work

Existing work in computational argumentation addresses a variety of different tasks. These include, for example, approaches for identifying reasoning type (Feng and Hirst 2011), argumentation style (Oraby et al. 2015), the stance of the author (Hasan and Ng 2014; Somasundaran and Wiebe 2009), the acceptability of arguments (Cabrio and Villata 2012), and appropriate support types (Park and Cardie 2014). Most relevant to our work, however, are approaches on argument mining that focus on the identification of argumentation structures in natural language texts. We categorize related approaches into the following three subtasks:

- **Component identification** focuses on the separation of argumentative from non-argumentative text units and the identification of argument component boundaries.
- **Component classification** addresses the function of argument components. It aims at classifying argument components into different types such as claims and premises.
- **Structure identification** focuses on linking arguments or argument components. Its objective is to recognize different types of argumentative relations such as support or attack relations.

2.1 Component Identification

Moens et al. (2007) identified argumentative sentences in various types of text such as newspapers, parliamentary records, and online discussions. They experimented with various different features and achieved an accuracy of 0.738 with word pairs, text statistics, verbs, and keyword features. Florou et al. (2013) classified text segments as argumentative or non-argumentative using discourse markers and several features extracted from the tense and mood of verbs. They report an F1 score of 0.764. Levy et al. (2014) proposed a pipeline including three consecutive steps for identifying context-dependent claims in Wikipedia articles. Their first component detects topic-relevant sentences including a claim. The second component detects the boundaries of each claim. The third component ranks the identified claims for identifying the most relevant claims for the given topic. They report a mean precision of 0.09 and a mean recall of 0.73 averaged over 32 topics for retrieving 200 claims. Goudas et al. (2014) presented a two-step approach for identifying argument components and their boundaries in social media texts. First, they classified each sentence as argumentative or non-argumentative and achieved 0.774 accuracy. Second, they segmented each argumentative sentence using a Conditional Random Field (CRF). Their best model achieved 0.424 accuracy.

2.2 Component Classification

The objective of the component classification task is to identify the type of argument components. Kwon et al. (2007) proposed two consecutive steps for identifying different types of claims in online comments. First, they classified sentences as claims and obtained an F1 score of 0.55 with a boosting algorithm. Second, they classified each claim as either support, oppose, or propose. Their best model achieved an F1 score of 0.67. Rooney, Wang, and Browne (2012) applied kernel methods for classifying text units as either claims, premises, or non-argumentative. They obtained an accuracy of 0.65. Mochales-Palau and Moens (2011) classified sentences in legal decisions as claim or premise. They achieved an F1 score of 0.741 for claims and 0.681 for premises using a Support Vector Machine (SVM) with domain-dependent key phrases, text statistics, verbs, and the tense of the sentence. In our previous work, we used a multiclass SVM for labeling text units of student essays as major claim, claim, premise, or non-argumentative (Stab and Gurevych 2014b). We obtained an F1 score of 0.726 using structural, lexical, syntactic, indicator, and contextual features. Recently, Nguyen and Litman (2015) found that argument and domain words from unlabeled data increase F1 score to 0.76 in the same experimental setup, and Lippi and Torroni (2015) achieved an F1 score of 0.714 for identifying sentences containing a claim in student essays using partial tree kernels.

2.3 Structure Identification

Approaches on structure identification can be divided into macro-level approaches and micro-level approaches. Macro-level approaches such as presented by Cabrio and Villata (2012), Ghosh et al. (2014), or Boltužić and Šnajder (2014) address relations between complete arguments and ignore the microstructure of arguments. More relevant to our work, however, are micro-level approaches, which focus on relations between argument components. Mochales-Palau and Moens (2009) introduced one of the first approaches for identifying the microstructure of arguments. Their approach is based on a manually created Context-Free Grammar and recognizes argument structures as trees. However, it is tailored to legal argumentation and does not recognize implicit argumentative relations (i.e., relations that are not indicated by discourse markers). In previous work, we considered the identification of argument structures as a binary classification task of ordered argument component pairs (Stab and Gurevych 2014b). We classified each pair as support or not-linked using an SVM with structural, lexical, syntactic, and indicator features. Our best model achieved an F1 score of 0.722. However, the approach recognizes argumentative relations locally and does not consider contextual information. Peldszus (2014) modeled the targets of argumentative relations along with additional information in a single tagset. His tagset includes, for instance, several labels denoting whether an argument component at position n is argumentatively related to preceding argument components $n-1, n-2$, and so forth, or following argument components $n+1$, $n+2$, and so on. Although his approach achieved a promising accuracy of 0.48, it is only applicable to short texts. Peldszus and Stede (2015) presented the first approach that globally optimizes argumentative relations. They jointly modeled several aspects of argumentation structures using a Minimum Spanning Tree model and achieved an F1 score of 0.720. They found that the function (support or attack) and the role (opponent and proponent) of argument components are the most useful dimensions for improving the identification of argumentative relations. However, the

texts in their corpus were created artificially using a guideline that promotes having one opposing argument component in each text (cf. Section 2.4). Therefore, it is unclear whether the results can be reproduced with real data, which may exhibit arguments with fewer opposing argument components (Wolfe and Britt 2009). Moreover, their approach links all argument components in a single tree structure. Thus, it is not capable of separating several arguments and recognizing unlinked components.

2.4 Existing Corpora Annotated with Argumentation Structures

Existing corpora in computational argumentation cover numerous aspects of argumentation analysis. There are, for instance, corpora that address argumentation strength (Persing and Ng 2015), factual knowledge (Beigman Klebanov and Higgins 2012), various properties of arguments (Walker et al. 2012), argumentative relations between complete arguments at the macro-level (Cabrio and Villata 2014; Boltužić and Šnajder 2014), different types of argument components (Mochales-Palau and Ieven 2009; Kwon et al. 2007; Habernal and Gurevych 2017), and argumentation structures over several documents (Aharoni et al. 2014). However, corpora annotated with argumentation structures at the level of discourse are still rare.

One prominent resource is AraucariaDB (Reed et al. 2008). It includes heterogenous text types such as newspaper editorials, parliamentary records, judicial summaries, and online discussions. It also includes annotations describing the type of reasoning according to Walton's argumentation schemes (Walton, Reed, and Macagno 2008) and implicit argument components that were added by the annotators during the analysis. However, the reliability of the annotations is unknown. Furthermore, recent releases of AraucariaDB are not appropriate for training end-to-end argumentation structure parsers because they do not include non-argumentative text units.

Kirschner, Eckle-Kohler, and Gurevych (2015) annotated argumentation structures in Introduction and Discussion sections of 24 German scientific articles. Their annotation scheme includes four argumentative relations (support, attack, detail, and sequence). However, the corpus does not contain annotations for argument component types.

Peldszus and Stede (2015) created a small corpus of 112 German microtexts with controlled linguistic and rhetoric complexity. Each document contains a single argument and does not include more than five argument components. Their annotation scheme models supporting and attacking relations as well as additional information like proponent and opponent. They obtained an inter-annotator agreement (IAA) of $\kappa = 0.83$[2] with three expert annotators. Recently, they translated the corpus to English, resulting in the first parallel corpus for computational argumentation. However, the corpus does not include non-argumentative text units. Therefore, the corpus is only of limited use for training end-to-end argumentation structure parsers. Because of the writing guidelines used (Peldszus and Stede 2013, page 197), it also exhibits an unusually high proportion of attack relations. In particular, 97 of the 112 arguments (86.6%) include at least one attack relation. This proportion is rather unnatural, since authors tend to support their standpoint instead of considering opposing views (Wolfe and Britt 2009).

2 The kappa coefficient is an IAA measure for categorical items that accounts for agreement by chance. The formal definition and a comprehensive overview of chance-corrected IAA measures can be found in the survey of Artstein and Poesio (2008).

Table 1
Existing corpora annotated with argumentation structures at the discourse-level (#Doc = number of documents; #Comp = number of argument components; NoArg = presence of non-argumentative text units).

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	$\alpha_U = 0.72$
(Peldszus and Stede 2015)	microtexts	112	576	no	clause	$\kappa = 0.83$
(Kirschner et al. 2015)	scientific articles	24	∼2,700	yes	sentence	$\kappa = 0.43$

In previous work, we created a corpus of 90 persuasive essays, which we selected randomly from essayforum.com (Stab and Gurevych 2014a). We annotated the corpus in two consecutive steps: First, we identified argument components at the clause level and obtained an agreement of $\alpha_U = 0.72$ between three annotators. Second, we annotated argumentative support and attack relations between argument components and achieved an agreement of $\kappa = 0.8$. Because the corpus also includes non-argumentative text units, it allows for training end-to-end argumentation structure parsers that separate argumentative from non-argumentative text units. Apart from this corpus, we are only aware of one additional study on argumentation structures in persuasive essays. Botley (2014) analyzed 10 essays using argument diagramming for studying differences in argumentation strategies. Unfortunately, the corpus is too small for computational purposes and the reliability of the annotations is unknown. Table 1 provides an overview of existing corpora annotated with argumentation structures at the discourse-level.

2.5 Discourse Analysis

The identification of argumentation structures is closely related to discourse analysis. Similar to the identification of argumentation structures, discourse analysis aims at identifying elementary discourse units and discourse relations between them. Existing approaches on discourse analysis mainly differ in the discourse theory utilized. RST (Mann and Thompson 1987), for instance, models discourse structures as trees by iteratively linking adjacent discourse units (Feng and Hirst 2014; Hernault et al. 2010) whereas approaches based on PDTB (Prasad et al. 2008) identify more shallow structures by linking two adjacent sentences or clauses (Lin, Ng, and Kan 2014). RST and PDTB are limited to discourse relations between adjacent discourse units, but SDRT (Asher and Lascarides 2003) also allows long distance relations (Afantenos and Asher 2014; Afantenos et al. 2015). However, similar to argumentation structure parsing, the main challenge of discourse analysis is to identify implicit discourse relations (Braud and Denis 2014, page 1694).

Marcu and Echihabi (2002) proposed one of the first approaches for identifying implicit discourse relations. In order to collect large amounts of training data, they exploited several discourse markers like "because" or "but". After removing the discourse markers, they found that word pair features are useful for identifying implicit discourse relations. Pitler, Louis, and Nenkova (2009) proposed an approach for identifying four implicit types of discourse relations in the PDTB and achieved F1 scores between 0.22 and 0.76. They found that using features tailored to each individual relation leads to the best results. Lin, Kan, and Ng (2009) showed that production rules collected from

parse trees yield good results and Louis et al. (2010) found that features based on named entities do not perform as well as lexical features.

Approaches to discourse analysis usually aim at identifying various different types of discourse relations. However, only a subset of these relations is relevant for argumentation structure parsing. For example, Peldszus and Stede (2013) proposed support, attack, and counter-attack relations for modeling argumentation structures, whereas our work focuses on support and attack relations. This difference is also illustrated by the work of Biran and Rambow (2011). They selected a subset of 12 relations from the RST Discourse Treebank (Carlson, Marcu, and Okurowski 2001) and argue that only a subset of RST relations is relevant for identifying justifications.

3. Argumentation: Theoretical Background

The study of argumentation is a comprehensive and interdisciplinary research field. It involves philosophy, communication science, logic, linguistics, psychology, and computer science. The first approaches to studying argumentation date back to the ancient Greek sophists and evolved in the 6th and 5th centuries BCE (van Eemeren, Grootendorst, and Snoeck Henkemans 1996). In particular, the influential works of Aristotle on traditional logic, rhetoric, and dialectics set an important milestone and are a cornerstone of modern argumentation theory. Because of the diversity of the field, there are numerous proposals for modeling argumentation. Bentahar, Moulin, and Bélanger (2010) categorize argumentation models into three types: (1) monological models, (2) dialogical models, and (3) rhetorical models. **Monological models** address the internal microstructure of arguments. They focus on the function of argument components, the links between them, and the reasoning type. Most monological models stem from the field of informal logic and focus on arguments as product (O'Keefe 1977; Johnson 2000). On the other hand, **dialogical models** focus on the process of argumentation and ignore the microstructure of arguments. They model the external macrostructure and address relations between arguments from several interlocutors. Finally, **rhetorical models** consider neither the micro- nor the macrostructure but rather the way arguments are used as a means of persuasion. They consider the audience's perception and aim at studying rhetorical schemes that are successful in practice. In this article, we focus on the monological perspective, which is well-suited for developing computational methods (Peldszus and Stede 2013).

3.1 Argument Diagramming

The laying out of argument structure is a widely used method in informal logic (Copi and Cohen 1990; Govier 2010). This technique, referred to as **argument diagramming**, aims at transferring natural language arguments into a structured representation for evaluating them in subsequent analysis steps (Henkemans 2000, page 447). Although argumentation theorists consider argument diagramming a manual activity, the diagramming conventions also serve as a good foundation for developing novel argument mining models (Peldszus and Stede 2013). An argument diagram is a node-link diagram whereby each node represents an argument component (i.e., a statement represented in natural language) and each link represents a directed argumentative relation indicating that the source component is a justification (or refutation) of the target component. Figure 1 shows some common argument structures. A **basic argument** includes a claim supported by a single premise. It can be considered the minimal form that an argument can take. A **convergent argument** comprises two premises that support the

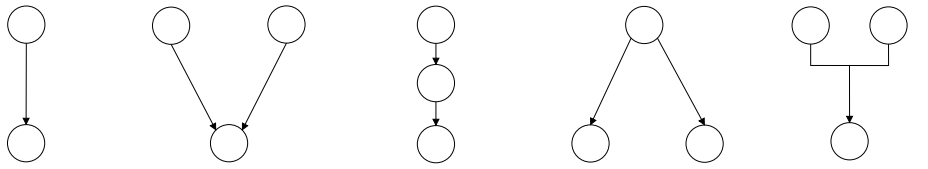

Figure 1
Microstructures of arguments: Nodes are argument components and links represent argumentative relations. Nodes at the bottom are the claims of the arguments.

claim individually; an argument is **serial** if it includes a reasoning chain and **divergent** if a premise supports several claims (Beardsley 1950). Complementarily, Thomas (1973) defined **linked arguments** (Figure 1e). Like convergent arguments, a linked argument includes two premises. However, neither of the two premises independently supports the claim. The premises are only relevant to the claim in conjunction. More complex arguments can combine any of the elementary structures illustrated in Figure 1.

On closer inspection, however, there are several ambiguities when applying argument diagramming to real texts: First, the distinction between convergent and linked structures is often ambiguous in real argumentation structures (Henkemans 2000; Freeman 2011). Second, it is unclear if the argumentation structure is a graph or a tree. Third, the argumentative type of argument components is ambiguous in serial structures. We discuss each of these questions in the following sections.

3.1.1 Distinguishing between Linked and Convergent Arguments. The question of whether an argumentation model needs to distinguish between linked and convergent arguments is still debated in argumentation theory (Conway 1991; Yanal 1991; van Eemeren, Grootendorst, and Snoeck Henkemans 1996; Freeman 2011). From a perspective based on traditional logic, linked arguments indicate deductive reasoning and convergent arguments represent inductive reasoning (Henkemans 2000, page 453). However, Freeman (2011, page 91ff.) showed that the traditional definition of linked arguments is frequently ambiguous in everyday discourse. Yanal (1991) argues that the distinction is equivalent to separating several arguments and Conway (1991) argues that linked structures can simply be omitted for modeling single arguments. From a computational perspective, the identification of linked arguments is equivalent to finding groups of premises or classifying the reasoning type of an argument as either deductive or inductive. Accordingly, it is not necessary to distinguish linked and convergent arguments during the identification of argumentation structures since this task can be solved in subsequent analysis steps.

3.1.2 Argumentation Structures as Trees. Defining argumentation structures as trees implies the exclusion of divergent arguments, to allow only one target for each premise and to neglect cycles. From a theoretical perspective, divergent structures are equivalent to several arguments (one for each claim) (Freeman 2011, page 16). As a result of this treatment, a great many of theoretical textbooks neglect divergent structures (Henkemans 2000; Reed and Rowe 2004) and also most computational approaches consider arguments as trees (Cohen 1987; Mochales-Palau and Moens 2009; Peldszus 2014). However, there is little empirical evidence regarding the structure of arguments. We are

only aware of one study, which showed that 5.26% of the arguments in political speeches (which can be assumed to exhibit complex argumentation structures) are divergent.

Essay writing usually follows a claim-oriented procedure (Kemper and Sebranek 2004; Shiach 2009; Whitaker 2009; Perutz 2010). Starting with the formulation of the standpoint on the topic, authors collect claims in support (or opposition) of their view. Subsequently, they collect premises that support or attack their claims. The following example illustrates this procedure. A major claim on abortion, for instance, is "abortion should be illegal"; a supporting claim could be "abortion is ethically wrong" and the associated premises "unborn babies are considered human beings" and "killing human beings is wrong". Because of this common writing procedure, divergent and circular structures are rather unlikely in persuasive essays. Therefore, we assume that modeling the argumentation structure of essays as a tree is a reasonable decision.

3.1.3 Argumentation Structures and Argument Component Types. Assigning argumentative types to the components of an argument is unambiguous if the argumentation structure is shallow. It is, for instance, obvious that an argument component c_1 is a premise and argument component c_2 is a claim, if c_1 supports c_2 in a basic argument (cf. Figure 1). However, if the tree structure is deeper (i.e., exhibits serial structures), assigning argumentative types becomes ambiguous. Essentially, there are three different approaches for assigning argumentative types to argument components. First, according to Beardsley (1950) a serial argument includes one argument component which is both a claim and a premise. Therefore, the inner argument component bears two different argumentative types (**multi-label approach**). Second, Govier (2010, page 24) distinguishes between "main claim" and "subclaim". Similarly, Damer (2009, page 17) distinguishes between "premise" and "subpremise" for labeling argument components in serial structures. Both approaches define specific 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 premises (**one-claim approach**). In order to define an argumentation model for persuasive essays, we propose a hybrid approach that combines the level approach and the one-claim approach.

3.2 Argumentation Structures in Persuasive Essays

We model the argumentation structure of persuasive essays as a connected tree structure. We use a level approach for modeling the first level of the tree and a one-claim approach for representing the structure of each individual argument. Accordingly, we model the first level of the tree with two different argument component types and the structure of individual arguments with argumentative relations.

The **major claim** is the root node of the argumentation structure and represents the author's standpoint on the topic. It is an opinionated statement that is usually stated in the introduction and restated in the conclusion of the essay. The individual body paragraphs of an essay include the actual arguments. They either support or attack the author's standpoint expressed in the major claim. Each argument consists of a claim and at least one premise. In order to differentiate between supporting and attacking arguments, each claim has a **stance attribute** that can take the values "for" or "against".

We model the structure of each argument with a one-claim approach. The claim constitutes the central component of each argument. The premises are the reasons of the argument. The actual structure of an argument comprises directed argumentative support and attack relations, which link a premise either to a claim or to another premise

(serial arguments). Each premise p has one **outgoing relation** (i.e., there is a relation that has p as source component) and none or several **incoming relations** (i.e., there can be a relation with p as target component). A claim can exhibit several incoming relations but no outgoing relation. The ambiguous function of inner premises in serial arguments is implicitly modeled by the structure of the argument. The inner premise exhibits one outgoing relation and at least one incoming relation. Finally, the stance of each premise is indicated by the type of its outgoing relation (support or attack).

The following example illustrates the argumentation structure of a persuasive essay.[3] The introduction of an essay describes the controversial topic and usually includes the major claim:

> *Ever since researchers at the Roslin Institute in Edinburgh cloned an adult sheep, there has been an ongoing debate about whether 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 argumentative content. The third sentence contains the major claim (boldfaced) and a claim that supports the major claim (underlined). The following body paragraphs of the essay include arguments that either support or attack the major claim. For example, the following body paragraph includes one argument that supports the positive standpoint 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 following three sentences (wavy underlined). The second sentence includes two premises, of which $Premise_1$ supports $Claim_2$ and $Premise_2$ supports $Premise_1$. $Premise_3$ in the third sentence supports $Claim_2$. The fourth sentence includes $Premise_4$ and $Premise_5$. Both support $Premise_3$. The next paragraph illustrates a body paragraph with 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, [infertile couples can bear children that are genetically related]$_{Premise7}$. [Even same sex couples can have children]$_{Premise8}$. Consequently, [cloning can help couples have children]$_{Claim4}$.*

The initial sentence includes the first argument, which consists of $Premise_6$ and $Claim_3$. The following three sentences include the second argument. $Premise_7$ and $Premise_8$ both support $Claim_4$ in the last sentence. Both arguments cover different aspects (development in science and cloning humans), which both support the author's standpoint on

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

cloning. This example illustrates that knowing argumentative relations is important for separating several arguments in a paragraph. The example also shows that argument components frequently exhibit preceding text units that are not relevant to the argument but helpful for recognizing the argument component type. For example, preceding discourse connectors like "therefore", "consequently", or "thus" can signal a subsequent claim. Discourse markers like "because", "since", or "furthermore" could indicate a premise. Formally, these **preceding tokens** of an argument component starting at token t_i are defined as the tokens $t_{i-m}, ..., t_{i-1}$ that are not covered by another argument component in the sentence $s = t_1, t_2, ..., t_n$ where $1 \leq i \leq n$ and $i - m \geq 1$. 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 $Claim_5$, which attacks the stance of the author. It is supported by $Premise_9$ in the second sentence. The third sentence includes two premises, both of which defend the stance of the author. $Premise_{11}$ is an attack of $Claim_5$, and $Premise_{10}$ supports $Premise_{11}$. The last paragraph (conclusion) 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 to humanity**]$_{MajorClaim2}$. It is likely that [this technology bears some important cures which will significantly improve life conditions]$_{Claim7}$.*

The conclusion of the essay starts with an attacking claim followed by the restatement of the major claim. The last sentence includes another claim that summarizes the most important points of the author's argumentation. Figure 2 shows the entire argumentation structure of the example essay.

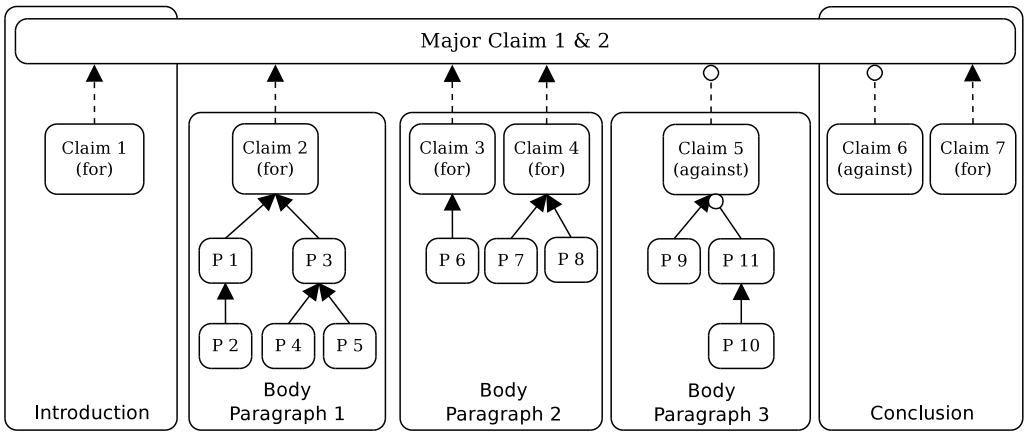

Figure 2
Argumentation structure of the example essay. Arrows indicate argumentative relations. Arrowheads denote argumentative support relations and circleheads attack relations. Dashed lines indicate relations that are encoded in the stance attributes of claims. "P" denotes premises.

4. Corpus Creation

The motivation for creating a new corpus is threefold: First, our previous corpus is relatively small. We believe that more data will improve the accuracy of our computational models. Second, we wanted to ensure the reproducibility of the annotation study and validate our previous results. Third, we improved our annotation guidelines. We added more precise rules for segmenting argument components and a detailed description of common essay structures. We expect that our novel annotation guidelines will guide annotators towards adequate agreement without collaborative training sessions. Our annotation guidelines comprise 31 pages and include the following three steps:

1. Topic and stance identification: We found in our previous annotation study that knowing the topic and stance of an essay improves inter-annotator agreement (Stab and Gurevych 2014a). For this reason, we ask the annotators to read the entire essay before starting with the annotation task.
2. Annotation of argument components: Annotators mark major claims, claims, and premises. They annotate the boundaries of argument components and determine the stance attribute of claims.
3. Linking premises with argumentative relations: The annotators identify the structure of arguments by linking each premise to a claim or another premise with argumentative support or attack relations.

Three non-native speakers participated in our annotation study. One of the three annotators had participated in our previous study (expert annotator).[4] The two other annotators learned the task by independently reading the annotation guidelines. We used the brat rapid annotation tool (Stenetorp et al. 2012). It provides a graphical web interface for marking text units and linking them.

4.1 Data

We randomly selected 402 English essays with a description of the writing prompt from essayforum.com. This online forum is an active community that provides correction and feedback about different texts such as research papers, essays, or poetry. For example, students post their essays in order to receive feedback about their writing skills while preparing for standardized language tests. The corpus includes 7,116 sentences with 147,271 tokens.

4.2 Inter-Annotator Agreement

All three annotators independently annotated a random subset of 80 essays. The remaining 322 essays were annotated by the expert annotator. We evaluate the inter-annotator agreement of the argument component annotations using two different strategies: First, we evaluate if the annotators agree on the presence of argument components in sentences using observed agreement and Fleiss' κ (Fleiss 1971). We consider each sentence as a markable and evaluate the presence of each argument component type

4 Although it would be preferable to have a group of annotators with similar annotation experience (e.g. all non-experts), because of lack of resources it is a common practice to have mixed annotator groups.

Table 2
Inter-annotator agreement of argument components.

Component type	Observed agreement	Fleiss' κ	α_U
MajorClaim	97.9%	0.877	0.810
Claim	88.9%	0.635	0.524
Premise	91.6%	0.833	0.824

$t \in \{MajorClaim, Claim, Premise\}$ in a sentence individually. Accordingly, the number of markables for each argument component type t corresponds to the number of sentences $N = 1{,}441$, the number of annotations per markable equals the number of annotators ($n = 3$), and the number of categories is $k = 2$ (t or *not t*). Evaluating the agreement at the sentence level is an approximation of the actual agreement since the boundaries of argument components can differ from sentence boundaries and a sentence can include several argument components.[5] Therefore, for the second evaluation strategy, we use Krippendorff's α_U (Krippendorff 2004). In contrast to common alpha coefficients, this coefficient allows us to evaluate the agreement of unitizing tasks by comparing the boundaries of the annotation units. We use the squared difference δ^2 between any two annotators' sections as proposed by Krippendorff (2004, page 9) and consider each essay as a single continuum at the token level. Accordingly, the length L of each continuum is the number of tokens in an essay. The number of annotators m that unitize the continuum is 3. We report the average α_U scores over 80 essays. For determining the inter-annotator agreement, we use DKPro Agreement, whose implementations of inter-annotator agreement measures are well-tested with various examples from the literature (Meyer et al. 2014).

Table 2 shows the inter-annotator agreement of each argument component type. The agreement is best for major claims. The IAA score of 97.9% and $\kappa = 0.877$ indicate that annotators are able to reliably identify major claims in persuasive essays. In addition, the unitized alpha measure of $\alpha_U = 0.810$ shows that there are only few disagreements about the boundaries of major claims. The results also indicate good agreement for premises ($\kappa = 0.833$ and $\alpha_U = 0.824$). We obtain the lowest agreement of $\kappa = 0.635$ for claims, which shows that the identification of claims is more complex than identifying major claims and premises. The joint unitized measure for all argument components is $\alpha_U = 0.767$, and thus the agreement improved by 0.043 compared with our previous study (Stab and Gurevych 2014b). Therefore, we tentatively conclude that overall, human annotators agree on the argument components in persuasive essays.

For determining the agreement of the stance attribute, we follow the same methodology as for the sentence-level agreement described above, but we consider each sentence containing a claim as "for" or "against" according to its stance attribute, and all sentences without a claim as "none" ($N = 1{,}441$; $n = 3$; $k = 3$). Consequently, the agreement of claims constitutes the upper bound for the stance attribute. We obtain an agreement of 88.5% and $\kappa = 0.623$, which is slightly below the agreement scores

5 In our evaluation set of 80 essays the annotators identified in 4.3% of the sentences several argument components of different types. Thus, evaluating the reliability of argument components at the sentence level is a good approximation of the inter-annotator agreement.

Table 3
Inter-annotator agreement of argumentative relations.

Relation type	Observed agreement	Fleiss' κ
Support	92.3%	0.708
Attack	99.6%	0.737

of claims (cf. Table 2). Therefore, human annotators can reliably differentiate between supporting and attacking claims.

We determined the markables for evaluating the agreement of argumentative relations by pairing all argument components in the same paragraph. For each paragraph with argument components $c_1, ..., c_n$, we consider each pair $p = (c_i, c_j)$ with $1 \leq i, j \leq n$ and $i \neq j$ as markable. Thus, the set of all markables corresponds to all argument component pairs that can be annotated according to our guidelines. The number of argument component pairs is $N = 4{,}922$, the number of ratings per markable is $n = 3$, and the number of categories $k = 2$.

Table 3 shows the inter-annotator agreement of argumentative relations. We obtain kappa scores above 0.7 for both argumentative support and attack relations, which allows tentative conclusions (Krippendorff 2004). On average, the annotators marked only 0.9% of the 4,922 pairs as argumentative attack relations and 18.4% as argumentative support relations. Although the agreement is usually much lower if a category is rare (Artstein and Poesio 2008, page 573), the annotators agree more on argumentative attack relations. This indicates that the identification of argumentative attack relations is a simpler task than the identification of argumentative support relations. The agreement scores for argumentative relations are approximately 0.10 lower compared with our previous study. This difference can be attributed to the fact that we did not explicitly annotate relations between claims and major claims, which are easy to annotate because claims are always linked to major claims (cf. Section 3.2).

4.3 Analysis of Human Disagreement

For analyzing the disagreements between the annotators, we determined Confusion Probability Matrices (CPMs) (Cinková, Holub, and Kríž 2012). Compared with traditional confusion matrices, a CPM also allows us to analyze confusion if more than two annotators are involved in an annotation study. A CPM includes conditional probabilities that an annotator assigns a category in the column given that another annotator selected the category in the row. Table 4 shows the CPM of argument component

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

	MajorClaim	Claim	Premise	NoArg
MajorClaim	0.771	0.077	0.010	0.142
Claim	0.036	0.517	0.307	0.141
Premise	0.002	0.131	0.841	0.026
NoArg	0.059	0.126	0.054	0.761

annotations. It shows that the highest confusion is between claims and premises. We observed that one annotator frequently did not split sentences including a claim. For instance, the annotator labeled the entire sentence as a claim although it includes an additional premise. This type of error also explains the lower unitized alpha score compared with the sentence-level agreements in Table 2. Furthermore, we found that concessions before claims were frequently not annotated as an attacking premise. For example, annotators often did not split sentences similarly to the following example:

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

The distinction between major claims and claims exhibits less confusion. This may be because major claims are relatively easy to locate in essays since they occur usually in introductions or conclusions, whereas claims can occur anywhere in the essay.

Table 5 shows the CPM of argumentative relations. There is little confusion between argumentative support and attack relations. The CPM also shows that the highest confusion is between argumentative relations (support and attack) and unlinked pairs. This can be attributed to the identification of the correct targets of premises. In particular, we observed that agreement on the targets decreases if a paragraph includes several claims or serial argument structures.

4.4 Creation of the Final Corpus

We created a partial gold standard of the essays annotated by all annotators. We use this partial gold standard of 80 essays as our test data (20%) and the remaining 322 essays annotated by the expert annotator as our training data (80%). The creation of our gold standard test data consists of the following two steps: First, we merge the annotation of all argument components. Thus, each annotator annotates argumentative relations based on the same argument components. Second, we merge the argumentative relations to compile our final gold standard test data. Because the argument component types are strongly related—the selection of the premises, for instance, depends on the selected claim(s) in a paragraph—we did not merge the annotations using majority voting as in our previous study. Instead, we discussed the disagreements in several meetings with all annotators for resolving the disagreements.

4.5 Corpus Statistics

Table 6 gives an overview of the size of the corpus. It contains 6,089 argument components, 751 major claims, 1,506 claims, and 3,832 premises. Such a large proportion of

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

	Support	Attack	Not-Linked
Support	0.605	0.006	0.389
Attack	0.107	0.587	0.307
Not-Linked	0.086	0.004	0.910

Table 6
Statistics of the final corpus.

		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
	MajorClaims	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

claims compared with premises is common in argumentative texts because writers tend to provide several reasons for ensuring a robust standpoint (Mochales-Palau and Moens 2011).

The proportion of non-argumentative text amounts to 47,474 tokens (32.2%) and 1,631 sentences (22.9%). The number of sentences with several argument components is 583, of which 302 include several components with different types (e.g., a claim followed by premise). Therefore, the identification of argument components requires the separation of argumentative from non-argumentative text units and the recognition of component boundaries at the token level. The proportion of paragraphs with unlinked argument components (e.g., unsupported claims without incoming relations) is 421 (23%). Thus, methods that link all argument components in a paragraph are only of limited use for identifying the argumentation structures in our corpus.

In total, the corpus includes 1,130 arguments (i.e., claims supported by at least one premise). Only 140 of them have an attack relation. Thus, the proportion of arguments with attack relations is considerably lower than in the microtext corpus from Peldszus and Stede (2015). Most of the arguments are convergent—that is, the depth of the argument is 1. The number of arguments with serial structure is 236 (20.9%).

5. Parsing Argumentation Structure

Our approach for parsing argumentation structures consists of five consecutive subtasks, depicted in Figure 3. The **identification model** separates argumentative from non-argumentative text units and recognizes the boundaries of argument components. The next three models constitute a joint model for recognizing the argumentation structure. We train two base classifiers. The **argument component classification model** labels each argument component as major claim, claim, or premise, and the **argumentative relation identification model** recognizes if two argument components are argumentatively linked or not. The **tree generation model** globally optimizes the results of the two base classifiers for finding a tree (or several ones) in each paragraph. Finally, the **stance recognition model** differentiates between support and attack relations.

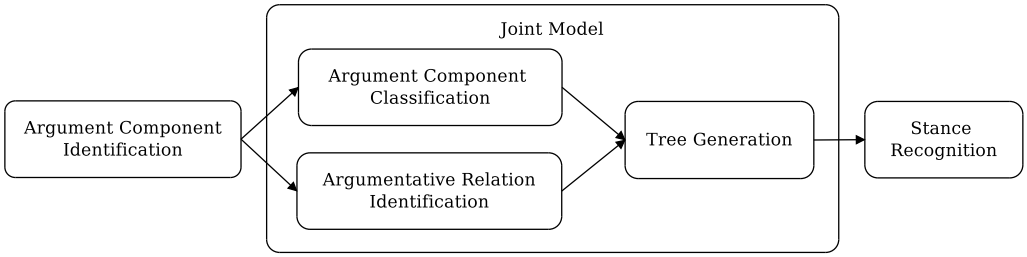

Figure 3
Architecture of the argumentation structure parser.

For preprocessing, we use several models from the DKPro Framework (Eckart de Castilho and Gurevych 2014). We identify tokens and sentence boundaries using the LanguageTool segmenter[6] and identify paragraphs 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 (POS) tagger (Toutanova et al. 2003), constituent and dependency parsers (Klein and Manning 2003), and sentiment analyzer (Socher et al. 2013). We use a discourse parser from Lin, Ng, and Kan (2014) for recognizing PDTB-style discourse relations. We use the DKPro TC text classification framework (Daxenberger et al. 2014) for feature extraction and experimentation.

In the following sections, we describe each model in detail. For finding the best-performing models, we conduct model selection on our training data using 5-fold cross-validation. Then, we conduct model assessment on our test data. We determine the evaluation scores of each cross-validation experiment by accumulating the confusion matrices of each fold into one confusion matrix, which has been shown to be the least biased method for evaluating cross-validation experiments (Forman and Scholz 2010). We use macro-averaging as described by Sokolova and Lapalme (2009) and report macro precision (P), macro recall (R), and macro F1 scores (F1). We use a two-sided Wilcoxon signed-rank test with $p = 0.01$ for significance testing. Because most evaluation measures for comparing system outputs are not normally distributed (Søgaard 2013), this non-parametric test is preferable to parametric tests, which make stronger assumptions about the underlying distribution of the random variables. We apply this test to all reported evaluation scores obtained for each of the 80 essays in our test set.

The remainder of this section is structured as follows: In the following section, we introduce the baselines and the upper bound for each task. In Section 5.2, we present the identification model that detects argument components and their boundaries. In Section 5.3, we propose a new joint model for identifying argumentation structures. In Section 5.4, we introduce our stance recognition model. In Section 5.5, we report the results of the model assessment on our test data and on the microtext corpus from Peldszus and Stede (2015). We present the results of the error analysis in Section 5.6. We evaluate the identification model independently and use the gold standard argument components for evaluating the remaining models.

6 www.languagetool.org.

5.1 Baselines and Upper Bound

For evaluating our models, we use two different types of baselines: First, we use **majority baselines** that label each instance with the majority class. Table A.1 in Appendix A shows the class distribution in our training data and test data for each task.

Second, we use **heuristic baselines**, which are motivated by the common structure of persuasive essays (Whitaker 2009; Perutz 2010). The heuristic baseline of the identification task exploits sentence boundaries. It selects all sentences as argument components except the first two and the last sentence of an essay.[7] The heuristic baseline of the classification task labels the first argument component in each body paragraph as claim, and all remaining components in body paragraphs as premise. The last argument component in the introduction and the first argument component in the conclusion are classified as major claim and all remaining argument components in the introduction and conclusion are labeled as claim. The heuristic baseline for the relation identification classifies an argument component pair as linked if the target is the first component of a body paragraph. We expect that this baseline will yield good results, because 62% of all body paragraphs in our corpus start with a claim. The heuristic baseline of the stance recognition classifies each argument component in the second to last paragraph as attack. The motivation for this baseline stems from essay writing guidelines, which recommend including opposing arguments in the second to last paragraph.

We determine the **human upper bound** for each task by averaging the evaluation scores of all three annotator pairs on our test data.

5.2 Identifying Argument Components

We consider the identification of argument components as a sequence labeling task at the token level. We encode the argument components using an IOB-tagset (Ramshaw and Marcus 1995) and consider an entire essay as a single sequence. Accordingly, we label the first token of each argument component as "Arg-B", the tokens covered by an argument component as "Arg-I", and non-argumentative tokens as "O". As a learner, we use a CRF (Lafferty, McCallum, and Pereira 2001) with the averaged perceptron training method (Collins 2002). Because a CRF considers contextual information, the model is particularly suited for sequence labeling tasks (Goudas et al. 2014, page 292). For each token, we extract the following features (Table 7):

Structural features capture the position of the token. We expect these features to be effective for filtering non-argumentative text units, since the introductions and conclusions of essays include few argumentatively relevant content. The punctuation features indicate if the token is a punctuation and if the token is adjacent to a punctuation.

Syntactic features consist of the token's POS as well as features extracted from the Lowest Common Ancestor (LCA) of the current token t_i and its adjacent tokens in the constituent parse tree. First, we define $LCA_{preceding}(t_i) = \frac{|lcaPath(t_i,t_{i-1})|}{depth}$, where $|lcaPath(u,v)|$ is the length of the path from u to the LCA of u and v, and $depth$ the depth of the constituent parse tree. Second, we define $LCA_{following}(t_i) = \frac{|lcaPath(t_i,t_{i+1})|}{depth}$, which considers the current token t_i and its following token t_{i+1}.[8] Additionally, we add the constituent types of both lowest common ancestors to our feature set.

7 Full stops at the end of a sentence are all classified as non-argumentative.

8 We set $LCA_{preceding} = -1$ if t_i is the first token in its covering sentence and $LCA_{following} = -1$ if t_i is the last token in its covering sentence.

Table 7
Features used for argument component identification (*indicates genre-dependent features).

Group	*Feature*	*Description*
Structural	Token position	Token present in introduction or conclusion*; token is first or last token in sentence; relative and absolute token position in document, paragraph and sentence
	Punctuation	Token precedes or follows any punctuation, full stop, comma and semicolon; token is any punctuation or full stop
	Position of covering sentence	Absolute and relative position of the token's covering sentence in the document and paragraph
Syntactic	Part-of-speech	The token's part-of-speech
	Lowest common ancestor (LCA)	Normalized length of the path to the LCA with the following and preceding token in the parse tree
	LCA types	The two constituent types of the LCA of the current token and its preceding and following token
LexSyn	Lexico-syntactic	Combination of lexical and syntactic features as described by Soricut and Marcu (2003)
Prob	Probability	Conditional probability of the current token being the beginning of a component given its preceding tokens

Lexico-syntactic features have been shown to be effective for segmenting elementary discourse units (Hernault et al. 2010). We adopt the features introduced by Soricut and Marcu (2003). We use lexical head projection rules (Collins 2003) implemented in the Stanford tool suite to lexicalize the constituent parse tree. For each token t, we extract its uppermost node n in the parse tree with the lexical head t and define a lexico-syntactic feature as the combination of t and the constituent type of n. We also consider the child node of n in the path to t and its right sibling, and combine their lexical heads and constituent types as described by Soricut and Marcu (2003).

The **probability feature** is the conditional probability of the current token t_i being the beginning of an argument component ("Arg-B") given its preceding tokens. We maximize the probability for preceding tokens of a length up to $n = 3$:

$$\underset{n \in \{1,2,3\}}{\operatorname{argmax}} P(t_i = \textit{Arg-B} | t_{i-n}, ..., t_{i-1})$$

To estimate these probabilities, we use maximum likelihood estimation on our training data.

5.2.1 Results of Argument Component Identification. The results of model selection show that using all features performs best. Table C.1 in Appendix C provides the detailed results of the feature analysis. Table 8 shows the results of the model assessment on the test data. The heuristic baseline achieves a macro F1 score of 0.642. It achieves an F1 score of 0.677 for non-argumentative tokens ("O") and 0.867 for argumentative tokens ("Arg-I"). Thus, the heuristic baseline effectively separates argumentative from non-argumentative text units. However, it achieves a low F1 score of 0.364 for

Table 8
Model assessment of argument component identification († = significant improvement over baseline heuristic).

	F1	P	R	F1 Arg-B	F1 Arg-I	F1 O
Human upper bound	0.886	0.887	0.885	0.821	0.941	0.892
Baseline majority	0.259	0.212	0.333	0	0.778	0
Baseline heuristic	0.642	0.664	0.621	0.364	0.867	0.677
CRF all features	†0.867	†0.873	†0.861	†0.809	†0.934	†0.857

identifying the beginning of argument components ("Arg-B"). Because it does not split sentences, it recognizes 145 fewer argument components than the number of gold standard components in the test data.

The CRF model with all features significantly outperforms the macro F1 score of the heuristic baseline ($p = 7.85 \times 10^{-15}$). Compared with the heuristic baseline, it performs significantly better in identifying the beginning of argument components ($p = 1.65 \times 10^{-14}$). It also performs better for separating argumentative from non-argumentative tokens ($p = 4.06 \times 10^{-14}$). In addition, the number of identified argument components differs only slightly from the number of gold standard components in our test data. It identifies 1,272 argument components, whereas the number of gold standard components in our test data amounts to 1,266. The human upper bound yields a macro F1 score of 0.886 for identifying argument components. The macro F1 score of our model is only 0.019 less. Therefore, our model achieves 97.9% of human performance.

5.2.2 Error Analysis. For identifying the most frequent errors of our model, we manually investigated the predicted argument components. The most frequent errors are false positives of "Arg-I". The model classifies 1,548 out of 9,403 non-argumentative tokens ("O") as argumentative ("Arg-I"). The reason for these errors is threefold: First, the model frequently labels non-argumentative sentences in the conclusion of an essay as argumentative. These sentences are, for instance, non-argumentative recommendations for future actions or summaries of the essay topic. Second, the model does not correctly recognize non-argumentative sentences in body paragraphs. It wrongly identifies argument components in 13 out of the 15 non-argumentative body paragraph sentences in our test data. The reason for these errors may be attributed to the high class imbalance in our training data. Third, the model tends to annotate lengthy non-argumentative preceding tokens as argumentative. For instance, it labels subordinate clauses preceding the actual argument component as argumentative in sentences similar to "In addition to the reasons mentioned above, [actual "Arg-B"] ..." (underlined text units represent the annotations of our model).

The second most frequent cause of errors are misclassified beginnings of argument components. The model classifies 137 of the 1,266 beginning tokens as "Arg-I". The model, for instance, 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"] ...". These examples also explain the false negatives of non-argumentative tokens which are wrongly classified as "Arg-B".

5.3 Recognizing Argumentation Structures

The identification of argumentation structures involves the classification of argument component types and the identification of argumentative relations. Both argumentative types and argumentative relations share information (Stab and Gurevych 2014b, p. 54). For instance, if an argument component is classified as claim, it is less likely to exhibit outgoing relations and more likely to have incoming relations. On the other hand, an argument component with an outgoing relation and few incoming relations is more likely to be a premise. Therefore, we propose a joint model that combines both types of information for finding the optimal structure. We train two local base classifiers. One classifier recognizes the type of argument components (Section 5.3.1), and another identifies argumentative relations between argument components (Section 5.3.2). For both models, we use an SVM (Cortes and Vapnik 1995) with a polynomial kernel implemented in the Weka machine learning framework (Hall et al. 2009). The motivation for selecting this learner stems from the results of our previous work, in which we found that SVMs outperform several other learners in both tasks (Stab and Gurevych 2014b, page 51). We globally optimize the outcomes of both classifiers in order to find the optimal argumentation structure using Integer Linear Programming (Section 5.3.3). In the following three sections, we first introduce the features of the two base classifiers before describing the Integer Linear Programming model.

5.3.1 Classifying Argument Components. We consider the classification of argument component types as multiclass classification and label each argument component as "major claim," "claim," or "premise." We experiment with the following feature groups:

Lexical features consist of binary lemmatized unigrams and the 2k most frequent dependency word pairs. We extract the unigrams from the component and its preceding tokens to ensure that discourse markers are included in the features.

Structural features capture the position of the component in the document and token statistics (Table 9). Because major claims occur frequently in introductions or conclusions, we expect that these features are valuable for differentiating component types.

Indicator features are based on four categories of lexical indicators that we manually extracted from 30 additional essays. **Forward indicators** such as "therefore", "thus", or "consequently" signal that the component following the indicator is a result of preceding argument components. **Backward indicators** indicate that the component following the indicator supports a preceding component. Examples of this category are "in addition", "because", or "additionally". **Thesis indicators** such as "in my opinion" or "I believe that" indicate major claims. **Rebuttal indicators** signal attacking premises or contra arguments. Examples are "although", "admittedly", or "but". The complete lists of all four categories are provided in Table B.1 in Appendix B. We define for each category a binary feature that indicates if an indicator of a category is present in the component or its preceding tokens. An additional binary feature indicates if first-person indicators are present in the argument component or its preceding tokens (Table 9). We assume that first-person indicators are informative for identifying major claims.

Contextual features capture the context of an argument component. We define eight binary features set to true if a forward, backward, rebuttal, or thesis indicator precedes or follows the current component in its covering paragraph. Additionally, we count the number of noun and verb phrases of the argument component that are also present in the introduction or conclusion of the essay. These features are motivated by the observation that claims frequently restate entities or phrases of the essay topic.

Table 9
Features of the argument component classification model (*indicates genre-dependent features).

Group	*Feature*	*Description*
Lexical	Unigrams	Binary and lemmatized unigrams of the component and its preceding tokens
	Dependency tuples	Lemmatized dependency tuples (2k most frequent)
Structural	Token statistics	Number of tokens of component, covering paragraph and covering sentence; number of tokens preceding and following the component in its sentence; ratio of component and sentence tokens
	Component position	Component is first or last in paragraph; component present in introduction or conclusion*; Relative position in paragraph; number of preceding and following components in paragraph
Indicators	Type indicators	Forward, backward, thesis or rebuttal indicators present in the component or its preceding tokens
	First-person indicators	"I", "me", "my", "mine", or "myself" present in component or its preceding tokens
Contextual	Type indicators in context	Forward, backward, thesis, or rebuttal indicators preceding or following the component in its paragraph
	Shared phrases*	Shared noun phrases or verb phrases with the introduction or conclusion (number and binary)
Syntactic	Subclauses	Number of subclauses in the covering sentence
	Depth of parse tree	Depth of the parse tree of the covering sentence
	Tense of main verb	Tense of the main verb of the component
	Modal verbs	Modal verbs present in the component
	POS distribution	POS distribution of the component
Probability	Type probability	Conditional probability of the component being a major claim, claim or premise, given its preceding tokens
Discourse	Discourse Triples	PDTB-discourse relations overlapping with the current component
Embedding	Combined word embeddings	Sum of the word vectors of each word of the component and its preceding tokens

Furthermore, we add four binary features indicating if the current component shares a noun or verb phrase with the introduction or conclusion.

Syntactic features consist of the POS distribution of the argument component, the number of subclauses in the covering sentence, the depth of the constituent parse tree of the covering sentence, the tense of the main verb of the component, and a binary feature that indicates whether a modal verb is present in the component.

The **probability features** are the conditional probabilities of the current component being assigned the type $t \in \{MajorClaim, Claim, Premise\}$ given the sequence of tokens p directly preceding the component. To estimate $P(t|p)$, we use maximum likelihood estimation on our training data.

Discourse features are based on the output of the PDTB-style discourse parser from Lin, Ng, and Kan (2014). Each binary feature is a triple combining the following information: (1) the type of the relation that overlaps with the current argument component, (2) whether the current argument component overlaps with the first or second elementary discourse unit of a relation, and (3) if the discourse relation is implicit or explicit. For instance, the feature `Contrast_imp_Arg1` indicates that the current component overlaps with the first discourse unit of an implicit contrast relation. The use of these features is motivated by the findings of Cabrio, Tonelli, and Villata (2013). By analyzing several example arguments, they hypothesized that general discourse relations could be informative for identifying argument components.

Embedding features are based on word embeddings trained on a part of the Google news data set (Mikolov et al. 2013). We sum the vectors of each word of an argument component and its preceding tokens and add it to our feature set. In contrast to common bag-of-words representations, embedding features have a continuous feature space that helped to achieve better results in several NLP tasks (Socher et al. 2013).

By experimenting with individual features and several feature combinations, we found that a combination of all features yields the best results. The results of the model selection can be found in Table C.2 in Appendix C.

5.3.2 Identifying Argumentative Relations. The relation identification model classifies ordered pairs of argument components as "linked" or "not-linked." In this analysis step, we consider both argumentative support and attack relations as "linked." For each paragraph with argument components $c_1, ..., c_n$, we consider $p = (c_i, c_j)$ with $i \neq j$ and $1 \leq i, j \leq n$ as an argument component pair. An argument component pair is "linked" if our corpus contains an argumentative relation with c_i as source component and c_j as target component. The class distribution is skewed towards "not-linked" pairs (Table A.1). We experiment with the following features:

Lexical features are binary lemmatized unigrams of the source and target component and their preceding tokens. We limit the number of unigrams for both source and target component to the 500 most frequent words in our training data.

Syntactic features include binary POS features of the source and target component and the 500 most frequent production rules extracted from the parse tree of the source and target component as described in our previous work (Stab and Gurevych 2014b).

Structural features consist of the number of tokens in the source and target component, statistics on the components of the covering paragraph of the current pair, and position features (Table 10).

Indicator features are based on the forward, backward, thesis, and rebuttal indicators introduced in Section 5.3.1. We extract binary features from the source and target component and the context of the current pair (Table 10). We assume that these features are helpful for modeling the direction of argumentative relations and the context of the current component pair.

Discourse features are extracted from the source and target component of each component pair as described in Section 5.3.1. Although PDTB-style discourse relations are limited to adjacent relations, we expect that the types of general discourse relations can be helpful for identifying argumentative relations. 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 are based on the assumption that particular words indicate incoming or outgoing relations. For instance, tokens like "therefore", "thus", or "hence" can signal incoming relations, whereas tokens such as "because", "since", or "furthermore"

Table 10
Features used for argumentative relation identification (*indicates genre-dependent features).

Group	*Feature*	*Description*
Lexical	Unigrams	Binary lemmatized unigrams of the source and target components including preceding tokens (500 most frequent)
Syntactic	Part-of-speech	Binary POS features of source and target components
	Production rules	Production rules extracted from the constituent parse tree (500 most frequent)
Structural	Token statistics	Number of tokens of source and target
	Component statistics	Number of components between source and target; number of components in covering paragraph
	Position features	Source and target present in same sentence; target present before source; source and target are first or last component in paragraph; pair present in introduction or conclusion*
Indicator	Indicator source/target	Indicator type present in source or target
	Indicators between	Indicator type present between source or target
	Indicators context	Indicator type follows or precedes source or target in the covering paragraph of the pair
Discourse	Discourse Triples	Binary discourse triples of source and target
PMI	Pointwise mutual information	Ratio of tokens positively or negatively associated with incoming or outgoing relations; Presence of words negatively or positively associated with incoming or outgoing relations
ShNo	Shared nouns	Shared nouns between source and target components (number and binary)

may indicate outgoing relations. To capture this information, we use Pointwise Mutual Information (PMI), which has been successfully used for measuring word associations (Turney 2002; Church and Hanks 1990). However, instead of determining the PMI of two words, we estimate the PMI between a lemmatized token t and the direction of a relation $d = \{incoming, outgoing\}$ as $PMI(t,d) = log\frac{p(t,d)}{p(t)\,p(d)}$. Here, $p(t,d)$ is the probability that token t occurs in an argument component with either incoming or outgoing relations. The ratio between $p(t,d)$ and $p(t)\,p(d)$ indicates the dependence between a token and the direction of a relation. We estimate $PMI(t,d)$ for each token in our training data. We extract the ratio of tokens positively and negatively associated with incoming or outgoing relations for both source and target components. Additionally, we extract four binary features, which indicate if any token of the components has a positive or negative association with either incoming or outgoing relations.

Shared noun features (shNo) indicate if the source and target components share a noun. We also add the number of shared nouns to our feature set. These features are motivated by the observation that claims and premises often share the same subject.

For selecting the best performing model, we conducted feature ablation tests and experimented with individual features. The results show that none of the feature groups is informative when used individually. We achieved the best performance by removing lexical features from our feature set (detailed results of the model selection can be found in Table C.3 in Appendix C).

5.3.3 Jointly Modeling Argumentative Relations and Argument Component Types. Both base classifiers identify argument component types and argumentative relations locally. Consequently, the results may not be globally consistent. For instance, the relation identification model does not link 37.1% of all premises in our model selection experiments. Therefore, we propose a joint model that globally optimizes the outcomes of the two base classifiers. We formalize this task as an Integer Linear Programming (ILP) problem. Given a paragraph including n argument components,[9] we define the following objective function

$$\underset{x}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_{ij} \tag{1}$$

with variables $x_{ij} \in \{0, 1\}$ indicating an argumentative relation from argument component i to argument component j.[10] Each coefficient $w_{ij} \in \mathbb{R}$ is a weight of a relation. It is determined by incorporating the outcomes of the two base classifiers. To ensure that the resulting structure is a tree, we define the following constraints:

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

$$\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \leq n - 1 \tag{3}$$

$$\forall i : x_{ii} = 0 \tag{4}$$

Equation (2) prevents an argument component i from having more than one outgoing relation. Equation (3) ensures that a paragraph includes at least one root node (i.e., a node without outgoing relation). Equation (4) prevents an argumentative relation from having the same source and target component.

To prevent cycles, we adopt the approach described by Kübler et al. (2008, page 92). We add the auxiliary variables $b_{ij} \in \{0, 1\}$ to our objective function (1) where $b_{ij} = 1$ if there is a directed path from argument component i to argument component j. The following constraints tie the auxiliary variables b_{ij} to the variables x_{ij}:

$$\forall i \, \forall j : x_{ij} - b_{ij} \leq 0 \tag{5}$$

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

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

The first constraint ensures that there is a path from i to j represented in variable b_{ij} if there is a direct relation between the argument components i and j. 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 argument component j ($b_{ij} = 1$) and another path from argument component j to argument component k ($b_{jk} = 1$) then there is also a path

9 We consider only claims and premises in our joint model, since argumentative relations between claims and major claims are modeled with a level approach (cf. Section 3.2).

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

from argument component i to argument component k. Thus, it iteratively covers paths of length $l + 1$ by having covered paths of length l. The third constraint avoids cycles by preventing all directed paths starting and ending with the same argument component.

Having defined the ILP model, we consolidate the results of the two base classifiers. We consider this task by determining the weight matrix $W \in \mathbb{R}^{n \times n}$ that includes the coefficients $w_{ij} \in W$ of our objective function. The weight matrix W can be considered an adjacency matrix. The greater the weight of a particular relation is, the higher the likelihood that the relation appears in the optimal structure found by the ILP-solver.

First, we incorporate the results of the relation identification model. Its result can be considered as an adjacency matrix $R \in \{0, 1\}^{n \times n}$. For each pair of argument components (i, j) with $1 \leq i, j \leq n$, 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.

Second, we derive a claim score cs_i for each argument component i from the predicted relations in R:

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

Here, $relin_i = \sum_{k=1}^{n} r_{ki}[i \neq k]$ is the number of predicted incoming relations of argument component i, $relout_i = \sum_{l=1}^{n} r_{il}[i \neq l]$ is the number of predicted outgoing 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 relations predicted in the current paragraph. The claim score cs_i is greater for argument components with many incoming relations and few outgoing relations. It becomes smaller for argument components with fewer incoming relations and more outgoing relations. By normalizing the score with the total number of predicted relations and argument components, it also accounts for contextual information in the current paragraph and prevents overly optimistic scores. For example, if all predicted relations point to 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 relation in a paragraph with four argument components and three predicted relations in R, cs_j is $\frac{1}{3}$. Because it is more likely that a 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 argumentative relation as:

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

By treating cs_j of the target component j as a positive term, we assign a higher weight to relations pointing to argument components that are likely to be a claim. 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.

Third, we incorporate the argument component types predicted by the classification model. We assign a higher score to the weight w_{ij} if the target component j is predicted as claim, because it is more likely that argumentative relations point to claims. Accordingly, we set $c_{ij} = 1$ if argument component j is labeled as claim and $c_{ij} = 0$ if argument component j is labeled as premise.

Finally, we combine all three scores to estimate the weights of the objective function:

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

Table 11
Features used for stance recognition.

Group	*Feature*	*Description*
Lexical	Unigrams	Binary and lemmatized unigrams of the component and its preceding token
Sentiment	Subjectivity clues	Presence of negative words; number of negative, positive, and neutral words; number of positive words subtracted by the number of negative words
	Sentiment scores	Five sentiment scores of covering sentence (Stanford sentiment analyzer)
Syntactic	POS distribution	POS distribution of the component
	Production rules	Production rules extracted from the constituent parse tree
Structural	Token statistics	Number of tokens of covering sentence; number of preceding and following tokens in covering sentence; ratio of component and sentence tokens
	Component statistics	Number of components in paragraph; number of preceding and following components in paragraph
	Component Position	Relative position of the argument component in paragraph
Discourse	Discourse Triples	PDTB discourse relations overlapping with the current component
Embedding	Combined word embeddings	Sum of the word vectors of each word of the component and its preceding tokens

Each ϕ represents a hyperparameter of the ILP model. In our model selection experiments, we found that $\phi_r = \frac{1}{2}$ and $\phi_{cr} = \phi_c = \frac{1}{4}$ yields the best performance. More detailed results of the model selection are provided in Table C.4 in Appendix C.

After applying the ILP model, we adapt the argumentative relations and argument types according to the results of the ILP-solver. We revise each relation according to the determined x_{ij} scores, set the type of all components without outgoing relation to "claim," and set the type of all remaining components to "premise."

5.4 Classifying Support and Attack Relations

The stance recognition model differentiates between argumentative support and attack relations. We model this task as binary classification and classify each claim and premise as "support" or "attack." The stance of each premise is encoded in the type of its outgoing relation, whereas the stance of each claim is encoded in its stance attribute. We use an SVM[11] and the features listed in Table 11.

5.5 Evaluation

Table 12 shows the F1 scores of the classification, relation identification, and stance recognition tasks using our test data. The ILP joint model significantly outperforms the macro F1 score of the heuristic baselines for component classification ($p = 1.49 \times 10^{-4}$)

11 For finding 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 SVM (Cortes and Vapnik 1995); we found that an SVM outperforms all other classifiers.

Table 12
Model assessment on persuasive essays († = significant improvement over baseline heuristic; ‡ = significant improvement over base classifier).

	Components				Relations			Stance recognition			
	F1	F1 MC	F1 Cl	F1 Pr	F1	F1 NoLi	F1 Li	F1	F1 Sup	F1 Att	Avg F1
Human upper bound	0.868	0.926	0.754	0.924	0.854	0.954	0.755	0.844	0.975	0.703	0.855
Baseline majority	0.260	0	0	0.780	0.455	0.910	0	0.478	0.957	0	0.398
Baseline heuristic	0.759	0.759	0.620	0.899	0.700	0.901	0.499	0.562	0.776	0.201	0.674
Base classifier	0.794	†0.891	0.611	0.879	0.717	0.917	0.508	†0.680	†0.947	†0.413	0.730
ILP joint model	†‡0.826	†0.891	‡0.682	‡0.903	†0.751	†0.918	†‡0.585	†0.680	†0.947	†0.413	0.752

and relation identification ($p = 0.003$). It also significantly outperforms the macro F1 score of the base classifier for component classification ($p = 7.45 \times 10^{-4}$). However, it does not yield a significant improvement over the macro F1 score of the base classifier for relation identification. The results show that the identification of claims and linked component pairs benefit most from the joint model. Compared with the base classifiers, the ILP joint model improves the F1 score of claims by 0.071 ($p = 1.84 \times 10^{-4}$) and the F1 score of linked component pairs by 0.077 ($p = 6.95 \times 10^{-5}$). The stance recognition model significantly outperforms the heuristic baseline by 0.118 macro F1 score ($p = 0.008$). It yields 0.947 F1 score for supporting components and 0.413 for attacking components.

The human upper bound yields macro F1 scores of 0.868 for component classification, 0.854 for relation identification, and 0.844 for stance recognition. The ILP joint model almost achieves human performance for classifying argument components. Its F1 score is only .042 lower than human upper bound. Regarding relation identification and stance recognition, the macro F1 scores of our model are 0.103 and 0.164 lower than human performance. Thus, our model achieves 95.2% of human performance for component identification, 87.9% for relation identification, and 80.5% for stance recognition.

In order to verify the effectiveness of our approach, we also evaluated the ILP joint model on the English microtext corpus (cf. Section 2.4). To ensure the comparability to previous results, we used the same repeated cross-validation set-up as described by Peldszus and Stede (2015). Because the microtext corpus does not include major claims, we removed the major claim label from our component classification model. Furthermore, it was necessary to adapt several features of the base classifiers, since the microtext corpus does not include non-argumentative text units. Therefore, we did not consider preceding tokens for lexical, indicator, and embedding features and removed

Table 13
Model assessment on microtext corpus from Peldszus and Stede (2015) († = significant improvement over baseline heuristic; ‡ = significant improvement over base classifier).

	Components			Relations			Stance recognition			
	F1	F1 Cl	F1 Pr	F1	F1 NoLi	F1 Li	F1	F1 Sup	F1 Att	Avg F1
Baseline heuristic	0.712	0.536	0.888	0.618	0.856	0.380	0.542	0.773	0.293	0.624
Base classifier	†0.830	†0.712	0.937	†0.650	†0.841	†0.446	†0.745	†0.855	†0.628	0.742
ILP joint model	†‡0.857	†‡0.770	†0.943	†‡0.683	†‡0.881	†‡0.486	†0.745	†0.855	†0.628	0.762
Best EG	0.869	-	-	0.693	-	0.502	0.710	-	-	0.757
MP+p	0.831	-	-	0.720	-	0.546	0.514	-	-	0.688

the probability feature of the component classification model. Additionally, we removed all genre-dependent features of both base classifiers.

Table 13 shows the evaluation results of our model on the microtext corpus. Our ILP joint model significantly outperforms the macro F1 score of the heuristic baselines for component classification ($p = 2.10 \times 10^{-10}$) and relation identification ($p = 1.5 \times 10^{-8}$). The results also show that our model yields significantly better macro F1 scores compared to the two base classifiers ($p = 0.002$ for component classification and $p = 7.52 \times 10^{-7}$ for relation identification). The stance recognition model achieves 0.745 macro F1 score on the microtext corpus. It significantly improves the macro F1 score of the heuristic baseline by 0.203 ($p = 7.55 \times 10^{-10}$).[12]

The last two rows in Table 13 show the results reported by Peldszus and Stede (2015) on the English microtext corpus. The Best EG model is their best model for component classification, and MP+p is their best model for relation identification. Compared with our ILP joint model, the Best EG model achieves better macro F1 scores for component classification and relation identification. However, because the outcomes of their systems are not available to us, we cannot determine if this difference is significant. The MP+p model achieves a better macro F1 score for relation identification, but yields lower results for component classification and stance recognition compared to our ILP joint model. These differences can be attributed to the additional information about the function and role attribute incorporated in their joint models (cf. Section 2.3). 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 corpus and the arguments in their corpus exhibit an unusually high proportion of attack relations. In particular, 86.6% of their arguments include attack relations, whereas the proportion of arguments with attack relations in our corpus amounts to only 12.4%. Therefore, we assume that incorporating function and role attributes will not be beneficial using our corpus.

Overall, the evaluation results show that our ILP joint model significantly outperforms challenging heuristic baselines and simultaneously improves the performance of component classification and relation identification on two different types of discourse.

5.6 Error Analysis

In order to analyze frequent errors of the ILP joint model, we investigated the predicted argumentation structures in our test data. The confusion matrix of the component classification task (Table 14) shows that the highest confusion is between claims and premises. The model classifies 74 actual premises as claims and 82 claims as premises. By manually investigating these errors, we found that the model tends to label inner premises in serial structures as claims and wrongly identifies claims in sentences containing two premises. Regarding the relation identification, we observed that the model tends to identify argumentation structures that are more shallow than the structures in our gold standard. The model correctly identifies only 34.7% of the 98 serial arguments in our test data. This can be attributed to the "claim-centered" weight calculation in our objective function. In particular, the predicted relations in matrix R are the only information about serial arguments, whereas the other two scores (c_{ij} and cr_{ij}) assign higher weights to relations pointing to claims.

12 The heuristic baseline for stance recognition on the microtext corpus classifies the fourth component as "attack" and all other components as "support."

Table 14
Confusion matrix of the ILP joint model of component classification on our test data.

		predictions		
		MajorClaim	Claim	Premise
actual	MajorClaim	**139**	12	2
	Claim	20	**202**	82
	Premise	0	74	**735**

In order to determine if the ILP joint model correctly models the relationship between component types and argumentative relations, we artificially improved the predictions of both base classifiers as suggested by Peldszus and Stede (2015). The dashed lines in Figure 4 show the performance of the artificially improved base classifiers. Continuous lines show the resulting performance of the ILP joint model. Figures 4a and 4b show the effect of improving the component classification and relation identification. They show that correct predictions of one base classifier are not maintained after applying the ILP model if the other base classifier exhibits less accurate predictions. In particular, less accurate argumentative relations have a more detrimental effect on the component types (Figure 4a) than less accurate component types do on the outcomes of the relation identification (Figure 4b). Thus, it is more reasonable to focus on improving relation identification than component classification in future work.

Figure 4c depicts the effect of improving both base classifiers, which illustrates that the ILP joint model improves the component types more effectively than argumentative relations. Figure 4c shows that the ILP joint model improves both tasks if the base classifiers are improved. Therefore, we conclude that the ILP joint model successfully captures the natural relationship between argument component types and argumentative relations.

6. Discussion

Our argumentation structure parser is a pipeline consisting of several consecutive steps. Therefore, potential errors of the upstream models are propagated and

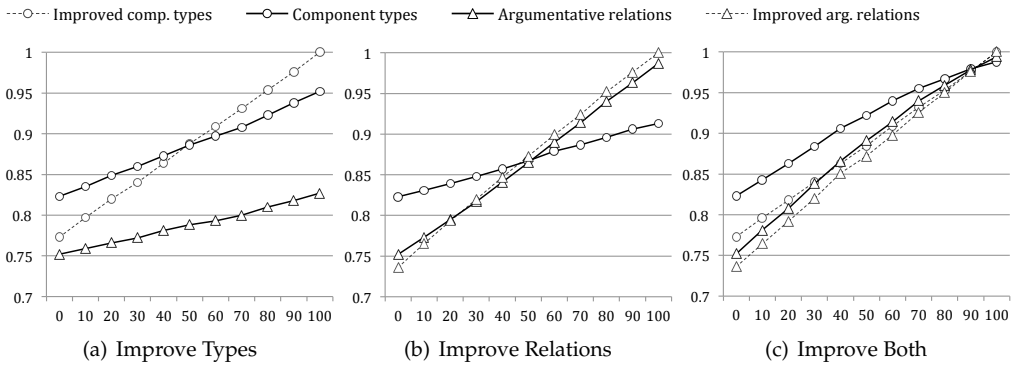

Figure 4
Influence of improving the base classifiers (*x*-axis shows the proportion of improved predictions and *y*-axis the macro F1 score).

negatively influence the results of the downstream models. For example, errors of the identification model can result in flawed argumentation structures if argumentatively relevant text units are not recognized or non-argumentative text units are identified as relevant. However, our identification model yields good accuracy and an α_U of 0.958 for identifying argument components. Therefore, it is unlikely that identification errors will significantly influence the outcome of the downstream models when applied to persuasive essays. However, as demonstrated by Levy et al. (2014) and Goudas et al. (2014), the identification of argument components is more complex in other text genres than it is in persuasive essays. Another potential issue of the pipeline architecture is that wrongly classified major claims will decrease the accuracy of the model because they are not integrated in the joint modeling approach. For this reason, it is worthwhile to experiment in future work with structured machine learning methods that incorporate several tasks in one model (Moens 2013).

In this work, we presented an approach for recognizing argumentation structures in persuasive essays. Other text genres, however, may exhibit less explicit arguments. Habernal and Gurevych (2017, page 27), for instance, showed that 48% of the arguments in user-generated Web discourse do not include explicit claims. These incomplete arguments, so called **enthymemes**, make both annotation and automatic analysis challenging. Although humans may be able to deduce the missing parts by interpreting the argument, existing argument mining methods fail on that task and may produce incomplete or even wrong argumentation structures. In particular, the presented approach is not able to recognize gaps in reasoning (i.e., missing premises) or to infer the missing components of implicit arguments. Inferring implicit argument components is challenging since it requires robust methods for capturing the semantics of natural language arguments and appropriate background knowledge for reconstructing the missing parts.

The presented argumentation structure parser is an important milestone for implementing argumentative writing support systems. For example, the recognized argumentation structures allow highlighting unwarranted claims, missing major claims, or different types of quantitative analyses on the number of arguments or their premises. It is still unknown, however, if this feedback provides an adequate guidance for improving students' argumentation skills. In order to answer this question, it is required to integrate the proposed model in writing environments and to investigate the effect of different feedback types on the argumentation skills of students in future research.

7. Conclusion

In this article, we presented an 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 argumentation structure parsing or rely on manually created rules. Consequently, previous approaches are only of limited use for parsing argumentation structures in real application scenarios. To the best of our knowledge, the presented work is the first approach that covers all required subtasks for identifying the global argumentation structure of documents. We showed that jointly modeling argumentation structures simultaneously improves the results of component classification and relation identification. Additionally, we introduced a novel annotation scheme and a new corpus of persuasive essays annotated with argumentation structures that represent the largest resource of its kind. Both the corpus and the annotation guidelines are freely available.

Appendix A. Class Distributions

Table A.1 shows the class distributions of the training and test data of the persuasive essay corpus for each analysis step.

Appendix B. Indicators

Table B.1 shows all of the lexical indicators we extracted from 30 persuasive essays. The lists include 24 forward indicators, 33 backward indicators, 48 thesis indicators, and 10 rebuttal indicators.

Appendix C. Detailed Results of Model Selections

The following tables show the model selection results for all five tasks using 5-fold cross-validation on our training data. Table C.1 shows the results of using individual feature groups for the argument component identification task. Lexico-syntactic features perform best regarding the macro F1 score, and they perform particularly well for recognizing the beginning of argument components ("Arg-B"). The second best features are structural features. They yield the best F1 score for separating argumentative from non-argumentative text units ("O").

Syntactic features are useful for identifying the beginning of argument components. The probability feature yields the lowest macro F1 score. Nevertheless, we observe a significant decrease compared with the macro F1 score of the model with all features when evaluating the system without the probability feature ($p = 0.003$). We obtain the best results by using all features. Because persuasive essays exhibit a particular paragraph structure, which may not be present in other text genres (e.g., user-generated Web discourse), we also evaluate the model without genre-dependent features (cf. Table 7). This yields a significant difference compared with the macro F1 score of the model with all features ($p = 2.24 \times 10^{-54}$).

Table A.1
Class distributions in training data and test data.

Class	Training data		Test data	
	Identification			
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%)
	Component classification			
MajorClaim	598	(12.4%)	153	(12.1%)
Claim	1,202	(24.9%)	304	(24.0%)
Premise	3,023	(62.7%)	809	(63.9%)
	Relation identification			
Not-Linked	14,227	(82.5%)	4,113	(83.5%)
Linked	3,023	(17.5%)	809	(16.5%)
	Stance recognition			
Support	3,820	(90.4%)	1,021	(91.7%)
Attack	405	(9.6%)	92	(8.3%)

Table B.1
List of lexical indicators.

Category	Indicators
Forward (24)	"As a result", "As the consequence", "Because", "Clearly", "Consequently", "Considering this subject", "Furthermore", "Hence", "leading to the consequence", "so", "So", "taking account on this fact", "That is the reason why", "The reason is that", "Therefore", "therefore", "This means that", "This shows that", "This will result", "Thus", "thus", "Thus, it is clearly seen that", "Thus, it is seen", "Thus, the example shows"
Backward (33)	"Additionally", "As a matter of fact", "because", "Besides", "due to", "Finally", "First of all", "Firstly", "for example", "For example", "For instance", "for instance", "Furthermore", "has proved it", "In addition", "In addition to this", "In the first place", "is due to the fact that", "It should also be noted", "Moreover", "On one hand", "On the one hand", "On the other hand", "One of the main reasons", "Secondly", "Similarly", "since", "Since", "So", "The reason", "To begin with", "To offer an instance", "What is more"
Thesis (48)	"All in all", "All things considered", "As far as I am concerned", "Based on some reasons", "by analyzing both the views", "considering both the previous fact", "Finally", "For the reasons mentioned above", "From explanation above", "From this point of view", "I agree that", "I agree with", "I agree with the statement that", "I believe", "I believe that", "I do not agree with this statement", "I firmly believe that", "I highly advocate that", "I highly recommend", "I strongly believe that", "I think that", "I think the view is", "I totally agree", "I totally agree to this opinion", "I would have to argue that", "I would reaffirm my position that", "In conclusion", "in conclusion", "in my opinion", "In my opinion", "In my personal point of view", "in my point of view", "In my point of view", "In summary", "In the light of the facts outlined above", "it can be said that", "it is clear that", "it seems to me that", "my deep conviction", "My sentiments", "Overall", "Personally", "the above explanations and example shows that", "This, however", "To conclude", "To my way of thinking", "To sum up", "Ultimately"
Rebuttal (10)	"Admittedly", "although", "Although", "besides these advantages", "but", "But", "Even though", "even though", "However", "Otherwise"

Table C.1
Argument component identification († = significant improvement over baseline heuristic).

	F1	P	R	F1 Arg-B	F1 Arg-I	F1 O
Baseline majority	0.259	0.212	0.333	0	0.778	0
Baseline heuristic	0.628	0.647	0.610	0.350	0.869	0.660
CRF only structural	†0.748	†0.757	†0.740	†0.542	†0.906	†0.789
CRF only syntactic	†0.730	†0.752	†0.710	†0.638	0.868	0.601
CRF only lexSyn	†0.762	†0.780	†0.744	†0.714	†0.873	0.620
CRF only probability	0.605	†0.698	0.534	†0.520	0.806	0.217
CRF w/o genre-dependent	†0.847	†0.851	†0.844	†0.778	†0.925	†0.835
CRF all features	†0.849	†0.853	†0.846	†0.777	†0.927	†0.842

Table C.2 shows the model selection results of the classification model. Structural features are the only features that significantly outperform the macro F1 score of the heuristic baseline when used individually ($p = 4.04 \times 10^{-6}$). They are the most effective features for identifying major claims and claims. The second-best features for identifying claims are discourse features. With this knowledge, we can confirm the assumption that general discourse relations are useful for component classification (cf. Section 5.3.1). Embedding features do not perform as well as lexical features. They yield lower F1 scores for major claims and claims. Contextual features are effective for identifying major claims, since they implicitly capture if an argument component is present in

Table C.2
Argument component classification († = significant improvement over baseline heuristic).

	F1	P	R	F1 MajorClaim	F1 Claim	F1 Premise
Baseline majority	0.257	0.209	0.333	0	0	0.771
Baseline heuristic	0.724	0.724	0.723	0.740	0.560	0.870
SVM only lexical	0.591	0.603	0.580	0.591	0.405	0.772
SVM only structural	†0.746	0.726	†0.767	†0.803	0.551	0.870
SVM only contextual	0.601	0.603	0.600	0.656	0.248	0.836
SVM only indicators	0.508	0.596	0.443	0.415	0.098	0.799
SVM only syntactic	0.387	0.371	0.405	0.313	0	0.783
SVM only probability	0.561	0.715	0.462	0.448	0.002	0.792
SVM only discourse	0.521	0.563	0.484	0.016	0.538	0.786
SVM only embeddings	0.588	0.620	0.560	0.560	0.355	0.815
SVM all w/o prob & emb	†0.771	†0.771	†0.772	†0.855	0.596	0.863
SVM w/o genre-dependent	†0.742	†0.745	0.739	†0.819	0.560	0.847
SVM all features	†0.773	†0.774	†0.771	†0.865	0.592	0.861

the introduction or conclusion (cf. Section 5.3.1). Indicator features are most effective for identifying major claims, but contribute only slightly to the identification of claims. Syntactic features are predictive of major claims and premises, but are not effective for recognizing claims. The probability features are not informative for identifying claims, probably because forward indicators may also signal inner premises in serial structures. Omitting probability and embedding features yields the best accuracy. However, we select the best system by means of the macro F1 score, which is more appropriate for imbalanced data sets. Accordingly, we select the model with all features (Table C.2).

The model selection results for relation identification are shown in Table C.3. We report the results of feature ablation tests, since none of the feature groups yields remarkable results when used individually. Structural features are the most effective features for identifying relations. The second- and third-most effective feature groups are indicator and PMI features. Removing the shared noun feature does not yield a significant difference in accuracy or macro F1 score compared with SVM all features. We achieve the best macro F1 score by removing lexical features from the feature set.

Table C.3
Argumentative relation identification († = significant improvement over baseline heuristic; ‡ = significant difference compared to SVM all features).

	F1	P	R	F1 Not-Linked	F1 Linked
Baseline majority	0.455	0.418	0.500	0.910	0
Baseline heuristic	0.660	0.657	0.664	0.885	0.436
SVM all w/o lexical	†0.736	†‡0.762	†0.711	†‡0.917	†0.547
SVM all w/o syntactic	†0.729	†‡0.764	†0.697	†‡0.917	†0.526
SVM all w/o structural	†‡0.715	†‡0.740	‡0.692	†‡0.911	‡0.511
SVM all w/o indicators	†‡0.719	†‡0.743	†‡0.697	†‡0.912	†‡0.520
SVM all w/o discourse	†0.732	†0.755	†0.709	†0.915	†0.540
SVM all w/o pmi	†‡0.720	†‡0.745	†‡0.697	†‡0.912	†‡0.521
SVM all w/o shNo	†0.733	†0.756	†0.712	†0.915	†0.545
SVM w/o genre-dependent	†0.722	†0.750	†0.700	†0.913	†0.520
SVM all features	†0.733	†0.756	†0.711	†0.915	†0.544

Table C.4
Joint modeling approach († = significant improvement over base heuristic; ‡ = significant improvement over base classifier; Cl→Pr = number of claims converted to premises; Pr→Cl = number of premises converted to claims; Trees = Percentage of correctly identified trees).

	Parameter			Components				Relations			Statistics		
	ϕ_r	ϕ_{cr}	ϕ_c	F1	F1 MC	F1 Cl	F1 Pr	F1	F1 NoLi	F1 Li	Cl→Pr	Pr→Cl	Trees
Base heuristic	-	-	-	0.724	0.740	0.560	0.870	0.660	0.885	0.436	-	-	100%
Base classifier	-	-	-	†0.773	†0.865	0.592	0.861	†0.736	†0.917	†0.547	-	-	20.9%
Base+heuristic	-	-	-	†0.776	†0.865	0.601	0.861	†0.739	†0.917	†0.555	0	31	24.2%
ILP-naïve	1	0	0	†0.765	†0.865	†0.591	0.761	†0.732	†‡0.918	†0.530	206	1,144	100%
ILP-relation	$\frac{1}{2}$	$\frac{1}{2}$	0	†‡0.809	†0.865	†‡0.677	‡0.875	†‡0.759	†‡0.919	†‡0.598	299	571	100%
ILP-claim	0	0	1	†0.740	†0.865	0.549	0.777	0.666	0.894	0.434	229	818	100%
ILP-equal	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	†‡0.822	†0.865	†‡0.699	†‡0.903	†‡0.751	†0.913	†‡0.590	294	280	100%
ILP-same	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	†‡0.817	†0.865	†‡0.687	†‡0.898	†‡0.738	†0.908	†‡0.569	264	250	100%
ILP-balanced	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	†‡0.823	†0.865	†‡0.701	†‡0.904	†‡0.752	†0.913	†‡0.591	297	283	100%

Table C.4 shows the model selection results of the ILP joint model. *Base+heuristic* shows the result of applying the baseline to all paragraphs in which the base classifiers identify neither claims nor argumentative relations. The heuristic baseline is triggered in 31 paragraphs, which results in 3.3% more trees identified compared with the base classifiers. However, the difference between *Base+heuristic* and the base classifiers is not statistically significant. For this reason, we can attribute any further improvements to the joint modeling approach. Moreover, Table C.4 shows selected results of the hyperparameter tuning of the ILP joint model. Using only predicted relations in the ILP-naïve model does not yield an improvement compared with the macro F1 score of the base classifiers. ILP-relation uses only information from the relation identification base classifier. It significantly outperforms the macro F1 score of both base classifiers ($p = 6.43 \times 10^{-12}$ for relations and $p = 7.23 \times 10^{-13}$ for components), but converts a large number of premises to claims. The ILP-claim model uses only the outcomes of the argument component base classifier and improves neither component classification nor relation identification. All three models identify a relatively high proportion of claims compared to the number of claims in our training data. The reason for this is that many weights in W are 0. Combining the results of both base classifiers yields a more balanced proportion of component type conversions. All three models (ILP-equal, ILP-same, and ILP-balanced) significantly outperform the macro F1 score of the base classifiers. We identify the best performing system by means of the average macro F1 score for both tasks. Accordingly, we select ILP-balanced as our best performing ILP joint model.

Table C.5 shows the model selection results for the stance recognition model. Using sentiment, structural, and embedding features individually does not yield an improvement over the majority baseline. Lexical features yield a significant improvement over the macro F1 score of the heuristic baseline when used individually ($p = 8.02 \times 10^{-10}$). Syntactic features significantly improve precision ($p = 1.81 \times 10^{-30}$), recall ($p = 1.95 \times 10^{-47}$), F1 Support ($p = 1.01 \times 10^{-27}$), and F1 Attack ($p = 1.53 \times 10^{-54}$) over the heuristic baseline, but do not yield a significant improvement over the macro F1 score of the heuristic baseline. Discourse features significantly outperform the heuristic baseline regarding precision ($p = 3.68 \times 10^{-28}$), recall ($p = 3.43 \times 10^{-49}$), and F1 Support ($p = 1.06 \times 10^{-32}$). Because omitting any of the feature groups yields a lower macro F1 score, we select the model with all features as the best performing model.

Table C.5
Stance recognition († = significant improvement over baseline heuristic; ‡ = significant difference compared to SVM all features).

	F1	P	R	F1 Support	F1 Attack
Baseline majority	0.475	0.452	0.500	0.950	0
Baseline heuristic	0.521	0.511	0.530	0.767	0.173
SVM only lexical	†0.663	†0.677	†0.650	†0.941	†0.383
SVM only syntactic	0.649	†0.725	†0.587	†0.950	†0.283
SVM only discourse	0.630	†0.746	†0.546	†0.951	0.169
SVM all w/o lexical	†0.696	†‡0.719	†0.657	†‡0.948	†‡0.439
SVM all w/o syntactic	†0.687	†‡0.691	†‡0.684	†‡0.941	†‡0.433
SVM all w/o sentiment	†0.699	†‡0.710	†0.688	†‡0.945	†‡0.451
SVM all w/o structural	†0.698	†‡0.710	†0.686	†‡0.946	†‡0.449
SVM all w/o discourse	†0.675	†‡0.685	†‡0.666	†‡0.941	†‡0.408
SVM all w/o embeddings	†0.692	†‡0.703	†‡0.682	†‡0.944	†‡0.439
SVM all features	†0.702	†0.714	†0.690	†0.946	†0.456

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 would like to thank the anonymous reviewers for their valuable feedback; Can Diehl, Ilya Kuznetsov, Todd Shore, and Anshul Tak for their valuable contributions; and Andreas Peldszus for providing details about his corpus.

References

Afantenos, Stergos and Nicholas Asher. 2014. Counter-argumentation and discourse: A case study. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 11–16, Bertinoro.

Afantenos, Stergos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon.

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.

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

Asher, Nicholas and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

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

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.

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, 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, 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, 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, MD.

Botley, Simon Philip. 2014. Argument structure in learner writing: a corpus-based analysis using argument mapping. *Kajian Malaysia*, 32(1):45–77.

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.

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

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, Elena and Serena Villata. 2012. Natural language arguments: A combined approach. In *Proceedings of the 20th European Conference on Artificial Intelligence*, ECAI '12, pages 205–210, Montpellier.

Cabrio, Elena and Serena Villata. 2014. NoDE: A benchmark of natural language arguments. In *Proceedings of COMMA*, pages 449–450, Pitlochry.

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.

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.

Eckart de Castilho, Richard and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin.

le Cessie, S. and J. C. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

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.

Cohen, Robin. 1987. Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13(1-2):11–24.

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.

Collins, Michael. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.

Conway, David A. 1991. On the distinction between convergent and linked arguments. *Informal Logic*, 13:145–158.

Copi, Irving M. and Carl Cohen. 1990. *Introduction To Logic*, 8th edition. Macmillan Publishing Company.

Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Damer, T. Edward. 2009. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Reasoning*, 6th edition. Wadsworth Cengage Learning.

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.

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.

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.

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, MD.
Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
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.
Forman, George and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explorations*, 12(1):49–57.
Freeman, James B. 2011. *Argument Structure: Representation and Theory*, volume 18 of *Argumentation Library*. Springer.
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, MD.
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.
Govier, Trudy. 2010. *A Practical Study of Argument*, 7th edition. Wadsworth, Cengage Learning.
Habernal, Ivan and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.
Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.
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.
Henkemans, A. Francisca Snoeck. 2000. State-of-the-art: The structure of argumentation. *Argumentation*, 14(4):447–473.
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.
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.
Johnson, Ralph H. 2000. *Manifest Rationality*. Lawrence Erlbaum.
Kemper, Dave and Pat Sebranek. 2004. *Inside Writing: Persuasive Essays*. Great Source Education Group.
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*, 1–11, Denver, CO.
Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Sapporo.
Krippendorff, Klaus. 2004. Measuring the reliability of qualitative text analysis data. *Quality & Quantity*, 38(6):787–800.
Kübler, Sandra, Ryan McDonald, Joakim Nivre, and Graeme Hirst. 2008. *Dependency Parsing*. Morgan and Claypool Publishers.
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.
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, 282–289, San Francisco, CA.
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)*, 1489–1500, Dublin.
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.

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, 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.

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.

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

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.

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, 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.

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.

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

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.

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.

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.

O'Keefe, Daniel J. 1977. Two concepts of argument. *Journal of the American Forensic Association*, 13(3):121–128.

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.

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, MD.

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

Peldszus, Andreas and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

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.

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.

Perutz, Vivien. 2010. *A Helpful Guide to Essay Writing!*, Student Services, Anglia Ruskin University.

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.

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.

Quinlan, Ross. 1993. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

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.

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.

Reed, Chris and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 14(4):961–980.

Rinott, Ruty, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence—an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 440–450, Lisbon.

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.

Sardianos, Christos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO.

Shiach, Don. 2009. *How to write essays*, 2nd ed. How To Books Ltd.

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.

Søgaard, Anders. 2013. Estimating effect size across datasets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 607–611, Atlanta.

Sokolova, Marina and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

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.

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, MD.

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.

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.

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.

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.

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.

Thomas, Stephen N. 1973. *Practical Reasoning in Natural Language*, Prentice-Hall.

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.

Turney, Peter D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 417–424, Philadelphia, PA.

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 Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 23–25, Istanbul.

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

Whitaker, Anne. 2009. *Academic Writing Guide 2010: A Step-by-Step Guide to Writing Academic Papers*. City University of Seattle.

Wolfe, Christopher R. and M. Anne Britt. 2009. Argumentation schema and the myside bias in written argumentation. *Written Communication*, 26(2):183–209.

Yanal, Robert J. 1991. Dependent and independent reasons. *Informal Logic*, 13(3):137–144.