# Scientia Potentia Est—On the Role of Knowledge in Computational Argumentation

**Anne Lauscher**[1*], **Henning Wachsmuth**[2*], **Iryna Gurevych**[3], **and Goran Glavaš**[4]
[1]MilaNLP, Bocconi University, Italy
[2]Department of Computer Science, Paderborn University, Germany
[3]Ubiquitous Knowledge Processing Lab, TU Darmstadt, Germany
[4]CAIDAS, University of Würzburg, Germany
anne.lauscher@unibocconi.it, henningw@upb.de,
gurevych@ukp.informatik.tu-darmstadt.de, goran.glavas@uni-wuerzburg.de

## Abstract

Despite extensive research efforts in recent years, computational argumentation (CA) remains one of the most challenging areas of natural language processing. The reason for this is the inherent complexity of the cognitive processes behind human argumentation, which integrate a plethora of different types of knowledge, ranging from topic-specific facts and common sense to rhetorical knowledge. The integration of knowledge from such a wide range in CA requires modeling capabilities far beyond many other natural language understanding tasks. Existing research on mining, assessing, reasoning over, and generating arguments largely acknowledges that much more knowledge is needed to accurately model argumentation computationally. However, a systematic overview of the types of knowledge introduced in existing CA models is missing, hindering targeted progress in the field. Adopting the operational definition of knowledge as any task-relevant normative information not provided as input, the survey paper at hand fills this gap by (1) proposing a taxonomy of types of knowledge required in CA tasks, (2) systematizing the large body of CA work according to the reliance on and exploitation of these knowledge types for the four main research areas in CA, and (3) outlining and discussing directions for future research efforts in CA.

## 1 Introduction

The phenomenon of argumentation, a direct reflection of human reasoning in natural language, has fascinated scholars across societies and cultures since ancient times (Aristotle, ca. 350 B.C.E./ translated 2007; Lloyd, 2007). The computational modeling of human argumentation, commonly referred to as *computational argumentation (CA)*, has evolved into one of the most prominent and at the same time most challenging areas in natural language processing (NLP) (Lippi and Torroni, 2015).

CA encompasses several families of tasks and research directions, the main ones in NLP being argument *mining*, *assessment*, *reasoning*, and *generation*. Although it bears some resemblance to other NLP tasks, such as opinion mining and natural language inference (NLI), it is widely acknowledged to be of much higher difficulty than the other tasks (Habernal et al., 2014). While opinion mining (Liu, 2012) assesses stances towards entities or controversies by asking *what* the opinions are, CA provides answers to a more difficult question: *Why* is the stance of an opinion holder the way it is? In a similar vein, while NLI focuses on detecting simple entailments between statement pairs (Bowman et al., 2015; Dagan et al., 2013), CA addresses more complex reasoning scenarios that involve multiple entailment steps, often over implicit premises (Boltužić and Šnajder, 2016).

CA targets reasoning processes that are only partially explicated in text. Its mastery thus requires advanced natural language understanding capabilities and a substantial amount of background knowledge (Moens, 2018; Paul et al., 2020). For example, the assessment of an argument's quality not only depends on the actual content of an argumentative text or speech but also on social and cultural context, such as speaker and audience characteristics, including their individual values, ideologies, and relationships (Wachsmuth

*Equal contribution.



Transactions of the Association for Computational Linguistics, vol. 10, pp. 1392–1422, 2022. https://doi.org/10.1162/tacl_a_00525
Action Editor: Mark Steedman. Submission batch: 4/2022; Revision batch: 7/2022; Published 12/2022.

et al., 2017a). Such contextual information remains most often implicit. For any concrete CA task, we here refer to all information that is not explicitly provided as input to models tackling the task but is (potentially) useful for it and (in most cases) normative in nature as *knowledge* (we detail this notion in §3.1).

Although there is ample awareness of the need for integrating various types of knowledge in CA models in the research community, there is no systematic overview of the types of knowledge that existing models and solutions for the different CA tasks rely on. This impedes targeted progress in pressing subareas of CA, such as argument generation. While general surveys on CA (e.g., Cabrio and Villata, 2018; Lawrence and Reed, 2020) and its subareas (e.g., Al Khatib et al., 2021; Schaefer and Stede, 2021) represent good starting points for targeted research along these lines, they lack a systematic analysis of the roles that different types of knowledge play in different CA tasks.

**Contributions.** In this work, we aim to systematically inform the research community about the types of knowledge that have—or have not yet—been integrated into computational models in different CA tasks. For this purpose, we **(1)** propose a pyramid-like taxonomy systematizing the relevant types of knowledge. The pyramid is organized by knowledge specificity, from linguistic knowledge and world and topic knowledge to argumentation-specific and task-specific knowledge. Starting from 162 CA publications, we **(2)** survey the existing body of work with respect to the level of integration of the various types of knowledge and respective methodology by which the knowledge of each type is integrated into models. To this end, we carry out an expert annotation study in which we manually label individual papers with the types from the knowledge pyramid. Finally, we **(3)** identify trends and challenges in the four most prominent CA subareas (mining, assessment, reasoning, and generation), summarizing them into three key recommendations for future CA research:

1. *All CA tasks are expected to benefit from more modeling of world and topic knowledge.* Although several studies report empirical gains from incorporating these types of knowledge, their inclusion is still an exception rather than a rule across the landscape of all CA tasks.

2. *Argument mining tasks are expected to benefit from more modeling of argumentation- and task-specific knowledge.* Such specialized knowledge has been proven effective in assessment, reasoning, and generation tasks. Yet, it has so far been exploited only sporadically in argument mining approaches.

3. *All CA tasks are expected to benefit from applying key techniques to other types of knowledge and data.* As an example, methods that represent symbolic input in a semantic vector space (e.g., pretrained word embeddings or language models) are still rarely applied to sources other than text (e.g., to knowledge bases). The bottleneck to a wider application of general-purpose techniques such as representation learning in CA is the lack of structured knowledge resources. We thus argue that significant progress in CA critically hinges on the availability of such resources at larger scale. Accordingly, based on the results of this survey effort, we strongly encourage the CA community to foster the creation of knowledge-rich argumentative corpora.

**Structure.** We start with an overview of the field of CA and its four most prominent subareas (§2). In §3, we describe our survey methodology, before we establish the knowledge pyramid and present the results of the survey with respect to the types of knowledge from the pyramid (§4). On this basis, we summarize emerging trends (§5) and offer recommendations for future progress in CA (§6).

## 2 Background

The study of argumentation in Western societies can be traced back to Ancient Greece. With the development of democracy and, thereby, the need to influence public decisions, the art of convincing others became an essential skill for successful participation in the democratic process (Aristotle, ca. 350 B.C.E./ translated 2007). In that period, rhetorical theories also started appearing in Eastern societies and cultures, such as *Nyaya Sutra* (Lloyd, 2007). Since then, a plethora of phenomena in the realm of argumentation, such

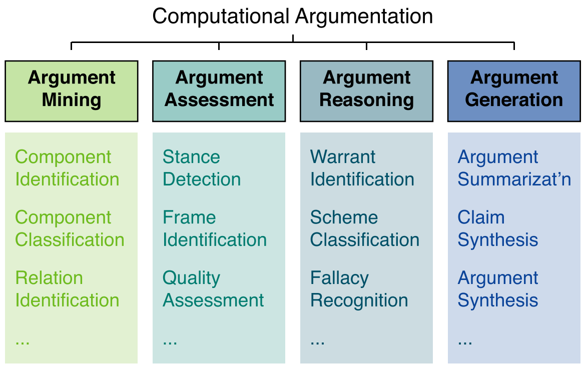


Figure 1: The four main subareas of computational argumentation (argument mining, argument assessment, argument reasoning, and argument generation) with three of their most prominent respective tasks each.

as fallacies (Hamblin, 1970) and argumentation schemes (Walton et al., 2008), have been studied extensively, usually focusing on specific domains, such as science (Gilbert, 1977) and law (Toulmin, 2003).

With the growing amount of argumentative data available publicly in Web debates, scientific articles, and other Internet sources, the computational modeling of argumentation, *computational argumentation (CA)*, gradually gained prominence and popularity in the NLP community. As depicted in Figure 1, CA can be divided into four main subareas that represent the main high-level types of tasks being tackled with computational models: mining, assessment, reasoning, and generation.

**Argument Mining.** Argument mining deals with the extraction of argumentative structures from natural language text (e.g., Stab and Gurevych, 2017a). Traditionally, it has been addressed with a pipeline of models each tackling one analysis task, most commonly *component identification*, *component classification*, and *relation identification* (Lippi and Torroni, 2015). The set of argument components and relations is defined by the selected underlying argument model which reflects the rhetorical, dialogical, or monological structure of argumentation (Bentahar et al., 2010b).

For instance, the model of Toulmin (2003), designed for the legal domain, encompasses six components: a claim with an optional qualifier, data (i.e., a fact supporting the claim) connected to the claim via a warrant (i.e., the reason why support is given) and its backing, and a rebuttal (i.e., a counterconsideration to the claim). Relations model the support or attack of components (or arguments) by others, sometimes with more fine-grained subtypes (Freeman, 2011). In contrast to argument reasoning (see below), the information needed for inferring argumentative relations is contained in the text.

**Argument Assessment.** Computational models that address tasks in this subarea typically focus on particular properties of arguments in their context and automatically assign discrete or numeric labels for these properties. This includes the *classification of stance* towards some target (Bar-Haim et al., 2017a) as well as the *identification of frames* (or aspects) covered by the argument (Ajjour et al., 2019). Arguably, the most popular family of tasks belongs to argument *quality assessment*, which has been studied under various conceptualizations, such as clarity (Persing and Ng, 2013) or convincingness (Habernal and Gurevych, 2016b). Wachsmuth et al. (2017a) propose a taxonomy that divides the overall quality of an argument into three complementary aspects: logic, rhetoric, and dialectic. Each of these three aspects further consists of several quality dimensions (e.g., the dimension of global acceptability for the dialectical aspect).

**Argument Reasoning.** In this subarea, the task is to understand the reasoning process behind an argument. In NLP, reasoning is instantiated in tasks such as predicting the entailment relationship between a premise and a hypothesis by means of natural language inference (Williams et al., 2017), or the more complex task of *warrant identification*, that is, to find (or even reconstruct) the missing warrant (Tian et al., 2018). Others have tried to *classify schemes* of inferences happening in arguments (Feng and Hirst, 2011) or to *recognize fallacies* of certain reasoning types in arguments, such as the common ad-hominem fallacy (Habernal et al., 2018c; Delobelle et al., 2019).

In argument reasoning, the challenge lies in inducing additional knowledge—not explicated in the text—from existing components, as opposed to relation identification, which focuses on recognizing argumentative content present in the text. In other words, argument mining structures explicated arguments and their connections, whereas argument reasoning infers knowledge missing from the text (e.g., a warrant that connects the premise to the claim). In practice, however, there is no guarantee that annotators for argument mining tasks (e.g., relation identification) do not resort to out-of-text reasoning, leveraging their commonsense and world knowledge to perform the task. However, from a structural point of view, a premise may still be given by an author to support a claim (e.g., indicated by lexical cues like *because*), while from a reasoning perspective, the premise might be irrelevant to the claim (e.g., the claim does not logically follow from the given premise).

**Argument Generation.** With conversational AI (i.e., dialogue systems) arguably becoming the most prominent application in modern NLP and AI, the research efforts on generating argumentative language have also been gaining traction. Main tasks in argument generation include the *summarization of arguments* given (Wang and Ling 2016), the *synthesis of new claims* and other argument components (Bilu and Slonim, 2016), and the *synthesis of entire arguments*, possibly conforming to some rhetorical strategy (El Baff et al., 2019).

The impact of argument generation is, for example, demonstrated by Project Debater (Slonim, 2018), a well-known argumentation system which combines models for several generation tasks.

## 3 Methodology

In this section, we first provide the definition of knowledge upon which we base this work. Then, we detail the methodology that we devised and pursued in order to organize the types of knowledge that CA approaches and models utilize.

### 3.1 An Operational Definition of Knowledge

Various definitions of ''knowledge'' have been proposed in the literature. One of the oldest is the tripartite definition of Plato (ca. 400 B.C.E.), who accepted as knowledge any *justified true belief*. This definition was later often challenged as being too narrow and was, accordingly, extended (e.g., Goldman, 1967; Hawthorne, 2002). As part of this effort, Dretske (1981) dressed Plato's view into an information-theoretic gown, defining knowledge as *information-caused belief*, specifying more narrowly the informational source of the belief as the only valid justification and de facto eliminating the veracity constraint.

Departing from attempts to define knowledge ontologically, Gottschalk-Mazouz (2013) adopted an impact-based viewpoint and argue that it is more important to understand *what knowledge can do* and *what it is like* than to ontologically answer *what knowledge is*. In their view, knowledge is thus normative and has practical implications. In the work at hand, we adopt this impact-oriented view on knowledge. We further operationalize the view, in the context of NLP and CA, as follows:

***Knowledge*** *is any kind of normative information that is considered to be relevant for solving a task at hand and that is not given as task input itself.*

In CA research, knowledge has been be modeled in a variety of forms that conform to this definition, ranging from lexicons, and engineered features to specially tailored pipelines, model components, or overall algorithm design (e.g., auxiliary tasks, or special training objectives). While this is not the primary dimension of our analysis (see §4.1), it is worth noting the difference between knowledge that is presented *explicitly*, namely, that can be rather directly used to shape the input representations for the task (e.g., lexicons, feature engineering, predictions of existing auxiliary models), and knowledge that is introduced *implicitly* through the algorithm or model design (e.g., auxiliary tasks in multi-task learning, or ordering of individual models in model pipelines). Both, we argue, conform to the above operational definition of knowledge to which we subscribe in this work. Finally, we emphasize that we consider the annotated corpora, leveraged in supervised task learning, to be input and not external knowledge brought to facilitate learning.

### 3.2 Analysis Scope

Generally, we focus on natural language argumentation and its computational treatment in NLP. Hence, we exclude work outside of this community, for example, studies on abstract argumentation (e.g., Vreeswijk, 1997), except if there is a strong link to natural language argumentation. For articles published in non-NLP venues, we made the decision based on the title. When unclear from the title whether the work primarily addresses natural language argumentation (e.g., as in the case of McBurney and Parsons, 2021), we analyzed the whole article before making the scope decision. Our survey covers the four subareas of CA in NLP from §2, with the following restrictions:

In *argument mining*, we do not include methods that have been designed strictly for a specific genre or domain and are not applicable elsewhere. *Argumentative zoning* (e.g., Teufel et al., 1999, 2009; Mo et al., 2020) and *citation analysis* (e.g., Athar, 2011; Lauscher et al., 2021), both specific to scientific publications, exemplify such methods. In contrast, we include methods that the general mining of argumentative structures, even if evaluated only in specific domains (e.g., Lauscher et al., 2018).

In *argument assessment*, we exclude work targeting sentiment analysis (e.g., Socher et al., 2013; Wachsmuth et al., 2014), as it is inherently more generic than other argumentation tasks and, accordingly, well-explored in general natural language understanding. Also, we exclude work on general-purpose natural language inference and common-sense reasoning (Bowman et al., 2015; Rajani et al., 2019; Ponti et al., 2020) in *argument reasoning*, and we do not cover the body of work on leveraging external structured knowledge for improved reasoning (e.g., Forbes et al., 2020; Lauscher et al., 2020a); we view these methods as more generic reasoning approaches that can, among others, also support argumentative reasoning (e.g., Habernal et al., 2018b), which we do cover in this survey. Finally, our overview of *argument generation* is limited strictly to argumentative text generation, as in argument summarization (e..g, Syed et al., 2020) and claim synthesis (e.g., Bilu and Slonim, 2016). The enormous body of work on (non-argumentative) natural language generation (Gatt and Krahmer, 2018) is out of our scope.

Note that some applications of CA are typically addressed through larger systems, which are composed of models tackling several of the tasks above. For instance, in *argument search*, a system might be composed of an argument extraction component (*mining*), a retrieval component that determines relevant arguments, as well as a quality rating component (*assessment*) to rank the mined arguments retrieved for given a topic (Wachsmuth et al., 2017b). In this work, we focus on core CA tasks and do not specifically discuss such composite systems. Within the described scope, we aim for comprehensiveness. However, given the immense body of work on natural language argumentation, we do not claim that this survey is complete.

### 3.3 Analysis and Annotation Process

We survey the state of the art in CA through the prism of the knowledge types leveraged in existing approaches. For each of the four CA subareas, we conducted our literature research in two steps: (1) in a *pre-study*, we collected *all* papers that we saw as relevant. To this end, we combined our expert knowledge of the field with extensive search in scientific search engines and proceedings of relevant conferences and workshops. On this basis, we established the knowledge pyramid. (2) In an *in-depth study*, we then selected the 10 most representative papers (according to scientometric indicators and our expert judgment) for each subarea and annotated them with the types of knowledge from the pyramid. We instructed three expert annotators to read each paper carefully. Based on our knowledge definition above and common forms of knowledge we identified in the pre-study, they were asked to decide *what types and what forms* of knowledge were involved, thus assigning all applicable types from the pyramid to each of the 40 sampled papers.

**Agreement.** We measured inter-annotator agreement (IAA) in a *top-level* and an *all-levels* variant across all sampled 40 papers (10 for each CA area) in terms of pair-wise averaged Cohen's $\kappa$ score. First, for each of the papers, we determined the most specific type of knowledge that it exploits (i.e., the one that is highest in the pyramid). Here, we observe a moderate IAA (Landis and Koch, 1977) with $\kappa = 0.54$. Second, across all categories, we observe a substantial IAA of $\kappa = 0.74$. All cases of disagreement were discussed thoroughly and resolved jointly.

The final distribution of knowledge types identified in papers for each CA subarea is shown in Figure 2b. As expected, almost all works (36 out of 40) leverage linguistic knowledge in some form. In contrast, world and topic knowledge (e.g., common-sense and factual knowledge, logic and rules) seem to be used least across the board. A reason for the latter may lie in the computational complexity of encoding such knowledge in a way that it can benefit concrete approaches to tasks—whereas this is often much

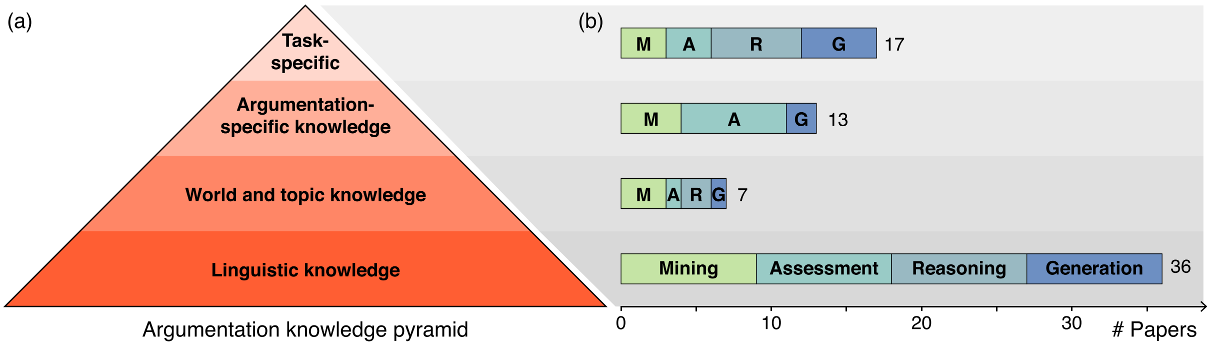


Figure 2: (a) Our proposed *argumentation knowledge pyramid*, encompassing four coarse-grained types of knowledge leveraged in CA research: The pyramid is organized according to increasing specificity of the knowledge types, from the bottom to the top. (b) Relative frequencies of the presence of knowledge types in the 40 representative papers (10 per CA subarea: mining, assessment, reasoning, and generation) selected for our in-depth study.

more straightforward for argumentation-specific knowledge (e.g., using lexicons) and task-specific knowledge (e.g., adopting a multitask learning setup). Moreover, topic knowledge is likely to make approaches more topic-dependent, namely, less broadly applicable, which is, more generally, often seen as an undesirable property for NLP approaches. We discuss the distribution in detail in the next section.

**Pre-Study.** Our aim was to collect as many relevant publications as we could for each of the four CA subareas. We first compiled a list of publications that we were personally aware of (i.e., leveraging ''expert knowledge''). Then, we augmented the list by firing queries with relevant keywords (again, compiled based on our expert knowledge) against the ACL Anthology[1] and Google Scholar.[2]

For example, we used the following queries for argument mining: *''argument[ation] mining''*, *''argument[ative] component''*, *''argument[ative] relation''*, and *''argument[ative] structure''*. For argument generation, we queried *''argument generation''*, *''argument synthesis''*, *''claim generation''*, *''claim synthesis''*, and *''argument summarization''*. In addition, we examined all publications from the proceedings of all seven editions (2014–2020) of the Argument Mining workshop series.

In each subarea, we included only publications that propose a computational approach to solving (at least) one CA task; in contrast, we did not consider publications describing shared tasks (Habernal et al., 2018b) or external knowledge resources for CA (Al Khatib et al., 2020a). With these rules in place, we ultimately collected a total of 162 CA papers, entirely listed in Table 1. By analyzing the types of knowledge used by approaches from collected publications, we induced the pyramid of knowledge types in Figure 2 with four coarse-grained knowledge types (§4.1), which was then the basis for our in-depth study (§4.2–§4.5).

**In-Depth Study.** In the second step, we used the knowledge pyramid as the basis for an in-depth analysis of a subset of 40 publications (10 per research area; bold in Table 1). Our selection of *prominent* papers for the in-depth study was guided by the following set of (sometimes mutually conflicting) criteria: (1) maximize the scientific impact of the publications in the sample, measured as a combination of the number of publication citations and our expert judgment of publication's overall impact on the CA field or subarea; (2) maximize the number of different methodological approaches in the sample;[3] and (3) maximize the representation of different researchers and research groups.

Once we had selected the 40 publications, three authors of this paper independently labeled all of

[1] https://aclanthology.org/.
[2] https://scholar.google.com/.

[3] Note that diversifying the sample with respect to methods is different than diversifying it according to knowledge types: two approaches may use the same type(s) of knowledge (e.g., linguistic) while adopting different methods (e.g., syntactic features vs. neural LMs). Our aim was to reduce the methodological redundancy of the sample.

| Task | Paper | Top Pyramid Level | Task | Paper | Top Pyramid Level |
|---|---|---|---|---|---|
| | | *Argument Mining* | | | |
| Comp. identification | Boltužić and Šnajder (2014) | World and topic | Multiple tasks | Stab and Gurevych (2014) | Arg.-specific |
| | **Ajjour et al. (2017)** | Linguistic | | Persing and Ng (2020) | Task-specific |
| | Spliethöver et al. (2019) | Linguistic | | Lawrence and Reed (2015) | Arg.-specific |
| | Petasis (2019) | Linguistic | | Sobhani et al. (2015) | Arg.-specific |
| | Trautmann et al. (2020) | Linguistic | | **Peldszus and Stede (2015)** | Task-specific |
| Comp. classification | Ong et al. (2014) | Linguistic | | Persing and Ng (2016a) | Arg.-specific |
| | Sobhani et al. (2015) | Arg.-specific | | **Eger et al. (2017)** | Linguistic |
| | Rinott et al. (2015) | Task-specific | | **Lawrence and Reed (2017a)** | Arg.-specific |
| | Al Khatib et al. (2016) | Linguistic | | Lawrence and Reed (2017b) | Arg.-specific |
| | Liebeck et al. (2016) | Linguistic | | Potash et al. (2017b) | Arg.-specific |
| | **Daxenberger et al. (2017)** | Linguistic | | Aker et al. (2017) | Arg.-specific |
| | **Levy et al. (2017)** | Arg.-specific | | **Niculae et al. (2017)** | Arg.-specific |
| | Shnarch et al. (2017) | Arg.-specific | | Stab and Gurevych (2017a) | Arg.-specific |
| | Habernal and Gurevych (2017) | Arg.-specific | | Saint-Dizier (2017) | Task-specific |
| | Dusmanu et al. (2017) | Arg.-specific | | Schulz et al. (2018) | Linguistic |
| | Lauscher et al. (2018) | Arg.-specific | | Shnarch et al. (2018) | Linguistic |
| | **Lugini and Litman (2018)** | Arg.-specific | | Eger et al. (2018) | Linguistic |
| | Stab et al. (2018b) | Arg.-specific | | Morio and Fujita (2018) | Arg.-specific |
| | Jo et al. (2019) | Linguistic | | Gemechu and Reed (2019) | Linguistic |
| | Mensonides et al. (2019) | Arg.-specific | | Lin et al. (2019) | Arg.-specific |
| | Reimers et al. (2019) | Arg.-specific | | Hewett et al. (2019) | Arg.-specific |
| | Hua et al. (2019b) | Arg.-specific | | Haddadan et al. (2019) | Arg.-specific |
| Relation identification | **Cabrio and Villata (2012)** | World and topic | | Eide (2019) | Arg.-specific |
| | Carstens and Toni (2015) | Arg.-specific | | Chakrabarty et al. (2019) | Arg.-specific |
| | Cocarascu and Toni (2017) | Linguistic | | Huber et al. (2019) | Arg.-specific |
| | Hou and Jochim (2017) | Task-specific | | Accuosto and Saggion (2019) | Task-specific |
| | **Galassi et al. (2018)** | Linguistic | | Morio et al. (2020) | Linguistic |
| | Paul et al. (2020) | World and topic | | Wang et al. (2020) | Arg.-specific |
| | | *Argument Assessment* | | | |
| Stance Detection | Ranade et al. (2013) | Arg.-specific | Quality assessment | **Habernal and Gurevych (2016b)** | Linguistic |
| | Hasan and Ng (2014) | Linguistic | | Ghosh et al. (2016) | Arg.-specific |
| | Sobhani et al. (2015) | Arg.-specific | | Wachsmuth et al. (2016) | Arg.-specific |
| | Persing and Ng (2016b) | Arg.-specific | | Wei et al. (2016) | Task-specific |
| | Toledo-Ronen et al. (2016) | Task-specific | | Tan et al. (2016) | Task-specific |
| | Sobhani et al. (2017) | Linguistic | | Chalaguine and Schulz (2017) | Linguistic |
| | **Bar-Haim et al. (2017a)** | Arg.-specific | | Stab and Gurevych (2017b) | Linguistic |
| | Boltužić and Šnajder (2017) | Task-specific | | Potash et al. (2017a) | Linguistic |
| | Bar-Haim et al. (2017b) | Task-specific | | **Wachsmuth et al. (2017c)** | Arg.-specific |
| | Rajendran et al. (2018a) | Linguistic | | Persing and Ng (2017) | Task-specific |
| | Sun et al. (2018) | Arg.-specific | | Lukin et al. (2017) | Task-specific |
| | Rajendran et al. (2018b) | Arg.-specific | | Wachsmuth et al. (2017a) | Task-specific |
| | Kotonya and Toni (2019) | Linguistic | | Simpson and Gurevych (2018) | Linguistic |
| | Durmus et al. (2019) | Linguistic | | Gu et al. (2018) | Linguistic |
| | Durmus and Cardie (2019) | Task-specific | | Passon et al. (2018) | Arg.-specific |
| | Toledo-Ronen et al. (2020) | Linguistic | | Ji et al. (2018) | Task-specific |
| | Kobbe et al. (2020a) | Arg.-specific | | **Durmus and Cardie (2018)** | Task-specific |
| | Sirrianni et al. (2020) | Arg.-specific | | El Baff et al. (2018) | Task-specific |
| | Somasundaran and Wiebe (2010) | Arg.-specific | | Dumani and Schenkel (2019) | Linguistic |
| | Porco and Goldwasser (2020) | Task-specific | | Potthast et al. (2019) | Linguistic |
| | Scialom et al. (2020) | Task-specific | | Gleize et al. (2019) | Linguistic |
| Frame identification | Ajjour et al. (2019) | Task-specific | | Toledo et al. (2019) | Linguistic |
| | **Trautmann (2020)** | Linguistic | | Potash et al. (2019) | Linguistic |
| Quality assessment | Liu et al. (2008) | Task-specific | | **Gretz et al. (2020b)** | Linguistic |
| | Persing et al. (2010) | Linguistic | | **El Baff et al. (2020)** | Linguistic |
| | Persing and Ng (2013) | Linguistic | | Wachsmuth and Werner (2020) | Linguistic |
| | Ong et al. (2014) | Linguistic | | Li et al. (2020) | Arg.-specific |
| | Persing and Ng (2014) | Linguistic | | **Al Khatib et al. (2020b)** | Task-specific |
| | Song et al. (2014) | Arg.-specific | | Lauscher et al. (2020b) | Task-specific |
| | **Persing and Ng (2015)** | Arg.-specific | | Skitalinskaya et al. (2021) | Linguistic |
| | Stab and Gurevych (2016) | Linguistic | Other tasks | **Kobbe et al. (2020b)** | Task-specific |
| | Habernal and Gurevych (2016a) | Linguistic | | Yang et al. (2019) | Linguistic |
| | | *Argument Reasoning* | | | |
| Warrant identification | **Boltužić and Šnajder (2016)** | Linguistic | Scheme classification | **Feng and Hirst (2011)** | Task-specific |
| | Sui et al. (2018) | Linguistic | | Song et al. (2014) | Task-specific |
| | Liebeck et al. (2018) | Linguistic | | **Lawrence and Reed (2015)** | Linguistic |
| | **Tian et al. (2018)** | Linguistic | | **Liga (2019)** | Linguistic |
| | Brassard et al. (2018) | Linguistic | | | |
| | Sui et al. (2018) | Linguistic | Fallacy Recognition | **Habernal et al. (2018c)** | Linguistic |
| | **Botschen et al. (2018)** | World and topic | | Habernal et al. (2018a) | Linguistic |
| | **Choi and Lee (2018)** | World and topic | | **Delobelle et al. (2019)** | Linguistic |
| | **Niven and Kao (2019)** | World and topic | Other tasks | Becker et al. (2021) | World and topic |
| | | *Argument Generation* | | | |
| Summarization | Egan et al. (2016) | Linguistic | Argument synthesis | **Zukerman et al. (2000)** | Task-specific |
| | **Wang and Ling (2016)** | Linguistic | | Carenini and Moore (2006) | Task-specific |
| | Syed et al. (2020) | Arg.-specific | | **Sato et al. (2015)** | Linguistic |
| | Alshomary et al. (2020a) | Arg.-specific | | Reisert et al. (2015) | Arg.-specific |
| | **Bar-Haim et al. (2020)** | Arg.-specific | | Hua and Wang (2018) | World and topic |
| Claim Synthesis | **Bilu and Slonim (2016)** | Task-specific | | Wachsmuth et al. (2018) | Arg.-specific |
| | Chen et al. (2018) | World and topic | | Le et al. (2018) | Arg.-specific |
| | Hidey and McKeown (2019) | Arg.-specific | | Hua et al. (2019a) | World and topic |
| | Alshomary et al. (2020b) | Arg.-specific | | **Hua and Wang (2019)** | World and topic |
| | **Gretz et al. (2020a)** | Arg.-specific | | **El Baff et al. (2019)** | Arg.-specific |
| | **Alshomary et al. (2021)** | Task-specific | | Bilu et al. (2019) | Task-specific |
| | | | | **Schiller et al. (2021)** | Task-specific |

Table 1: List of all publications surveyed in this study, across the four subareas of CA (argument mining, argument assessment, argument reasoning, and argument generation) sorted by task and year of publication, with the indication of the most specific level of knowledge used. Publications in bold are those selected for the in-depth study (§3).

them with the knowledge types from the pyramid. This allowed us to measure the inter-annotator agreement and to test the extent of shared understanding of the knowledge types captured by the pyramid and their usage in individual methodological approaches in CA. While we are aware that we cannot draw statistically significant conclusions based on a sample of such a limited size, we believe that our findings and this in-depth perspective will still be informative for the CA community.

## 4 Knowledge in Argumentation

As a result of our survey, we now introduce the *argumentation knowledge pyramid*, our proposed taxonomy encompassing four coarse-grained types of knowledge leveraged in CA. We then profile the large body of papers from the four CA subareas through the lens of the pyramid.

### 4.1 Argumentation Knowledge Pyramid

Based on the findings of our pre-study, we identify four coarse-grained types of knowledge being leveraged in CA research, which we organize in a taxonomy, as depicted in Figure 2. We chose to visualize our organization as a pyramid because it allows us to express a hierarchical generality-specificity relationship between the different types of knowledge.

**Linguistic Knowledge.** At the bottom of the pyramid is linguistic knowledge, leveraged by virtually all CA models and needed in practically all NLP tasks. In our pyramid, linguistic knowledge is a broad category that includes features derived from word *n*-grams, information about linguistic structure (e.g., part-of-speech tags, dependency parses), as well as features based on models of distributional semantics, such as (pre-trained) word embedding spaces (e.g., Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017) or representation spaces spanned by neural language models (LMs) (e.g., Clark et al., 2020; Devlin et al., 2019). We also consider leveraging distributional spaces (word embeddings or pre-trained LMs) built for specific (argumentative) tasks and domains as a form of linguistic knowledge, since such representation spaces are induced purely from textual corpora without any external supervision signal.

**World and Topic Knowledge.** Above the linguistic knowledge, we place the category of *world and topic knowledge*, in which we bundle all types of knowledge that are generally considered useful for various natural language understanding tasks, but that are not (or even cannot be) directly derived from textual corpora. This includes all types of common-sense knowledge, task-independent world knowledge (also known as factual knowledge), logical general-purpose axioms and rules, and similar. In most cases, such knowledge is collected from external structured or semi-structured resources (Sap et al., 2020; Lauscher et al., 2020a; Ji et al., 2021). Knowledge about a specific *debate topic* (e.g., legalization of marijuana) falls under this category, since topics encompass a set of real-world concepts (e.g., marijuana) and related facts (e.g., medical aspects of marijuana usage). Some systems explicitly require the debate topic as input, in order to gather topic knowledge from external sources.

**Argumentation-Specific Knowledge.** The third category in our knowledge pyramid encompasses knowledge about *what constitutes argumentation, arguments, and argumentative language*, including knowledge about subjective language (Stede and Schneider, 2018). This includes models of argumentation and argumentative structures (Toulmin, 2003; Bentahar et al., 2010a), models of cultural aspects and moral values (Haidt and Joseph, 2004; Graham et al., 2013), lexicons with terms indicating subjective, psychological, and moral categories (Hu and Liu, 2004; Tausczik and Pennebaker, 2010; Graham et al., 2009), predictions of subjectivity and sentiment classification models (Socher et al., 2013), and so forth. While sentiment, emotions, and affect are not argumentative *per se*, subjectivity is ingrained in argumentation and strongly influences argumentative manifestations (or lack thereof).

**Task-Specific Knowledge.** As the most specific type of knowledge, this category covers the types of knowledge that are relevant only for a specific CA task or a small set of tasks. For instance, leveraging discourse structure is considered beneficial for argumentative relation identification (Stab and Gurevych, 2014; Persing and Ng, 2016a; Opitz and Frank, 2019), a common argument mining task.

Table 2 illustrates the four types of knowledge from the pyramid by means of concrete examples.

| Knowledge | Source | CA Subarea (Task) | Introduced | Explanation |
|---|---|---|---|---|
| Linguistic | Habernal et al. (2018c) | Argument reasoning (fallacy recognition) | Explicitly | Semantic associations between lexical units in the word embedding space enable generalization across different *lexicalizations* of ad hominem arguments (e.g., *''pretentions [explanation]''* vs. *''narcissistic [idiot]''*) and wordings that point to fallacious reasoning (e.g., *''[if only you wouldn't rely on] fallacious arguments''* vs. *''[another] unsubstantiated statement)''*. |
| World and topic | Hua and Wang (2019) | Argument generation (argument synthesis) | Implicitly | The structure of the argument – sequence of *Premise*, *Claim*, and *Functional* utterances – is conditioned by the *topic* of debate. For example, Reddit arguments in political topics (e.g., *''US cutting off foreign aid''* tend to start with a *Claim* (*''It can be a useful political bargaining chip''*), continue with supporting *Premises* (e.g., *''US cut financial aid to Uganda due to its plans to make homosexuality a crime''*) and finish with *Functional* utterances (e.g., *''Please change your mind!''*). |
| Arg.-specific | Wachsmuth et al. (2017c) | Argument assessment (quality assessment) | Explicitly | *Argument relevance* is determined in an ''objective'' way. Argument ''reuse'', where one argument leverages the conclusion of another argument is the base for the induction of a large-scale (directed) argument graph. Running a PageRank algorithm on that graphs yields relevance scores for all arguments. Such objective and content-agnostic argument relevance score can be *useful for a wide variety of CA tasks*; knowledge about argument reuse thus represents argumentation-specific knowledge. |
| Task-specific | Peldszus and Stede (2015) | Argument mining (multiple tasks) | Implicitly | Argumentative structure of the text assumed to be a *tree*: There is one central claim for the text which is the root of the tree, other argumentative components are the nodes of the tree, and edges reflect the *support* or *attack* relations between argumentative discourse components. |

Table 2: Concrete examples for the four types of knowledge distinguished in the knowledge pyramid (see Figure 2). We additionally indicate for each example whether the knowledge is introduced in an explicit or implicit manner (column ''Introduced''; see §3.1).

### 4.2 Knowledge in Argument Mining

**Pre-Study.** From the 162 papers we surveyed, 56 belong to the subarea of argument mining, which is the second-largest subarea after argument assessment. The publications that we analyzed were published in the period from 2012 to 2020. Of these 56 publications, 17 relied purely on linguistic knowledge, three exploited world and topic knowledge as the most specific knowledge type, 30 leveraged argumentation-specific knowledge, and six task-specific knowledge. We next describe the detailed findings of our in-depth analysis.

**In-Depth Study.** Table 3 shows the results of our assignment of all applicable knowledge types to 10 sampled argument mining papers, published between 2012 and 2018. All but one rely on linguistic knowledge: Earlier approaches leveraged traditional linguistic features, such as $n$-grams and syntactic features (e.g., Peldszus and Stede, 2015; Lugini and Litman, 2018), whereas later work resorted to word embeddings as the dominant representation (e.g., Eger et al., 2017; Niculae et al., 2017; Daxenberger et al., 2017; Galassi et al., 2018).

A few papers exploit other types of knowledge. Cabrio and Villata (2012), for example, leverage a pretrained NLI model to analyze online debate interactions.[4] While they resort to the abstract argumentation framework of Dung (1995), they do so only for the purposes of the evaluation, which is why we do not judge their approach as reliant on argumentation-specific knowledge. Lawrence and Reed (2017b) use, in addition to word embeddings, world and topic knowledge from WordNet and argumentation-specific knowledge in the form of structural assumptions for mining large-scale debates. Ajjour et al. (2017) combine linguistic knowledge in the form of GloVe embeddings (Pennington et al., 2014) and other linguistic features with an argumentation-specific lexicon of discourse markers. Task-specific mining knowledge is mostly leveraged

[4]Note that our judgments reflect only the types of knowledge that the approach presented in the paper directly exploits: this is why, for example, we judge the reliance of the approach of Cabrio and Villata (2012) on a pretrained NLI model as exploitation of world and topic knowledge only, even though the NLI model itself (Kouylekov and Negri, 2010) had been trained using a range of linguistic features.

| Approach | Linguistic | World and Topic | Arg. specific | Task. Specific |
|---|---|---|---|---|
| *Argument Mining* | | | | |
| Cabrio and Villata (2012) | ✗ | ✓ | ✗ | ✗ |
| Peldszus and Stede (2015) | ✓ | ✗ | ✓ | ✓ |
| Daxenberger et al. (2017) | ✓ | ✗ | ✗ | ✗ |
| Eger et al. (2017) | ✓ | ✗ | ✗ | ✓ |
| Niculae et al. (2017) | ✓ | ✗ | ✗ | ✗ |
| Lawrence and Reed (2017b) | ✓ | ✓ | ✓ | ✗ |
| Levy et al. (2017) | ✓ | ✓ | ✓ | ✗ |
| Ajjour et al. (2017) | ✓ | ✗ | ✓ | ✗ |
| Galassi et al. (2018) | ✓ | ✗ | ✗ | ✗ |
| Lugini and Litman (2018) | ✓ | ✗ | ✗ | ✓ |
| *Argument Assessment* | | | | |
| Persing and Ng (2015) | ✓ | ✗ | ✓ | ✗ |
| Habernal and Gurevych (2016b) | ✓ | ✗ | ✓ | ✗ |
| Wachsmuth et al. (2017c) | ✗ | ✗ | ✓ | ✗ |
| Bar-Haim et al. (2017a) | ✓ | ✓ | ✓ | ✗ |
| Durmus and Cardie (2018) | ✓ | ✗ | ✓ | ✓ |
| Trautmann (2020) | ✓ | ✗ | ✗ | ✗ |
| Kobbe et al. (2020b) | ✓ | ✗ | ✓ | ✗ |
| El Baff et al. (2020) | ✓ | ✗ | ✓ | ✓ |
| Al Khatib et al. (2020b) | ✓ | ✗ | ✗ | ✓ |
| Gretz et al. (2020b) | ✓ | ✗ | ✗ | ✗ |
| *Argument Reasoning* | | | | |
| Feng and Hirst (2011) | ✗ | ✗ | ✗ | ✓ |
| Lawrence and Reed (2015) | ✓ | ✗ | ✗ | ✓ |
| Boltužić and Šnajder (2016) | ✓ | ✗ | ✗ | ✗ |
| Habernal et al. (2018c) | ✓ | ✗ | ✗ | ✗ |
| Choi and Lee (2018) | ✓ | ✓ | ✗ | ✗ |
| Tian et al. (2018) | ✓ | ✗ | ✗ | ✗ |
| Botschen et al. (2018) | ✓ | ✓ | ✗ | ✗ |
| Delobelle et al. (2019) | ✓ | ✗ | ✗ | ✗ |
| Niven and Kao (2019) | ✓ | ✗ | ✗ | ✗ |
| Liga (2019) | ✓ | ✗ | ✗ | ✗ |
| *Argument Generation* | | | | |
| Zukerman et al. (2000) | ✗ | ✗ | ✓ | ✓ |
| Sato et al. (2015) | ✓ | ✗ | ✓ | ✗ |
| Bilu and Slonim (2016) | ✓ | ✗ | ✓ | ✓ |
| Wang and Ling (2016) | ✓ | ✗ | ✗ | ✗ |
| El Baff et al. (2019) | ✓ | ✗ | ✗ | ✓ |
| Hua et al. (2019b) | ✓ | ✓ | ✗ | ✗ |
| Bar-Haim et al. (2020) | ✓ | ✗ | ✓ | ✗ |
| Gretz et al. (2020a) | ✓ | ✗ | ✗ | ✗ |
| Alshomary et al. (2021) | ✓ | ✗ | ✗ | ✓ |
| Schiller et al. (2021) | ✓ | ✗ | ✓ | ✓ |

Table 3: The types of knowledge involved in the approaches of all publications that we included in the second stage of our literature survey (in-depth study) ordered by the high-level task they tackle and the year.

in multi-task learning scenarios (Lugini and Litman, 2018) or when aiming to extract arguments of more complex structures, that is, with multiple components and/or chains of claims (Eger et al., 2017; Peldszus and Stede, 2015). For instance, Peldszus and Stede (2015) jointly predict different aspects of the argument structure and then apply minimum spanning tree decoding, exploiting that mining of argument structure bears similarities with discourse parsing. The only template-based approach we cover is that of Levy et al. (2017), who construct queries using templates and use ground sentences in Wikipedia concepts (i.e., world and topic knowledge) for unsupervised claim detection. Their approach also leverages an argumentation-specific lexicon of claim-related words (i.e., arg.-specific knowledge), next to the linguistic and world/topic knowledge.

### 4.3 Knowledge in Argument Assessment

**Pre-Study.** The largest portion of the 162 publications, 64 in total, belong to the area of argument assessment, spanning the time period from 2008 to 2021. Of those publications, 29 leverage only linguistic knowledge, but almost 20 rely on task-specific knowledge as the most specific knowledge type. Interestingly, none of the surveyed papers use world and topic knowledge as the most specific knowledge type. That is, if they rely on world and topic knowledge, they also leverage argumentation-specific and/or task-specific knowledge.

**In-Depth Study.** The 10 assessment papers analyzed in-depth (period 2015–2020) reveal that, much like in argument mining, most of the work models linguistic knowledge (e.g., Trautmann, 2020; Kobbe et al., 2020b). For example, Gretz et al. (2020b) assess argument quality based on a representation that combines bag-of-words (i.e., sparse symbolic text representation) with latent embeddings, both derived from static GloVe word embeddings (Pennington et al., 2014) and produced by a pretrained BERT model (Devlin et al., 2019). Most of the papers at the linguistic knowledge level of the pyramid, however, predominantly rely on sparse symbolic (i.e., word-based) linguistic features (e.g., Persing and Ng, 2015; Bar-Haim et al., 2017b; Durmus and Cardie, 2018; Al Khatib et al., 2020b; El-Baff et al., 2020).

Only one of the 10 selected publications resorts to world and topic knowledge: Bar-Haim et al. (2017a) map the content of claims to Wikipedia concepts for stance classification. A common technique in argument assessment is to include argumentation-specific knowledge about sentiment or subjectivity: this is motivated by the intuition that these features directly affect argumentation quality and correlate with stances. For instance, Wachsmuth et al. (2017a) note that *emotional appeal*, which is clearly correlated with the sentiment of the text, may affect the rhetorical effectiveness of arguments. Technically, the information on subjectivity is introduced either by means of subjective lexica (e.g., Bar-Haim et al., 2017a; Durmus and Cardie, 2018; El-Baff et al., 2020) or via predictions of pretrained sentiment classifiers (Habernal and Gurevych, 2016b). In a different example of the use of argumentation-specific knowledge, Wachsmuth et al. (2017c) exploit reuses between arguments (e.g., a premise of one argument uses the claim of another) to quantify argument relevance by means of graph-based propagation with PageRank.

A notable task-specific knowledge category is the use of *user information* for argument quality assessment. According to theory (Wachsmuth et al., 2017a), argument quality does not only depend on the text utterance itself but also on the speaker and the audience, for example, on their prior beliefs and their cultural context. To model this, Durmus and Cardie (2018) include information about users' prior beliefs as predictors of arguments' persuasiveness, Al Khatib et al. (2020b) predict persuasiveness using user-specific feature vectors, and El Baff et al. (2020) train audience-specific classifiers.

### 4.4 Knowledge in Argument Reasoning

**Pre-Study.** According to our pre-study, argument reasoning is the smallest subarea of CA, with only 17 (out of 162) papers published (in the period between 2011 and 2021). The tasks in this subarea include argumentation scheme classification (Feng and Hirst, 2011; Lawrence and Reed, 2015), warrant identification and exploitation (Habernal et al., 2018b; Boltužić and Šnajder, 2016), and fallacy recognition (Habernal et al., 2018c; Delobelle et al., 2019). Linguistic knowledge denotes the most commonly used type of knowledge in reasoning as well (11 out of 17 papers rely on some type of linguistic knowledge), and four papers in this subarea exploit world and topic knowledge.

**In-Depth Study.** In our subset from argument reasoning, general-domain embeddings are by far the most frequently employed type of knowledge injection approach (Boltužić and Šnajder, 2016; Habernal et al., 2018c; Choi and Lee, 2018; Tian et al., 2018; Botschen et al., 2018; Delobelle et al., 2019; Niven and Kao, 2019). In contrast, Lawrence and Reed (2015) use traditional linguistic features, and Liga (2019) models syntactic features with tree kernels to recognize specific reasoning structures in arguments. Task-specific knowledge is modeled by Feng and Hirst (2011), who design specific features for classifying argumentation schemes, and Lawrence and Reed (2015) utilize features specific to individual types of premises and conclusions. Choi and Lee (2018) use a pretrained natural language inference model to select the correct warrant in warrant identification.[5] For the same task, Botschen et al. (2018) leverage event knowledge about common situations (from FrameNet) and factual knowledge about entities (from Wikidata).

### 4.5 Knowledge in Argument Generation

**Pre-Study.** Finally, we surveyed 23 generation papers, ranging from 2000 to 2021. Argumentation-specific knowledge is the most specific knowledge type in most (10) publications. Six publications have task-specific knowledge as the most specific knowledge type, and four do not employ anything more specific than world and topic knowledge. Unlike in other subareas, only a few publications (3) in argument generation rely purely on linguistic knowledge. Common argument generation tasks include argument summarization (Egan et al., 2016; Bar-Haim et al., 2020), claim synthesis (Bilu et al., 2019; Alshomary et al., 2021), and argument synthesis (Zukerman et al., 2000; Sato et al., 2015).

**In-Depth Study.** As in the case of argument reasoning, many generation approaches employ linguistic knowledge in the form of general-purpose embeddings (Wang and Ling, 2016; Hua et al., 2019a; Bar-Haim et al., 2020; Gretz et al., 2020a; Schiller et al., 2021). Only Sato et al. (2015) report using traditional (i.e., sparse, symbolic) linguistic features; Bilu and Slonim (2016) used traditional linguistic features for predicting the suitability of candidate claims.

World and topic knowledge is utilized by Hua et al. (2019a), who retrieve Wikipedia passages as claim candidates. As argumentation-specific knowledge, Bar-Haim et al. (2020) use an external quality classifier. In a similar vein, Schiller et al. (2021) incorporate the output from argument and stance classifiers from the ArgumenText API (Stab et al., 2018a) and condition the generation model on control codes encoding topic, stance, and aspect of the argument. Alshomary et al. (2021) condition their model on a audience beliefs by deriving bag-of-words representations

[5] As in the case of Cabrio and Villata (2012) in argument mining, we consider a pretrained NLI model to represent *world and topic knowledge*.

| Type | Common Modeling Techniques |
|---|---|
| Task-specific | Structure (e.g., multitask learning), user information (e.g., features), . . . |
| Argumentation-specific | Sentiment (e.g., lexicon, external classifier), argumentation (e.g., fine-tuning), . . . |
| World and topic | Inference knowledge (e.g., infusion), world knowledge (e.g., linking to Wikipedia), . . . |
| Linguistic | $n$-grams (e.g., traditional features), general semantics (e.g., GloVe embeddings), . . . |

Table 4: Common techniques used for modeling the four types of knowledge from the proposed knowledge pyramid.

from the authors' texts and then fine-tuning a pre-trained language model. Sato et al. (2015) model (argumentation-specific) knowledge about values. Predicate and sentiment lexica are employed by Bilu and Slonim (2016), whereas El Baff et al. (2019) learn likely sequences of argumentative units from features computed from argumentation-specific knowledge. They additionally include task-specific knowledge by using a knowledge base with components of claims. A pioneering work that stands out is the approach of Zukerman et al. (2000), which uses argumentation-specific knowledge about micro-structure in combination with task-specific discourse templates.

## 5 Emerging Trends and Discussion

We now summarize the emerging trends and open challenges in the four CA areas, abstracted from our analyses of the use of knowledge types.

**General Observations.** Most of the 162 publications that we reviewed aim to capture some type of ''advanced'' knowledge, that is, knowledge beyond what can be inferred from the text data alone: 60 publications rely purely on linguistic knowledge, whereas the remaining 102 model at least one of the other three higher knowledge types. This empirically confirms the intuition that success in CA crucially depends on complex knowledge that is external to the text. Also, unsurprisingly, argumentation-specific knowledge is overall the most common type of external knowledge used in CA approaches: Argumentation-specific knowledge can, in principle, facilitate any computational argumentation task. In comparison, world and common-sense knowledge are fairly underrepresented: Only seven of the 40 publications in our in-depth study rely on some variant of it. This is surprising, given that the approaches that leverage such knowledge consistently report substantial performance gains.

**Comparison across Types of Knowledge.** We observe differences in the form in which the different knowledge types (e.g., linguistic vs. argument-specific knowledge) are commonly provided and incorporated in methodological approaches. We provide examples in Table 4.

**Comparison across Areas.** We also note substantial differences across the four high-level CA subareas. The predominant most specific knowledge types vary across the areas: in argument mining and assessment, linguistic and argumentation-specific knowledge are most commonly employed, whereas in argument reasoning approaches, world and topic knowledge (e.g., knowledge about reasoning mechanisms) represents the most common top-level category from the pyramid. In argument generation, argumentation-specific and task-specific knowledge were the most common top-level categories. We believe that this variance is due to the nature of the tasks in each area: Predicting argumentative structures in argument mining is strongly driven by lexical cues (linguistic knowledge) and structural aspects (argumentation-specific knowledge). Despite being studied most extensively, argument mining rarely exploits world and topic knowledge (e.g., from knowledge bases or lexico-semantic resources): There is possibly room for progress in argument mining from more extensive exploitation of structured knowledge sources.

As previously suggested by Wachsmuth et al. (2017a), we find that argument assessment relies on a combination of linguistic features and higher-level argumentation-related properties that are assessed independently, such as sentiment. Argument reasoning, in contrast, strongly relies on basic inference rules and general world knowledge. Finally, the knowledge used in argument generation seems to be highly task- and domain-dependent.

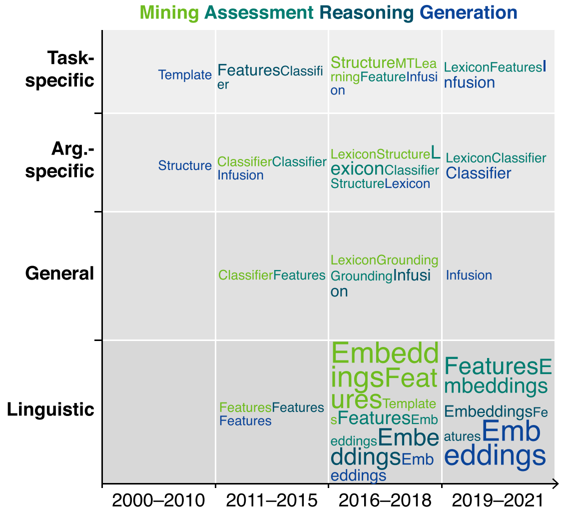


Figure 3: Techniques of employing knowledge in CA organized by defined time periods ($x$-axis), knowledge category ($y$-axis), and area (color). The size of the term indicates the number of occurrences of the techniques (between 1 and 7) in our sample of 40 papers.

Not only the types of knowledge but also the techniques employed for injecting that knowledge into CA models substantially differ across the subareas. Considering linguistic knowledge, for example, argument assessment approaches predominantly use lexical cues and traditional symbolic text representations, whereas the body of work on argument reasoning primarily relies on latent semantic representations (i.e., embeddings). Most variation in terms of knowledge modeling techniques is found in the argument generation area. Here, the techniques range from template- and structure-based approaches to external lexica and classifiers to embeddings and infusion.

**Diachronic Analysis.** Figure 3 depicts the temporal development of knowledge modeling techniques in CA, with year, CA subarea, and knowledge type as dimensions. We analyze four time periods, corresponding to pioneering work (2000–2010), the rise of CA in NLP (2011–2015), the shift to distributional methods (2016–2018), and the most recent trends (2019–2021).

This diachronic analysis reveals that CA is roughly aligned with trends observed in other NLP areas: in the pre-neural era before 2016, knowledge has traditionally been modeled via features, sometimes using knowledge from external resources and outputs or previously trained classifiers (i.e., the pipelined approaches). Later, more advanced techniques such as grounding, infusion, and above all embeddings became more popular. However, we note that distinct techniques are used for the different knowledge types; embeddings, in particular, have been used exclusively to encode linguistic knowledge. Although representation learning can be applied to other argumentative resources, CA efforts in this direction have been few and far between (e.g., Toledo-Ronen et al., 2016; Al Khatib et al., 2020a). This warrants more CA work on embedding structured knowledge and towards a unified argumentative representation space that would support the whole spectrum of CA tasks.

## 6 Where Should We Go from Here?

Mastering argumentative discourse requires various types of advanced knowledge (Moens, 2018), making CA one of the most complex problems in AI (Atkinson et al., 2017). This raises the question of a suitable path to reaching argumentative proficiency for computational models. In this survey, we identified empirical evidence that integrating advanced knowledge can lead to performance improvements on a range of CA tasks. In the following, we pick out those that we see as key ideas toward the goal of mastering argumentation computationally.

Argument mining is often seen as a structure-oriented task. Lawrence and Reed (2017a) brought up the notion that topic knowledge may actually predict relations between argument components. Eger et al. (2017), on the other hand, formulated mining of argument structure as an end-to-end task. Integrating these two views and combining respective methods could hold much promise.

Despite an abundance of work on encoding and leveraging common sense knowledge (e.g., Lauscher et al., 2020a; Lin et al., 2021), argument assessment methods fail to decompose arguments into concepts, with the work of Bar-Haim et al. (2017a) on stance classification as the positive exception. Despite some evidence of difficulty of integration of common-sense knowledge in argument reasoning tasks (Botschen et al., 2018), there is no alternative to accurately representing/encoding common-sense knowledge, if we are to build reliable CA systems. Beyond that, Kobbe et al. (2020b) looked at the impact of morals on argument quality. Such research on modeling fine-grained and socially and culturally-dependent knowledge, such as values and social norms—across languages, is still in its infancy in NLP in general. Systematic research on building respective knowledge sources and benchmarks could push CA to the next level.

As emphasized by existing work (e.g., Stede and Schneider, 2018), argumentation is inherently social and thus highly dependent on the relationship between the speaker and her audience. A more straightforward integration of knowledge about the speaker could prove beneficial: The work of Alshomary et al. (2021), encoding speaker's belief in argument generation, is a step in this direction.

In sum, what we believe is missing in existing work and what could drive the future of CA is a *unified knowledge representation space* that would aggregate and consolidate all CA-relevant knowledge, and be universally beneficial across CA tasks. As shown in this survey, CA-relevant knowledge is fragmented across heterogeneous sources (e.g., corpora, knowledge bases, lexicons) and coupled only sporadically and in an ad-hoc (not principled) manner. Considering the modest sizes of existing CA resources, a methodological orientation to modular and sample-efficient learning and adaptation (Houlsby et al., 2019; Gururangan et al., 2020; Ponti et al., 2022) could provide means to this end.

## 7 Conclusion

Motivated by the theoretical importance of knowledge in argumentation and by previous work pointing to the need for more research on incorporating advanced types of knowledge in computational argumentation, we have studied the role of knowledge in the body of research works in the field. In total, we surveyed 162 publications spanning the subareas of argument mining, assessment, reasoning, and generation. To organize the approaches described in these works, we proposed a pyramid-like knowledge taxonomy systematizing the types of knowledge according to their specificity, from basic linguistic to task-specific knowledge.

Our survey yields important findings. Many approaches employing advanced knowledge types (e.g., world and argumentation-specific knowledge) report empirical gains. Still, reliance on such external knowledge types is far from uniform across CA areas: While exploitation of such knowledge is pervasive in argument reasoning and generation, it is far less present in argument mining. We hope that our findings lead to more systematic consideration of different knowledge sources for CA tasks.

## References


Pablo Accuosto and Horacio Saggion. 2019. Transferring knowledge from discourse to arguments: A case study with scientific abstracts. In *Proceedings of the 6th Workshop on Argument Mining*, pages 41–51, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4505

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1290

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5115

Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, and Mina Ghobadi. 2017. What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 91–96, Copenhagen, Denmark, Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5112

Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument mining for scholarly document processing: Taking stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.sdp-1.7

Khalid Al Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020a. End-to-end argumentation knowledge graph construction. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7367–7374. AAAI. https://doi.org/10.1609/aaai.v34i05.6231

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020b. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.632

Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1165

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.17

Milad Alshomary, Nick Düsterhus, and Henning Wachsmuth. 2020a. Extractive snippet generation for arguments. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1969–1972, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3397271.3401186

Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2020b. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.399

Aristotle. ca. 350 B.C.E./ translated 2007. *On Rhetoric: A Theory of Civic Discourse*, Oxford University Press. Oxford, UK. Translated by George A. Kennedy.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, pages 81–87, Stroudsburg, PA, USA. Association for Computational Linguistics.

Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. 2017. Towards artificial argumentation. *AI Magazine*, 38(3):25–36. https://doi.org/10.1609/aimag.v38i3.2704

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017a. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-1024

Roy Bar-Haim, Lilach Edelstein, Charles Jochim, and Noam Slonim. 2017b. Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, Copenhagen, Denmark, Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5104

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

*Processing (EMNLP)*, pages 39–49, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.3

Maria Becker, Siting Liang, and Anette Frank. 2021. Reconstructing implicit knowledge with language models. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24, Online, Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.deelio-1.2

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010a. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259. https://doi.org/10.1007/s10462-010-9154-1

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010b. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259. https://doi.org/10.1007/s10462-010-9154-1

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 2019. Argument invention from first principles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1013–1026, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1097

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2085

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. https://doi.org/10.1162/tacl_a_00051

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland, Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-2107

Filip Boltužić and Jan Šnajder. 2016. Fill the gap! Analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2815

Filip Boltužić and Jan Šnajder. 2017. Toward stance classification based on claim microstructures. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 74–80, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5210

Teresa Botschen, Daniil Sorokin, and Iryna Gurevych. 2018. Frame- and entity-based knowledge for common-sense argumentative reasoning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 90–96, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5211

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. https://doi.org/10.18653/v1/D15-1075

Ana Brassard, Tin Kuculo, Filip Boltužić, and Jan Šnajder. 2018. TakeLab at SemEval-2018 task12: Argument reasoning comprehension with skip-thought vectors. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1133–1136, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1192

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433. https://doi.org/10.24963/ijcai.2018/766

Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952. https://doi.org/10.1016/j.artint.2006.05.003

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO, Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0504

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1291

Lisa Andreevna Chalaguine and Claudia Schulz. 2017. Assessing convincingness of arguments in online debates with limited number of features. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 75–83, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-4008

Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88, Tilburg University, The Netherlands. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6509

HongSeok Choi and Hyunju Lee. 2018. GIST at SemEval-2018 task 12: A network transferring inference knowledge to argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 773–777, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1122

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1144

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220. https://doi.org/10.1007/978-3-031-02151-0

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? Cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1218

Pieter Delobelle, Murilo Cunha, Eric Massip Cano, Jeroen Peperkamp, and Bettina Berendt. 2019. Computational ad hominem detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 203–209, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-2028

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fred I. Dretske. 1981. *Knowledge and the Flow of Information*. MIT Press.

Lorik Dumani and Ralf Schenkel. 2019. A systematic comparison of methods for finding good premises for claims. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 957–960, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3331184.3331282

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357. https://doi.org/10.1016/0004-3702(94)00041-X

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1094

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1057

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1456

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1245

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2816

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1002

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stian Rødven Eide. 2019. The Swedish PoliGraph: A semantic graph for argument mining of Swedish parliamentary data. In *Proceedings of the 6th Workshop on Argument Mining*, pages 52–57, Florence, Italy. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein.

2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-8607

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-1044

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.287

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

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.48

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Springer.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2018. Argumentative link prediction using residual networks and multi-objective learning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5201

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170. https://doi.org/10.1613/jair.5477

Debela Gemechu and Chris Reed. 2019. Decompositional argument mining: A general purpose approach for argument graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 516–526, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1049

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2089

G. Nigel Gilbert. 1977. Referencing as persuasion. *Social Studies of Science*, 7(1):113–122. https://doi.org/10.1177/030631277700700112

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? Choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1093

Alvin I. Goldman. 1967. A causal theory of knowing. *The Journal of Philosophy*, 64(12):357–372. https://doi.org/10.2307/2024268

Niels Gottschalk-Mazouz. 2013. Internet and the flow of knowledge: Which ethical and political challenges will we face? *From ontos verlag: Publications of the Austrian Ludwig Wittgenstein Society-New Series (Volumes 1–18)*, 7.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029. https://doi.org/10.1037/a0015141

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020a. The workweek is the best time to start a family – a study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 528–544, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.47

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020b. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813. https://doi.org/10.1609/aaai.v34i05.6285

Yunfan Gu, Zhongyu Wei, Maoran Xu, Hao Fu, Yang Liu, and Xuanjing Huang. 2018. Incorporating topic aspects for online comment convincingness evaluation. In *Proceedings of the 5th Workshop on Argument Mining*, pages 97–104, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*. Forlì-Cesena, Italy.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1129

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1150

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179. https://doi.org/10.1162/COLI_a_00276

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018a. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1175

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018c. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1036

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! Mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1463

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66. https://doi.org/10.1162/0011526042365555

Charles L. Hamblin. 1970. *Fallacies*. Methuen, London, UK.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

John Hawthorne. 2002. Deeply contingent a priori knowledge. *Philosophy and Phenomenological Research*, 65(2):247–269. https://doi.org/10.1111/j.1933-1592.2002.tb00201.x

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4512

Christopher Hidey and Kathy McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1756–1767, Minneapolis, Minnesota. Association for Computational Linguistics.

Yufang Hou and Charles Jochim. 2017. Argument relation classification using a joint inference model. In *Proceedings of the 4th Workshop on Argument Mining*, pages 60–66, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5107

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. *AAAI*, 4(4):755–760.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019a. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1255

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019b. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

https://doi.org/10.18653/v1/P18-1021

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 591–602, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1055

Laurine Huber, Yannick Toussaint, Charlotte Roze, Mathilde Dargnat, and Chloé Braud. 2019. Aligning discourse and argumentation structures using subtrees and redescription mining. In *Proceedings of the 6th Workshop on Argument Mining*, pages 35–40, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4504

Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuanjing Huang. 2018. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3703–3714, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S. Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. 2019. A cascade model for proposition extraction in argumentation. In *Proceedings of the 6th Workshop on Argument Mining*, pages 11–24, Florence, Italy. Association for Computational Linguistics.

Jonathan Kobbe, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020a. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.4

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020b. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2019. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *Proceedings of the 6th Workshop on Argument Mining*, pages 156–166, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4518

Milen Kouylekov and Matteo Negri. 2010. An open-source package for recognizing textual entailment. In *Proceedings of the ACL 2010 System Demonstrations*, pages 42–47.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174. https://doi.org/10.2307/2529310, PubMed: 843571

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2018. Investigating the role of argumentation in the rhetorical analysis of scientific publications with neural multi-task learning models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3326–3338, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1370

Anne Lauscher, Brandon Ko, Bailey Kuhl, Sophie Johnson, David Jurgens, Arman Cohan, and Kyle Lo. 2021. Multicite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. *arXiv preprint arXiv:2107.00414*. https://doi.org/10.18653/v1/2022.naacl-main.137

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020a. Common sense or world knowledge? Investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First*

*Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.deelio-1.5

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020b. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.402

John Lawrence and Chris Reed. 2015. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO. Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0516

John Lawrence and Chris Reed. 2017a. Mining argumentative structure from natural language text using automatically generated premise-conclusion topic models. In *Proceedings of the 4th Workshop on Argument Mining*, pages 39–48, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5105

John Lawrence and Chris Reed. 2017b. Using complex argumentative interactions to reconstruct the argumentative structure of large-scale debates. In *Proceedings of the 4th Workshop on Argument Mining*, pages 108–117, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5114

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818. https://doi.org/10.1162/coli_a_00364

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: A retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130, Brussels, Belgium. Association for Computational Linguistics.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5110

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? Mining arguments in the German online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2817

Matthias Liebeck, Andreas Funke, and Stefan Conrad. 2018. HHU at SemEval-2018 task 12: Analyzing an ensemble-based deep learning approach for the argument mining task of choosing the correct warrant. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1114–1119, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1188

Davide Liga. 2019. Argumentative evidences classification and argument scheme detection using tree kernels. In *Proceedings of the 6th Workshop on Argument Mining*, pages 92–97, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4511

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on*

*Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287.

Jian-Fu Lin, Kuo Yu Huang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Lexicon guided attentive neural network model for argument mining. In *Proceedings of the 6th Workshop on Argument Mining*, pages 67–73, Florence, Italy. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2015. Argument mining: A machine learning perspective. In *International Workshop on Theory and Applications of Formal Argumentation*, pages 163–176. Springer. https://doi.org/10.1007/978-3-319-28460-6_10

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167. https://doi.org/10.1007/978-3-031-02145-9

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *2008 Eighth IEEE International Conference on Data Mining*, pages 443–452. https://doi.org/10.1109/ICDM.2008.94

Keith Lloyd. 2007. Rethinking rhetoric from an indian perspective: Implications in the ‘‘nyaya sutra’’. *Rhetoric Review*, 26(4):365–384. https://doi.org/10.1080/07350190701577892

Luca Lugini and Diane Litman. 2018. Argument component classification for classroom discussions. In *Proceedings of the 5th Workshop on Argument Mining*, pages 57–67, Brussels, Belgium, Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5208

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-1070

Peter McBurney and Simon Parsons. 2021. Argument schemes and dialogue protocols: Doug Walton's legacy in artificial intelligence. *IfCoLoG Journal of Logics and their Applications*, 8(1):263–290.

Jean-Christophe Mensonides, Sébastien Harispe, Jacky Montmain, and Véronique Thireau. 2019. Automatic detection and classification of argument components using multi-task deep neural network. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 25–33, Trento, Italy. Association for Computational Linguistics.

Tomas Mikolov, 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*, pages 3111–3119.

Wang Mo, Cui Yunpeng, Chen Li, and Li Huan. 2020. A deep learning-based method of argumentative zoning for research articles. *Data Analysis and Knowledge Discovery*, 4(6):60–68.

Marie-Francine Moens. 2018. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument & Computation*, 9(1):1–14. https://doi.org/10.3233/AAC-170025

Gaku Morio and Katsuhide Fujita. 2018. End-to-end argument mining for discussion threads based on parallel constrained pointer architecture. In *Proceedings of the 5th Workshop on Argument Mining*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5202

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. 2020. Towards better non-tree argument mining: Proposition-level biaffine parsing with task-specific parameterization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3259–3266, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.298

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1091

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1459

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, Baltimore, Maryland. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4503

Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. 2018. Predicting the usefulness of Amazon reviews using off-the-shelf argumentation mining. In *Proceedings of the 5th Workshop on Argument Mining*, pages 35–39, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5205

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020, Argumentative relation classification with background knowledge. *Computational Models of Argument*, pages 319–330. IOS Press.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1110

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland, Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1144

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1053

Isaac Persing and Vincent Ng. 2016a. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1164

Isaac Persing and Vincent Ng. 2016b. Modeling stance in student essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2174–2184, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1205

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? Modeling weaknesses in unpersuasive arguments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4082–4088. https://doi.org/10.24963/ijcai.2017/570

Isaac Persing and Vincent Ng. 2020. Unsupervised argumentation mining in student essays. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6795–6803, Marseille, France. European Language Resources Association.

Georgios Petasis. 2019. Segmentation of argumentative texts with contextualised word representations. In *Proceedings of the 6th Workshop on Argument Mining*, pages 1–10, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4501

Plato. ca. 400 B.C.E. *Theaetetus*. 2014 edition. Oxford University Press. Translated by John McDowell.

Edoardo M. Ponti, Alessandro Sordoni, Yoshua Bengio, and Siva Reddy. 2022. Combining modular skills in multitask learning. *arXiv preprint arXiv:2202.13914*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Aldo Porco and Dan Goldwasser. 2020. Predicting stance change using modular architectures. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 396–406, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.35

Peter Potash, Robin Bhattacharya, and Anna Rumshisky. 2017a. Length, interchangeability, and external knowledge: Observations from predicting argument convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 342–351, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Peter Potash, Adam Ferguson, and Timothy J. Hazen. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*, pages 146–155, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4517

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017b. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1364–1373, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1143

Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument search: Assessing argument relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1117–1120, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3331184.3331327

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! Leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. 2018a. Is something better than nothing? Automatically predicting stance-based arguments using deep learning and small labelled dataset. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 28–34, New

Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2005

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. 2018b. Sentiment-stance-specificity (SSS) dataset: Identifying support-based entailment among opinions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sarvesh Ranade, Rajeev Sangal, and Radhika Mamidi. 2013. Stance classification in online debates by recognizing users' intentions. In *Proceedings of the SIGDIAL 2013 Conference*, pages 61–69, Metz, France. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1054

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2015. A computational approach for generating toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55, Denver, CO. Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0507

Ruty Rinott, 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*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1050

Patrick Saint-Dizier. 2017. Using question-answering techniques to implement a knowledge-driven argument mining approach. In *Proceedings of the 4th Workshop on Argument Mining*, pages 85–90, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5111

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing. https://doi.org/10.3115/v1/P15-4019

Robin Schaefer and Manfred Stede. 2021. Argument mining on Twitter: A survey. *it-Information Technology*, 63(1):45–58. https://doi.org/10.1515/itit-2020-0053

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.34

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2006

Thomas Scialom, Serra Sinem Tekiroğlu, Jacopo Staiano, and Marco Guerini. 2020. Toward stance-based personas for opinionated dialogues. In

*Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2625–2635, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.238

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2095

Eyal Shnarch, Ran Levy, Vikas Raykar, and Noam Slonim. 2017. GRASP: Rich patterns for argumentation mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1345–1350, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1140

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371. https://doi.org/10.1162/tacl_a_00026

Joseph Sirrianni, Xiaoqing Liu, and Douglas Adams. 2020. Agreement prediction of arguments in cyber argumentation for detecting stance polarity and intensity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5746–5758, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.509

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.147

Noam Slonim. 2018. Project Debater. In *Proceedings of COMMA*, 4 pages. https://doi.org/10.3233/978-1-61499-906-5-4

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, Denver, CO. Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0509

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2088

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew 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, Washington, USA. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

Yi Song, 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, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-2110

Maximilian Spliethöver, Jonas Klaff, and Hendrik Heuer. 2019. Is it worth the attention? A comparative evaluation of attention layers for argument unit segmentation. In *Proceedings of the 6th Workshop on Argument Mining*, pages 74–82, Florence, Italy. Association for

Computational Linguistics. https://doi.org/10.18653/v1/W19-4509

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018a. Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25. https://doi.org/10.18653/v1/N18-5005

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1006

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2813

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659. https://doi.org/10.1162/COLI_a_00295

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-1092

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018b. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1402

Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191. https://doi.org/10.1007/978-3-031-02169-5

Guobin Sui, Wenhan Chao, and Zhunchen Luo. 2018. Joker at SemEval-2018 task 12: The argument reasoning comprehension with neural attention. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1129–1132, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1191

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shahbaz Syed, Roxanne El Baff, Johannes Kiesel, Khalid Al Khatib, Benno Stein, and Martin Potthast. 2020. News editorials: Towards summarizing long argumentative texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5384–5396, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.470

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54. https://doi.org/10.1177/0261927X09351676

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics. https://doi.org/10.3115/977035.977051

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502. https://doi.org/10.3115/1699648.1699696

Junfeng Tian, Man Lan, and Yuanbin Wu. 2018. ECNU at SemEval-2018 task 12: An end-to-end attention-based neural network for the argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1094–1098, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1184

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment—new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1564

Orith Toledo-Ronen, Roy Bar-Haim, and Noam Slonim. 2016. Expert stance graphs for computational argumentation. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 119–123, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2814

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.29

Stephen E. Toulmin. 2003. *The Uses of Argument*, updated edition. Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

Dietrich Trautmann. 2020. Aspect-based argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 41–52, Online. Association for Computational Linguistics.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9048–9056. https://doi.org/10.1609/aaai.v34i05.6438

Gerard A. W. Vreeswijk. 1997. Abstract argumentation systems. *Artificial Intelligence*, 90(1–2):225–279. https://doi.org/10.1016/S0004-3702(96)00041-0

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-1017

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In

*Proceedings of the 4th Workshop on Argument Mining*, pages 49–59. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5106

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017c. ''PageRank'' for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-1105

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014. Modeling review argumentation for robust sentiment analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 553–564.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.592

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802034

Hao Wang, Zhen Huang, Yong Dou, and Yu Hong. 2020. Argumentation mining on essays at multi scales. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5480–5493, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.478

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1007

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1364

Ingrid Zukerman, Richard McConachy, and Sarah George. 2000. Using argumentation strategies in automated argument generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 55–62, Mitzpe Ramon, Israel. Association for Computational Linguistics. https://doi.org/10.3115/1118253.1118262