Argument Mining: A Survey

John Lawrence
University of Dundee, UK
Centre for Argument Technology
j.lawrence@dundee.ac.uk

Chris Reed
University of Dundee, UK
Centre for Argument Technology

Argument mining is the automatic identification and extraction of the structure of inference and reasoning expressed as arguments presented in natural language. Understanding argumentative structure makes it possible to determine not only what positions people are adopting, but also why they hold the opinions they do, providing valuable insights in domains as diverse as financial market prediction and public relations. This survey explores the techniques that establish the foundations for argument mining, provides a review of recent advances in argument mining techniques, and discusses the challenges faced in automatically extracting a deeper understanding of reasoning expressed in language in general.

1. Introduction

With online fora increasingly serving as the primary media for argument and debate, the automatic processing of such data is rapidly growing in importance. Unfortunately, though data science techniques have been extraordinarily successful in many natural language processing tasks, existing approaches have struggled to identify more complex structural relationships between concepts. For example, although opinion mining and sentiment analysis provide techniques that are proving to be enormously successful in marketing and public relations, and in financial market prediction, with the market for these technologies currently estimated to be worth around $10 billion, they can only tell us *what* opinions are being expressed and not *why* people hold the opinions they do.

Justifying opinions by presenting reasons for claims is the domain of argumentation theory, which studies arguments in both text and spoken language; in specific domains and in general; with both normative and empirical methodologies; and from philosophical, linguistic, cognitive and computational perspectives. Though an enormous field with a long and distinguished pedigree (see van Eemeren et al. [2014] for a compendious review), we begin with an intuitive understanding of argument as reason-giving (and refine it later on), and focus initially on how to go about manually identifying arguments in the wild.

Submission received: 2 August 2017; revised version received: 11 August 2019; accepted for publication: 15 September 2019.

https://doi.org/10.1162/COLI_a_00364

Argument analysis aims to address this issue by turning unstructured text into structured argument data, giving an understanding not just of the individual points being made, but of the relationships between them and how they work together to support (or undermine) the overall message. Although there is evidence that argument analysis aids comprehension of large volumes of data, the manual extraction of argument structure is a skilled and time-consuming process. For example, Robert Horn, talking about the argument maps he produced on the debate as to whether computers can think, quotes a student as saying "These maps would have saved me 500 hours of time my first year in graduate school"; [1] however, Metzinger (1999) notes that over 7,000 hours of work was required in order for Horn and his team to create these maps.

Although attempts have been made to increase the speed of manual argument analysis, it is clearly impossible to keep up with the rate of data being generated across even a small subset of areas and, as such, attention is increasingly turning to **argument mining**,[2] the automatic identification and extraction of argument components and structure. The field of argument mining has been expanding rapidly in recent years (with ACL workshops on the topic being held annually, from the first in 2014,[3] up to the most recent in 2019,[4] which received a record number of 41 submissions. These have been complemented by further workshops organized in Warsaw,[5] Dundee,[6] Dagstuhl,[7] and tutorials at IJCAI,[8] ACL 2016,[9] ACL 2019,[10] and ESSLLI.[11]) This increasing activity makes a comprehensive review of both timely and practical value.

Previous reviews, including Palau and Moens (2009) and Peldszus and Stede (2013a), predated this explosion in the volume of work in the area, whereas more contemporary reviews are aimed at different audiences: Budzynska and Villata (2017) at the computational argumentation community and Lippi and Torroni (2016) at a general computational science audience. Most recently, Stede and Schneider (2018) have, in their 2018 tour de force, assembled an extensive review of performance on tasks in, and related to, argument mining. Our goal here is to update and extend, introducing reorganization where more recent results suggest different ways of conceptualizing the field. Our intended audience are those already familiar with computational linguistics, so we spend proportionally more time on those parts of the story that may be less familiar to such an audience, and rather less on things that represent mainstays of modern research in computational linguistics. With this goal in mind we also move on from Stede and Schneider (2018) in three ways. First, we bring the discussion up to date with the newest results based on approaches such as Integer Linear Programming, transfer learning, and new attention management methods, and cover a much larger range of data sources: For a discipline that is so increasingly data-hungry, we review annotated data sources covering over 2.2 million words. Second, we provide greater depth in discussion of foundational topics—covering both the rich heritage of philosophical research in the analysis and understanding of argumentation, as well as those areas and techniques in computational linguistics that lay the groundwork for

1 http://www.stanford.edu/~rhorn/a/topic/phil/artclTchngPhilosphy.html.
2 Sometimes also referred to as argumentation mining.
3 http://www.uncg.edu/cmp/ArgMining2014/.
4 https://argmining19.webis.de/.
5 http://argdiap.pl/argdiap2014.
6 http://www.arg-tech.org/swam2014/.
7 https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=16161.
8 http://www.i3s.unice.fr/~villata/tutorialIJCAI2016.html.
9 http://acl2016tutorial.arg.tech/.
10 http://arg.tech/~chris/acl2019tut/index.html.
11 https://www.irit.fr/esslli2017/courses/20.

much current argument mining work. Thirdly and finally, the simple pipeline view of argument mining, which characterizes a lot of both older research work and reviews, is increasingly being superceded by more sophisticated and interconnected techniques; here we adopt a more network view of subtasks in argument mining and focus on the interconnections and dependencies between them.

We look first, in Section 2, at existing work in areas that form the foundation for many of the current approaches to argument mining, including sentiment analysis, citation mining, and argumentative zoning. In Section 3 we look at the task of manual argument analysis, considering the steps involved and tools available, as well as the limitations of manually analyzing large volumes of text. Section 4 discusses the argumentation data available to those working in the argument mining field, as well as the limitations and challenges that this data presents. In Section 5, we provide an overview of the tasks involved in argument mining before giving a comprehensive overview of each in Sections 6, 7, and 8.

2. Foundational Areas and Techniques

In this section, we look at a range of different areas that constitute precursors to the task of argument mining. Although these areas are somewhat different in their goals and approach, they all offer techniques that at least form a useful starting point for determining argument structure. We do not aim to present a comprehensive review of these techniques in this section, but, instead, to highlight their key features and how they relate to the task of argument mining.

In Section 2.1, we present an overview of opinion mining, focusing specifically on its connection to argument mining. Section 2.2 looks at Controversy Detection, an extension of opinion mining that aims to identify topics where opinions are polarized. Citation mining, covered in Section 2.3, looks at citation instances in scientific writing and attempts to label them with their rhetorical roles in the discourse. Finally, in Section 2.4, we look at argumentative zoning, where scientific papers are annotated at the sentence level with labels that indicate the rhetorical role of the sentence (criticism or support for previous work, comparison of methods, results or goals, etc.).

2.1 Opinion Mining

As the volume of online user-generated content has increased, so too has the availability of a wide range of text offering opinions about different subjects, including product reviews, blog posts, and discussion groups. The information contained within this content is valuable not only to individuals, but also to companies looking to research customer opinion. This demand has resulted in a great deal of development in techniques to automatically identify opinions and emotions.

Opinion mining is "the computational study of opinions, sentiments, and emotions expressed in text" (Liu 2010). The terms "opinion mining" and "sentiment analysis" are often used interchangeably; however, sentiment analysis is specifically limited to positive and negative views, whereas opinion mining may encompass a broader range of opinions.

The link between sentiment, opinion, and argumentative structure is described in Hogenboom et al. (2010), where the role that argumentation plays in expressing and promoting an opinion is considered and a framework proposed for incorporating information on argumentation structure into the models for economic sentiment discovery

in text. Based on their role in the argumentation structure, text segments are assigned different weights relating to their contribution to the overall sentiment. Conclusions, for example, are hypothesized to be good summaries of the main message in a text and therefore key indicators of sentiment. The interesting point here, from an argument mining perspective, is that this theory could equally be reversed and sentiment be used as an indicator of the argumentative process found in a text. Taking the example of conclusions, those segments that align with the overall sentiment of the document are more likely to be a conclusion than those that do not.

Many applications of sentiment analysis are carried out at the document level to determine an overall positive or negative sentiment. For example, in Pang, Lee, and Vaithyanathan (2002), topic-based classification using the two "topics" of positive and negative sentiment is carried out. To perform this task, a range of different machine learning techniques (including support vector machines [Cortes and Vapnik 1995], maximum entropy, and naïve Bayes [Lewis 1998]) are investigated. Negation tagging is also performed using a technique from Das and Chen (2001) whereby the tag `NOT_` is prepended to each of the words between a negation word ("not," "isn't," "didn't," etc.) and the first punctuation mark occurring after the negation word. In terms of relative performance, the support vector machines (SVMs) achieved the best results, with average 3-fold cross-validation accuracies over 0.82 using the presence of unigrams and bigrams as features.

Shorter spans of text are also considered in Grosse, Chesñevar, and Maguitman (2012), who look at microblogging platforms such as Twitter with the aim of mining opinions from individual posts to build an "opinion tree" that can be built recursively by considering arguments associated with incrementally extended queries. Sentiment analysis tools are used to determine the overall sentiment for an initial one word query, which is then extended and the change in overall sentiment recalculated. By following this procedure, it is possible to see where extending the query results in a change of overall sentiment and, as such, to determine those terms that introduce conflict with the previous query. Conflicting elements in an opinion tree are then used to generate a "conflict tree," similar to the dialectical trees (Prakken 2005) used traditionally in defeasible argumentation (Pollock 1987).

Opinion mining, however, is not limited to just determining positive and negative views. In Kim and Hovy (2006b) sentences from online news media texts are examined to determine the topic and proponent of opinions being expressed. The approach uses semantic role labeling to attach an opinion holder and topic to an opinion-bearing word in each sentence using FrameNet[12] (a lexical database of English, based on manual annotation of how words are used in actual texts). To supplement the FrameNet data, a clustering technique is used to predict the most probable frame for words that FrameNet does not include. This method is split into three subtasks:

1. Collection of opinion words and opinion-related frames—1,860 adjectives and 2,011 verbs classified into positive, negative, and neutral. Clustering By Committee (Pantel 2003) is used to find the closest frame. This method uses the hypothesis that words that occur in the same context tend to be similar.
2. Semantic role labeling for those frames. A maximum entropy model is used to classify frame element types (Stimulus, Degree, Experiencer, etc.)

12 `https://framenet.icsi.berkeley.edu/`.

3. Mapping of semantic roles to the opinion holder and topic. A manually built mapping table maps Frame Elements to a holder or topic.

Results show an increase from the baseline of 0.30 to 0.67 for verb target words and of 0.38 to 0.70 for adjectives, with the identification of opinion holders giving a higher F-score[13] than topic identification.

Although understanding the sentiment of a document as a whole could be a useful step in extracting the argument structure, the work carried out on sentiment analysis at a finer-grained level perhaps offers greater benefit still. In Wilson, Wiebe, and Hoffmann (2005), an approach to phrase-level sentiment analysis is presented, using a two-step process: first, applying a machine learning algorithm to classify a phrase as either neutral or polar (for which an accuracy of 0.76 is reported); and then looking at a variety of features in order to determine the contextual polarity (*positive*, *negative*, *both*, or *neutral*) of each polar phrase (with an accuracy of 0.62–0.66, depending on the features used).

In Sobhani, Inkpen, and Matwin (2015), we see an example of extending simple pro and con sentiment analysis, to determine the stance which online comments take toward an article. Each comment is identified as "Strongly For," "For," "Other," "Against," and "Strongly Against" the original article. These stances are then linked more clearly to the argumentative structure by using a topic model to determine what is being discussed in each comment, and classify it to a hierarchical structure of argument topics. This combination of stance and topic hints at possible argumentative relations—for example, comments about the same topic that have opposing stance classifications are likely to be connected by conflict relations, whereas those with similar stance classifications are more likely to connect through support relations.

In Kim and Hovy (2006a), the link between argument mining and opinion mining is clearer still. Instead of looking solely at whether online reviews are positive or negative, a system is developed for extracting the reasons *why* the review is positive or negative. Using reviews from epinions.com, which allows a user to give their review as well as specific positive and negative points, these specific positive and negative phrases were first collected and then the main review searched for sentences that covered most of the words in the phrase. Using this information, sentences were classified as "pro" or "con" with unmatched sentences classified as "neither." Sentences from further reviews were then classified as, first, "pro" and "con" against "neither" followed by classification into "pro" or "con." The best feature selection results in an F-score of 0.71 for reason identification and 0.61 for reason classification.

2.2 Controversy Detection

One extension to the field of opinion mining that has particular relevance to argument mining is controversy detection, where the aim is to identify controversial topics and text where conflicting points of view are being presented. The most clear link between controversy and argument detection can be seen in Boltužić and Šnajder (2015), where argumentative statements are clustered based on their textual similarity, in order to identify prominent arguments in online debates. Controversy detection to date has

13 F-score refers to the equally weighted harmonic mean of the precision and recall measured for a system. When the system is applied to several sets of data, the micro-average F-score is obtained by first summing up the individual true positives, false positives, and false negatives and then calculating precision and recall using these figures, whereas the macro-average F-score is calculated by averaging the precision and recall of the system on the individual sets (van Rijsbergen 1979).

largely targeted specific domains: Kittur et al. (2007), for example, look at the cost of conflict in producing Wikipedia articles, where conflict cost is defined as "excess work in the system that does not directly lead to new article content." Conflict revision count (CRC), a measure counting the number of revisions in which the "controversial" tag was applied to the article, is developed and used to train a machine learning model for predicting conflict. Computing the CRC for each revision of every article on Wikipedia resulted in 1,343 articles for which the CRC score was greater than zero (meaning they had at least one "controversial" revision). Of these, 272 articles were additionally marked as being controversial in their most recent revision. A selection of these 272 articles is then used as training data for an SVM classifier. Features are calculated from the specific page such as the length of the page, how many revisions were carried out, links from other articles, and the number of unique editors. Of these features, the number of revisions carried out is determined to be the most important indicator of conflict; and by predicting the CRC scores using a combination of page metrics, the classifier is able to account for approximately 90% of the variation in scores. It is reasonable to assume that the topics covered on those pages with a high CRC are controversial and, therefore, topics for which more complex argument is likely to occur.

The scope of controversy detection is broadened slightly in Choi, Jung, and Myaeng (2010) and Awadallah, Ramanath, and Weikum (2012), who both look at identifying controversy in news articles. In Choi, Jung, and Myaeng (2010), a controversial issue is defined as "a concept that invokes conflicting sentiments or views" and a subtopic as "a reason or factor that gives a particular sentiment or view to the issue." A method is proposed for the detection of controversial issues, based on the magnitude of sentiment information and the difference between the magnitudes for two different polarities. First, noun and verb phrases are identified as candidate issues using a mixture of sentiment models and topical information. The degree of controversy for these issues is calculated by measuring the volume of both positive and negative sentiment and the difference between them. For subtopic extraction, noun phrases are identified as candidates and, for these phrases, three statistical features (contextual similarity between the issue and a subtopic candidate, relatedness of a subtopic to sentiment, and the degree of physical vicinity between the issue and the candidate phrases) as well as two positional features are calculated. The results for subtopic identification are poor, with an F-score of 0.50; however, identifying controversial issues is considerably more successful, with a precision of 0.83.[14]

Awadallah, Ramanath, and Weikum (2012) present the OpinioNetIt system, which aims to automatically derive a map of the opinions-people network from news and other Web documents. The network is constructed in four stages. First, generic terms are used to identify sample controversial topics; next, opinion holders are identified for each topic, and their opinions extracted; the acquired topics and opinion holders are then used to construct a lexicon of phrases indicating support or opposition. Finally, this process is performed iteratively using the richer lexicon to identify more opinion holders, opinions, and topics. Using this approach a precision of 0.72 is achieved in classifying controversial opinions.

Despite the specific domain limitations of this controversy detection work, Dori-Hacohen and Allan (2013) extend their scope to detecting controversy on the Web as a whole, enabling users to be informed of controversial issues and alerted when alternative

14 The precision is calculated based upon a user study where the participants are asked to confirm if an issue is controversial; as such, recall is not reported.

viewpoints are available. This is achieved by first mapping a given Web page to a set of neighboring Wikipedia articles labeled on a controversiality metric, then combining the labels to give an estimate of the page's controversiality, which is finally converted into a binary value using a threshold. This approach gives a 22% increase in accuracy over a sentiment-based approach, indicating that, although closely related, detecting controversy is more complex than simply detecting opinions and looking at where they differ.

Such widespread use of controversy detection offers the ability to address potential hotspot issues as they arise and the possibility of dealing with conflict in a debate at an early stage, before the quality of discussion can be negatively impacted. Rumshisky et al. (2017), for example, take advantage of both content- and graph-based features to analyze the dynamics of social or political conflict as it develops over time, using a combination of measures of conflict intensity derived from social media data. Such methods for determining controversial issues can play a significant role in determining the argumentative structure inherent in a piece of text. Those points that are controversial are likely to attract not only more attention, but also a more even mix of supporting and attacking views, than those on which there is broad consensus. Lawrence et al. (2017) make this connection explicit, showing how the divisiveness, or controversiality, of a proposition might be based upon the relative number of its supports and conflicts. A proposition with many of both might be taken to be divisive, whereas few of either might suggest only limited divisiveness. Alternatively, given a pair of propositions that are in conflict, the divisiveness of this conflict is shown to be a measure of the amount of support on both sides. It is easy to see how this process could be reversed, meaning that if we are able to identify controversial points in a piece of text, we already know something about the argumentative structure.

2.3 Citation Mining

Citation mining involves the labeling of citation instances in scientific writing with their rhetorical roles in the discourse. The techniques used to automatically determine the motivating factors behind each citation map closely to applications in argument mining, where text spans are labeled based on their argumentative role. For example, if a citation is being used to highlight a gap or deficiency in the referenced work, then the language used will be suggestive of conflict relations between the two; if a citation is being used to back up the current work, then there are likely argumentative support relations between the two.

There are a broad range of manual schemes for classifying citation motivation and citation function (the reason why an author chooses to cite a paper), and Teufel, Siddharthan, and Tidhar (2006) look at how this classification can be automated. A classification scheme is first developed using guidelines for twelve different categories (explicit statement of weakness, four types of contrast/comparison, six types of agreement/usage, and neutral). Human annotators testing this scheme achieve a κ[15] of 0.72 and, when implemented as an automatic procedure with the features listed below, an accuracy of 0.77 and κ of 0.57 is achieved (or accuracy 0.83, κ 0.58 for 3-way classification

15 κ is a statistical measure of inter-rater agreement, measuring pairwise agreement among a set of coders and correcting for expected chance agreement (Carletta 1996). An interpretation of kappa values is offered by Landis and Koch (1977), who describe values between 0.01 and 0.20 as showing slight agreement, 0.21 and 0.40 fair agreement, 0.41 and 0.60 moderate agreement, 0.61 and 0.80 substantial agreement, and 0.81 and 1.00 almost perfect agreement.

positive/negative/neutral) based on an evaluation corpus of 116 articles, containing 2,829 citations.

- Cue phrases
- Cues identified by annotators: 892 cue phrases identified by annotators (around 75 per category)
- Verb tense and voice used for recognizing statements of previous/future/current work
- Location in paper/sentence/paragraph
- Self citations identified by author name

Kappa is even higher for the top-level distinction; collapsing the similar categories into just four (statement of weakness, contrast/comparison, agreement/usage, and neutral) gives a κ value of 0.59. By comparison, the human agreement for this configuration is $\kappa = 0.76$. Although this leaves a significant gap between automated and human performance, it nevertheless suggests "moderate agreement" using the automated approach, an encouraging result for a complex task.

An attempt to classify the opinion an author holds toward a work that they cite (for example, positive/negative attitudes or approval/disapproval) is presented in Piao et al. (2007), where semantic lexical resources and NLP tools are used to create a network of opinion polarity relations. Sentences containing citations are extracted first, before determining the opinion orientation of the subjective words in the context of the citation. From these opinion orientations, the attitude of the author toward the work that they are citing is labeled.

Athar (2011) takes a similar approach, whereby analysis is performed on a corpus of scientific texts taken from the ACL Anthology, and consisting of 8,736 citations from 310 research papers manually annotated for their sentiment. Sentences are labeled as positive, negative, or objective, with 1,472 used for development and training. Each citation is represented as a feature set in a SVM and processed using WEKA (Holmes, Donkin, and Witten 1994) and the WEKA LibSVM library with the following features:

- **Word Level Features** Unigrams and bigrams as well as 3-grams to capture longer technical terms. POS tags are also included using two approaches: attaching the tag to the word by a delimiter, and appending all tags at the end of the sentence. A science-specific sentiment lexicon is also added, consisting of 83 polar phrases such as efficient, popular, successful, state-of-the-art, and effective.
- **Contextual Polarity Features** Sentence-based features, for example, presence of subjectivity clues that have been compiled from several sources along with the number of adjectives, adverbs, pronouns, modals, and cardinals.
- **Dependency Structures** Typed dependency structures (De Marneffe and Manning 2008) describing the grammatical relationships between words. For instance, in the sentence "CITE showed that the results for French-English were competitive to state-of-the-art alignment systems,"

the relationship between results and competitive will be missed by trigrams but the dependency representation captures it in a single feature `nsubj_competitive_results`.

- **Sentence Splitting** Each sentence is split by trimming its parse tree. Walking from the citation node toward the root, the subtree rooted at the first sentence node is selected and the rest ignored.
- **Negation** All words inside a *k*-word window of any negation term are suffixed with a token `_neg` to distinguish them from their non-polar versions.

The results show that 3-grams and dependencies perform best in this task with macro F-score 0.76 and micro F-score 0.89.

2.4 Argumentative Zoning

Argumentative zoning (AZ) is the classification of sentences by their rhetorical and argumentative role within a scientific paper. For example, criticism or support for previous work, comparison of methods, and results or goals. Although this approach of labeling a sentence by its role is slightly removed from the goal of identifying the argumentation structure contained within the document, it is clear that the information obtained by AZ provides a useful step toward determining the structure.

In Teufel, Siddharthan, and Batchelor (2009), an annotation scheme covering 14 possible roles is used to classify sentences into mutually exclusive categories. These categories extend the original seven categories presented in Teufel, Carletta, and Moens (1999) and are designed to be applied to material from the life sciences domain as well as to the *Computational Linguistics* (CL) material considered in the earlier work. This categorization highlights the link between AZ and argument mining. The 'AIM' (statement of specific research goal, or hypothesis of current paper) and 'OWN CONC' (findings, conclusions (non-measurable) of own work) categories, for example, are suggestive of conclusions. 'NOV ADV' (novelty or advantage of own approach) and 'SUPPORT' (other work supports current work or is supported by current work) are suggestive of support relations, and 'GAP WEAK' (lack of solution in field, problem with other solutions) and 'ANTISUPP' (clash with somebody else's results or theory) are suggestive of conflict relations.

Teufel et al. use a domain expert to encode basic knowledge about the subject, such as terminology and domain specific rules for individual categories, as part of the annotation guidelines. The produced guidelines include a decision tree, descriptions of the semantic nature of each category, rules for pairwise distinction of the categories, and a large range of examples taken from both chemistry and computational linguistics. Human coders with background knowledge in computational linguistics, and varied experience in chemistry, applied these guidelines, achieving inter-annotator agreement for chemistry with $\kappa = 0.71$ (N=3745, n=15, k=3). For CL, the inter-annotator agreement was $\kappa = 0.65$ (N=1629, n=15, k=3). As a comparison, the inter-annotator agreement for Teufel's original, CL-specific AZ with seven categories (Teufel, Carletta, and Moens 1999) was $\kappa = 0.71$ (N=3420, n=7, k=3). This level of agreement between the three annotators is acceptable overall and supports the hypothesis that the task definition is domain-knowledge free. However, agreements involving the semi-expert are higher than the agreement between expert and non-expert, indicating that a general

understanding of basic chemistry was not sufficiently adequate to ensure that the non-expert understood enough of the material to achieve the highest-possible agreement.

Merity, Murphy, and Curran (2009) present a maximum entropy classifier with each sentence of an article classified into one of the seven basic rhetorical structures from Teufel, Carletta, and Moens (1999). A maximum entropy model combined with the addition of new features to those used by Teufel et al. gives an increase from 0.76 to 0.97 F-score on Teufel's *Computational Linguistics* conference paper corpus (48 computational linguistics papers, taken from the proceedings of the COLING, ANLP, and ACL conferences between April 1994 and April 1995). The features used are described below:

- **Unigrams, bigrams, and *n*-grams** Unigram and bigram features were included and reported individually and together (as *n*-grams). These features include all of the unigrams and bigrams above the feature cutoff.
- **First** The first four words of a sentence, added individually.
- **Section** A section counter that increments on each heading to measure the distance into the document.
- **Location** The position of a sentence between two headings (representing a section).
- **Paragraph** The position of the sentence within a paragraph.
- **Length** Length of sentence grouped into multiples of 3.
- **Teufel et al.'s (1999) features** To compare with previous work, most of the features that gave Teufel et al. the best performance are also implemented.
- **Feature cutoff** Instead of including every possible feature, a cutoff was used to remove features that occur less than four times.
- **History features** History features were used and AZ treated as a sequence labeling task with history lengths ranging from previous label to the previous four labels.

The results show that *n*-grams have by far the largest impact, with a 21.39% reduction in accuracy when they are removed (the next largest impact being 1.24% for the first four words of the sentence). The history features also have an impact of just over 1%. It is shown that none of Teufel et al.'s individual features alone make a substantial contribution to the results when using the maximum entropy model. To evaluate the wider applicability of AZ, a corpus of *Astronomy* journal articles was also annotated with a modified zone and content scheme, and a similar level of performance (around 0.96 accuracy) was achieved.

3. Manual Argument Analysis

In this section we look at the task of manual argument analysis, considering the steps involved and tools available, as well as the limitations of manually analyzing large volumes of text. Understanding manual analysis can offer unique insight into how this task can be automated and provides a valuable insight into how an analyst unpicks the complex argumentative relationships represented in natural language texts.

Although the argumentative structure contained within a piece of text (van Eemeren et al. 2014) can be diagrammed manually using pen and paper or simple

graphics software, a wide range of specific argument diagramming tools (Scheuer et al. 2010) has been developed to allow an analyst to identify the argumentative sections of the text and diagram the structure that they represent (Kirschner, Buckingham-Shum, and Carr 2003; Okada, Shum, and Sherborne 2008). The advantages of this approach, as opposed to the use of non-specialized software, are discussed in Harrell (2005), though there is varied (and conflicting evidence of) impact on the the day-to-day activity within domains in which these tools are applied, such as law, pedagogy, scientific writing (Lauscher, Glavaš, and Eckert 2018; Lauscher, Glavaš, and Ponzetto 2018), and design (Scheuer et al. 2010). The majority of these tools, such as Araucaria (Reed and Rowe 2004), Rationale (van Gelder 2007), OVA (Bex et al. 2013), and Carneades (Gordon, Prakken, and Walton 2007), require the analyst to manually identify the propositions involved in the argument being made and then connect them identifying the premises and conclusion. In many cases, this simple structure can then be extended with more specialized information, depending on the nature of the analysis task being performed; for example, giving details of the argumentation schemes (Walton, Reed, and Macagno 2008) used or details of the participants and their dialogical moves (for example, *questioning* or *asserting*) when analyzing dialogue.

Generally, manual argument analysis, as carried out using the tools previously mentioned, can be split into four distinct stages, as shown in Figure 1.

Though both manual and automated analysis techniques may develop a more complex, hybrid approach in practice, the pipeline model presented here offers a good starting point from which to introduce the range of techniques currently available. Then in Section 5 we further dissect these steps, presenting a more detailed view of the individual argument mining steps and how they relate to the manual annotation process, explaining how increasingly the pipeline view oversimplifies complex interdependencies.

3.1 Text Segmentation

Text segmentation involves the extraction of the fragments of text from the original piece that will form the constituent parts of the resulting argument structure. Text segmentation can be considered as the identification of a form of elementary discourse

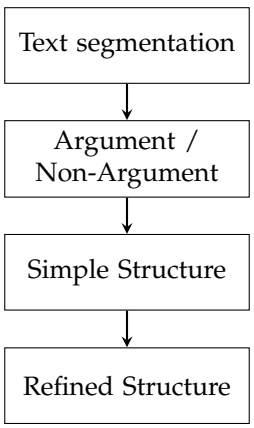


Figure 1
Steps in argument analysis.

units (EDUs). Though there are competing hypotheses about what constitutes an EDU (for example, Grimes [1975] and Givón [1983] view them as clauses, whereas Hirschberg and Litman [1993] view them as prosodic units, Sacks, Schegloff, and Jefferson [1974] as turns of talk, Polanyi [1988] as sentences, and Grosz and Sidner [1986] as intentionally defined discourse segments), all agree that EDUs are non-overlapping spans of text corresponding to the atomic units of discourse. Peldszus and Stede (2013a) refer to these argument segments as "argumentative discourse units" (ADUs), and define an ADU as a "minimal unit of analysis," pointing out that an ADU may not always be as small as an EDU, for example, "when two EDUs are joined by some coherence relation that is irrelevant for argumentation, the resulting complex might be the better ADU" (page 20).

Generally speaking, in argument analysis, the sections that the analyst extracts correspond to the propositions contained within the text; however, some knowledge of the argument being made is often required in order to determine the exact boundaries of these propositions and how fine-grained the segmentation needs to be. In some cases, for example, propositional content can occur nested in reported speech, such as the sentence "Simon said this is a blue pen." The rest of the argument structure may refer to either the whole sentence ("Simon didn't say that"), to the statement "this is a blue pen" ("it's clearly a black pen"), or to both parts separately ("Yes, I heard Simon say that, but he's wrong, it's a black pen"). Another challenging example is **dislocation** which, similar to cleft constructions in syntax (Lasnik and Uriagereka 1988), occurs when one segment is embedded into another, such as the example given in Saint-Dizier (2012): "Products X and Y because of their toxicity are not allowed in this building." In this case the conclusion, "Products X and Y are not allowed in this building," is split around the premise "because of their toxicity." As these examples show, robustly identifying the text segments required for an analysis can be challenging even for a human analyst.

An additional complication can occur in cases where some reconstruction of the argument is required in order to identify the points being made. There is a tendency for arguers to leave implicit an assumption required in order for their conclusion to follow from their premises. This can often occur when the omitted proposition is believed to be obvious; however, it can also happen for a range of other reasons, for example, to increase the rhetorical force of the argument, or to conceal its unsoundness. Such missing premises are referred to as **enthymemes** (Hitchcock 1985), and can cause difficulties for both automatic and manual segmentation due to the requirement of knowledge that may be outside the scope of that expressed in the text.

3.2 Argument / Non-Argument Classification

This step involves determining which of the segments previously identified are part of the argument being presented and which are not. For most manual analysis tools this step is performed as an integral part of segmentation: The analyst simply avoids segmenting any parts of the text that are not relevant to the argument. However, in some cases, for example, where segmentation has been performed automatically or by a different analyst, this step must be carried out independently. In these cases the judgment as to whether a particular segment is argumentative can be made as a preliminary step in determining the structure, or left until the end of the analysis, when any segments left unconnected to the rest of the structure can simply be discarded.

Looking at the text shown in Example (1), we can see that the majority of Michael Buerk's introduction of Nick Dearden is non-argumentative, with only the single claim identified that Mr Dearden would like people not to have to pay their debts. Meanwhile, almost the entirety of the response (excluding brief connectives) forms part of the argument structure.

> **Michael Buerk:** John Lamiday, thank you very much indeed for joining us this evening. Our third witness is Nick Dearden, who is director of the Jubilee Debt Campaign. *Mr Dearden, you'd like people not to have to pay their debts*. Where's the morality in that?
>
> **Nick Dearden:** *I wouldn't like people not to have to pay their debts across the board*. But I think what we say is that *this isn't simply a matter of individual morality. Debt is used time and again as a set of economic decisions, and political decisions, to achieve certain things in society*. And *very often what high levels of debt can mean, and especially when the debt is on very unjust terms, is a massive redistribution of wealth in society, from the poorest to the richest*.

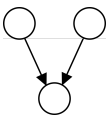
Example (1) Excerpt from the BBC Moral Maze 'Money' corpus (`http://corpora.aifdb.org/Money`). Argumentative segments are *highlighted*.

In some cases, however, this task can be remarkably demanding. Letters to the Editor contributions, for example, can sometimes offer rich pickings for the argument analyst, but such letters can often be little more than frivolity or wit masquerading as argument and inference. Distinguishing argument from non-argument in this domain is extremely demanding, even for a highly trained human analyst.

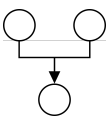
3.3 Simple Structure

Once the elements of the argument have been determined, the next step is to examine the links between them. This can be as simple as noting segments that are thematically related, but usually involves the identification of support and attack relations between segments. Although these relations can be simply labeled pairs, it is common to consider the varying ways in which components can work together (Groarke, Tindale, and Fisher 1997):

Convergent Arguments In a convergent argument, multiple premises are used to independently support a single conclusion. In this case the premises act on their own and the removal of one premise from the argument does not weaken the others. From Example (1) we can see that "what we say is that this isn't simply a matter of individual morality" and "I wouldn't like people not to have to pay their debts across the board" independently support "Mr Dearden would like people not to have to pay their debts."

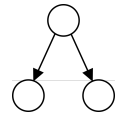


Linked Arguments In a linked argument, multiple premises work together to support a conclusion. The important point here is that each premise requires the others in order to work fully. In Example 1, the statements "Debt is used time and again as a set of economic decisions, and political decisions, to achieve certain things in society" and "very often what high levels of debt can mean, and especially when the debt is on very unjust terms, is a massive redistribution



of wealth in society, from the poorest to the richest" work together to support the point "what we say is that this isn't simply a matter of individual morality."

Divergent Arguments In some cases the same premise may support multiple conclusions. Divergent arguments are somewhat less common and, as such, are not supported by those analysis tools which, for example, are limited to analyzing arguments in a tree structure.



Though Example (1) does not include a divergent argument, Dearden might have said, "And if it's not individual morality, then the state should take some of the responsibility," which would have offered a second conclusion based on the premise of individual morality.

Sequential (or Serial) Arguments The final way in which multiple premises can support a conclusion is in a sequential argument. In this case, one premise leads to another and this, in turn, leads to the conclusion. In Example (1), the statements "very often what high levels of debt can mean, and especially when the debt is on very unjust terms, is a massive redistribution of wealth in society, from the poorest to the richest," "what we say is that this isn't simply a matter of individual morality," and "Mr Dearden would like people not to have to pay their debts" follow a sequential structure.

Hybrid Argument Structure More complicated arguments, such as that in Example (1), usually involve several instances and combinations of the above elements into a larger, hybrid, argument structure. The complete analyzed structure of Example (1) can be seen in Figure 2. We must also consider conflict, or attack, relations between propositions. These include both standard conflict relations where one proposition directly conflicts with another, as well as more complex forms of defeating an argument (Pollock 1986):

Rebutting Attacks Rebutting arguments express a position that is directly incompatible with a conclusion (Pollock 1986, page 38). Later in the debate from which Example (1) is drawn, an opponent, Michael Portillo, says, "People who lend money, that is to say, people who save money, say through building societies, are very ordinary

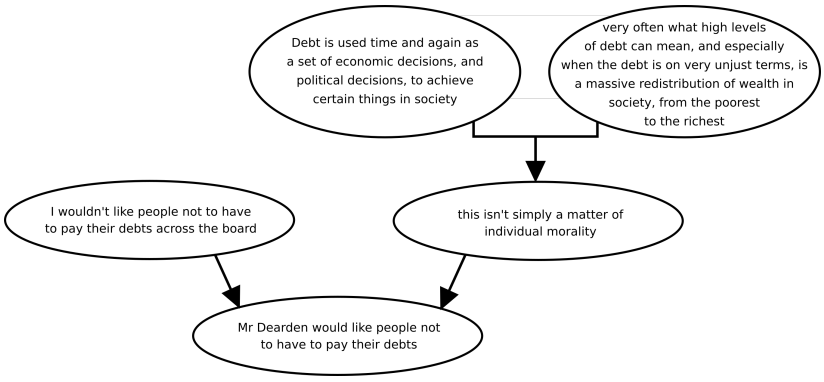


Figure 2
Simple argument structure of the text in Example (1).

people." This offers a direct, rebutting attack to Dearden's conclusion, expressed in Example (1), that debt is a massive redistribution of wealth from the poorest to the richest.

Undercutting Attacks Undercutting arguments attack or conflict with the inference between a premise and a conclusion, and, as such, offer a reason for no longer believing the conclusion, rather than for believing the negation of the conclusion (Pollock 1986, page 39). Though the fragment of debate from which Example (1) is drawn does not offer clear examples of undercutting, Portillo might have retorted with, "If there were politicial decisions being taken, they are being taken by elected officers—so state actions don't require more than individual morality." Such an attack does not directly counter the conclusion, but instead focuses on the robustness of the passage from premise to conclusion.

Although this approach to identifying argument structure is by far the most common, other methodologies, such as Toulmin (1958) are also widely used; perhaps the clearest synthesis for computational purposes is presented by the philosopher J. B. Freeman (Freeman 1991, 2011). For argument mining, successful extraction of argument structure in one form can often be translated, modulo expressivity constraints, into others (we discuss different argument representations and formats as well as the translation between them in 4).

3.4 Refined Structure

Having determined the basic argumentative structure, some analysis tools allow this to be refined further. For example, Araucaria, Carneades, Rationale, and OVA allow the analyst to identify the argumentation scheme related to a particular structure. Argumentation schemes are patterns of inference, connecting a set of premises to a conclusion, that represent stereotypical patterns of human reasoning. Such schemes were originally viewed as rhetorical methods by which a speaker could influence their audience; later, they have also been adopted as a way to distinguish good arguments from bad. Argumentation schemes can thus be seen as a historical descendant of the topics of Aristotle (1958), and, much like Aristotle's topics, play a valuable role in both the construction and evaluation of arguments. Arguments are evaluated based on a set of critical questions corresponding to the scheme which, if not answered adequately, result in the argument to which the scheme corresponds defaulting.

The Argument from Expert Opinion scheme (Walton 1996) is commonly used to illustrate the concept:

Major Premise: Source E is an expert in subject domain S containing proposition A.
Minor Premise: E asserts that proposition A is true (false).
Conclusion: A is true (false).

with the associated critical questions:

1. *Expertise Question:* How credible is E as an expert source?
2. *Field Question:* Is E an expert in the field F that A is in?
3. *Opinion Question:* What did E assert that implies A?
4. *Trustworthiness Question:* Is E personally reliable as a source?
5. *Consistency Question:* Is A consistent with what other experts assert?
6. *Backup Evidence Question:* Is E's assertion based on evidence?

Recent study has resulted in the identification and analysis of the most important and commonly used schematic structures (Hastings 1963; Perelman and Olbrechts-Tyteca 1969; Kienpointer 1992; Pollock 1995; Walton 1996; Grennan 1997; Katzav and Reed 2004; Walton, Reed, and Macagno 2008). Although there is much overlap in these classifications, they often differ in their granularity: Pollock identifies fewer than 10 schemes; Walton, nearly 30; Grennan, more than 50; and Katvaz and Reed, more than 100. Due to these differences, it is common for analysis tools to retain the grouping of schemes into sets. Araucaria, for example, supports the Walton, Grennan, Perelman and Olbrechts-Tyteca, Katzav and Reed, and Pollock scheme sets.

Experiments on the annotation of Walton schemes by annotators with a strong background in linguistics but who were provided with only the description of the schemes given in Walton, Reed, and Macagno (2008) have shown that this is an exceptionally difficult task, with results differing in both numbers of arguments annotated and the distributions of units (Lindahl, Borin, and Rouces 2019). However, recent developments in annotation guidelines for these schemes, including the decision tree–based method described in Lawrence, Visser, and Reed (2019), suggest that this situation can be improved and offer hope for the construction of scheme annotated corpora.

3.5 Limitations of Manual Analysis

Although these tools can be used for the analysis of small sections of text, analyzing large volumes of text quickly, and certainly in anything approaching real time, is beyond their scope. Compendium[16] IBIS map facilitators are the closest, but the analysis involved is at a much higher level. The major limitation is the amount of information that can be handled by a single analyst. Efforts have been made to overcome this obstacle by both crowdsourcing of annotation (Ghosh et al. 2014) and using hardware designed to allow multiple trained annotators to collaborate on the same analysis (Bex et al. 2013). In the first case, by applying a clustering technique to identify which pieces of text were easier or harder for trained experts to annotate, it was determined that the crowdsourced results were only accurate for those segments that were identified as being easier for expert annotators. In the second case, although the AnalysisWall (a touchscreen measuring 11 feet by 7 feet running bespoke analysis software) has been used to analyze several hour-long radio programs in real time, it still does not come close to allowing for the analysis of the vast volumes of data produced every day.

4. Argument Data

One of the challenges faced by current approaches to argument mining is the lack of large quantities of appropriately annotated arguments to serve as training and test data. Several recent efforts have been made to improve this situation by the creation of corpora across a range of different domains.

For example, Green (2014) aims to create a freely available corpus of open-access, full-text scientific articles from the biomedical genetics research literature, annotated to support argument mining research. However, there are challenges to creating such corpora, such as the extensive use of biological, chemical, and clinical terminology in the BioNLP domain. These challenges are highlighted in Green (2015), where preliminary

16 http://compendium.open.ac.uk/.

work on guidelines for the manual identification of 10 custom argumentation schemes targeted at genetics research articles is presented. For example, one of the schemes presented, Failed to Observe Effect of Hypothesized Cause, looks for situations where specific properties were not observed, and where it is assumed that a specific condition that would result in those properties is present, leading to the conclusion that the condition may not be present. Twenty-three students were assessed on their ability to identify instances of these schemes after having read the guidelines, and the results show a mean accuracy of only 49%. It can be seen from these results that the classification of such nuanced argument schemes is not a straightforward task. This suggests the need for both more rigorous scheme definitions, with particular attention given to error analysis of those schemes that are commonly confused, as well as the development of guidelines taking these issues into account.

In Houngbo and Mercer (2014), a straightforward feature of co-referring text—presence of the lexeme, "this"—is used to build a self-annotating corpus extracted from a large biomedical research paper data set. This is achieved by collecting pairs of sequential sentences where the second sentence begins with "This method. . . ," "This result. . . ," or "This conclusion. . . ," and then categorizing the first sentence in each pair respectively as Method, Result, or Conclusion sentences. In order to remove outliers in the data set, a multinomial naïve Bayes classifier was trained to classify sentences from the corpus, and sentences that were classified with less than 98% confidence were removed. This reduced corpus was then used as training data to identify Method, Result, and Conclusion sentences using both SVM and naïve Bayes classifiers. These classifiers show an average F-score of 0.97 with naïve Bayes and 0.99 with SVM, and are further tested on the corpus used by Agarwal and Yu (2009), where sentences are classified in the same way. By using this approach, Houngbo and Mercer are able to improve on the results from Agarwal and Yu, whose results show an F-score of 0.92 using 10-fold cross-validation. Despite the limited nature of this task, only identifying specific types of sentences and not giving any idea of the relations between them, these results show that by extending the training data available, substantial improvements in classifying sentences can be made.

Lawrence and Reed (2017) take a similar approach to Houngbo and Mercer, using discourse indicators (connectives such as "because," "however," etc.) in place of "this." In this work, the topic of a given text is first identified and a Web search carried out to retrieve related documents. Sentences containing discourse indicators showing support relations are then found within the retrieved documents and these sentences are split on either side of the indicator to give possible premise conclusion pairs. Despite this being a noisy data set, with potential off-topic sentences and cases where the indicator has been used for a different reason, it is shown that a topic model can be built from large numbers of these pairs, resulting in stereotypical patterns of support on the given topic.

Similarly, Habernal and Gurevych (2015) use large volumes of unlabeled data from online debate portals. By identifying clusters of both sentences and posts from these debate portals that contain similar phrases, and then finding the centroids of these clusters, "prototypical arguments" are identified. Al-Khatib et al. (2016) likewise leverage online debate portals, generating annotations by automatically mapping source data, in this case the labeled text components from the idebate.org (e.g., "Introduction," "point," "counterpoint"), to a set of predefined class labels to create a large corpus with argumentative and non-argumentative text segments from several domains.

The Argument Annotated Essays Corpus (AAEC), presented in Stab and Gurevych (2014a) and updated in Stab and Gurevych (2017), consists of argument-annotated

persuasive essays, and features topic and stance identification, annotation of argument components, and argumentative relations. Drawn from 402 English language essays, the final corpus contains 751 major claims, 1,506 claims, and 3,832 premises, connected by 3,613 support and 219 attack relations. A random sample of 102 essays taken from the AAEC have been further annotated, as described in Carlile et al. (2018), to also include a persuasiveness score for each argument as well as scores for attributes that potentially impact persuasiveness (Eloquence, Specificity, Relevance, and Evidence), the means of persuasion (Ethos, Pathos, or Logos), and the types of both claims and premises. This addition to AAEC has already shown potential in developing automated persuasiveness scoring for essays (Ke et al. 2018) and, similarly, such annotations of Ethos, Pathos, or Logos as found in the AAEC have been shown to closely reflect the persuasive strength of arguments (Duthie, Budzynska, and Reed 2016; Wachsmuth et al. 2018).

Kirschner, Eckle-Kohler, and Gurevych (2015) present a corpus of 24 German language articles, which were selected from the education research domain, and annotated using a custom designed tool (DiGAT). The annotation scheme used identifies binary relations between argument components, which in this work correspond to sentences from the original texts. Four types of relation are identified: support, attack, detail, and sequence. The first two of these relations are argumentative, whereas the latter two are discourse relations similar to the sequence and background relations of Rhetorical Structure Theory (Mann and Thompson 1987). The results of annotation using this scheme are represented as graph structures, and a range of methods to determine inter annotator agreement for these structures are considered. Despite the complexity of the articles being analyzed, the results show multi-κ values up to 0.63. Although this result is fair for such a complex annotation task, several specific areas are identified that reduce agreement. Similar categories were particularly problematic; for example, in many cases disagreement was due to confusion between support and detail or support and sequence relation. Although these differences could potentially be improved by more detailed annotation guidelines, the authors argue that in many cases several correct solutions exist, with both labelings being correct.

Legal texts are the focus of Walker, Vazirova, and Sanford (2014), where a type system is developed for marking up successful and unsuccessful patterns of argument in U.S. judicial decisions. Building on a corpus of vaccine-injury compensation cases that report factfinding about causation, based on both scientific and non-scientific evidence and reasoning, patterns of reasoning are identified and used to illustrate the difficulty of developing a type or annotation system for characterizing these patterns. A further example of legal material is the ECHR corpus (Mochales and Ieven 2009), a set of documents extracted from legal texts of the European Court of Human Rights (ECHR). The ECHR material, although not annotated specifically for argumentative content, contains a standard type of reasoning and structure of argumentation that means that the corpus can be easily adapted to serve as data for argument mining.

A different domain is considered in Kiesel et al. (2015), who present a corpus of 200 newspaper editorials annotated for their argumentative structure. The annotation is based on a model consisting of explicit argumentative units, and the implicit argumentative relations (i.e., support or attack) between them. In this case, an argumentative unit is understood to be a segment of the original text containing at least one proposition. Argumentative relations are considered as the links from one unit to the unit that it most directly supports or attacks.

The Internet Argument Corpus (IAC) (Walker et al. 2012) is a corpus for research in political debate on Internet forums. It consists of approximately 11,000 discussions, 390,000 posts, and some 73,000,000 words. Subsets of the data have been annotated

for topic, stance, agreement, sarcasm, and nastiness, among others. The IAC is further developed in the IAC version 2 (Abbott et al. 2016), a collection of corpora for research in political debate on Internet forums. It consists of three data sets: 4forums (414K posts), ConvinceMe (65K posts), and a sample from CreateDebate (3K posts). It includes topic annotations, response characterizations (4forums), and stance, though argument annotation in both IAC data sets is rather limited by comparison to that available in other data sets.

Such efforts add to the volume of currently available data for which at least some elements of the argumentative structure have been identified. The most comprehensive and completely annotated existing collection of such data is the openly accessible database, AIFdb[17] (Lawrence et al. 2012), containing over 14,000 Argument Interchange Format (AIF) argument maps, with over 1.6m words and 160,000 claims in 14 different languages.[18] These numbers are growing rapidly, thanks to both the increase in analysis tools interacting directly with AIFdb and the ability to import analyses produced with the Rationale and Carneades tools (Bex et al. 2012). Indeed, AIFdb aims to provide researchers with a facility to store large quantities of argument data in a uniform way. AIFdb Web services allow data to be imported and exported in a range of formats to encourage re-use and collaboration between researchers independent of the specific tools and data format that they require.

Additionally, several online tools such as DebateGraph,[19] TruthMapping,[20] Debatepedia,[21] Agora,[22] Argunet[23] and Rationale Online[24] allow users to create and share argument analyses. Although these tools are helping to increase the volume of analyzed argumentation, they generally do not offer the ability to access this data and each use their own formats for its annotation and storage. At the moment, some research projects continue to introduce ad hoc, idiosyncratic data representation languages for argumentation and debate, which can limit reusability, integration, and longevity of the data sets.

Whereas the previously discussed data sets can be viewed as "fully" structured argument data, there is an increasing usage of larger "semi-structured" argumentative data sources. The most striking example of such are recent data sets gathered from the ChangeMyView (CMV) Reddit subcommunity[25] (Tan et al. 2016; Hidey and McKeown 2018; Musi, Ghosh, and Muresan 2018). These data take the form of discussion threads where the original poster of a thread provides a viewpoint on a specific topic, and other users reply with comments aiming to change this view. If the original poster finds that a comment succeeds in changing their viewpoint, they can reply with a 'delta' symbol indicating this. Although this data is not strictly argumentative, there are strong indicators of argumentative structure: Direct responses, for example, often include counterarguments to the original post. Indeed, Hua and Wang (2017) use CMV data to both train and evaluate a model for automatically generating arguments of the opposing stance for a given statement.

17 http://www.aifdb.org.
18 Chinese, Dutch, English, French, German, Hindi, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, and Ukrainian.
19 http://debategraph.org
20 https://www.truthmapping.com
21 http://www.debatepedia.org
22 http://agora.gatech.edu/
23 http://www.argunet.org/
24 https://www.rationaleonline.com/
25 https://www.reddit.com/r/changemyview/

In addition to these corpora of structured argument data, there are large corpora of unstructured data available that are rich in argumentative structure, from, for example, Wikipedia, Twitter, Google Books, meeting data from the AMIDA Meeting Corpus[26] annotated using the Twente Argumentation Scheme (Rienks, Heylen, and Weijden 2005), and product reviews from Web sites such as Amazon and epinions.com. Although these corpora may be useful for certain argument mining techniques, such as those using unsupervised learning methods, there are limits on their utility imposed, inevitably, by their lack of annotation.

Despite the lack of marked argument structure, Wikipedia, in particular, represents a considerable amount of data rich in argumentative content. In Aharoni et al. (2014), work toward annotating articles from Wikipedia using a meticulously monitored manual annotation process is discussed. The result is a corpus of 2,683 argument elements, collected in the context of 33 predefined controversial topics, and organized under a simple structure detailing a claim and its associated supporting evidence.

In their far-ranging work on Project Debater,[27] IBM has made extensive use of Wikipedia and other data to create the first AI system that can debate humans on complex topics. Debater can respond to a given topic by automatically constructing a set of relevant pro/con arguments phrased in natural language. For example, when asked for responses to the topic "The sale of violent video games to minors should be banned," an early prototype of Debater scanned approximately 4 million Wikipedia articles and determined the ten most relevant articles, scanned all 3,000 sentences in those articles, detected sentences that contain candidate claims, assessed their pro and con polarity, and then presented three relevant pro and con arguments,[28] with more recent developments also working toward choosing the most convincing of these arguments (Gleize et al. 2019), expanding the topic of the debate (Bar-Haim et al. 2019), and providing "first principle" debate points, commonplace arguments that are relevant to many topics, where specific data are lacking (Bilu et al. 2019). These abilities are the result of ongoing work to extract meaningful argument data from large corpora. In Levy et al. (2014), the challenge of detecting Context Dependent Claims (CDCs) in Wikipedia articles was first addressed, showing how, given a topic and a selection of relevant articles, a selection of "general, concise statements that directly support or contest the given topic" can be found. This work was followed in Rinott et al. (2015), where extracting supporting evidence from Wikipedia data for a given CDC was addressed. Bar-Haim et al. (2017) introduced the task of claim stance classification, that is, detecting the target of a given CDC, and determining the stance toward that target. Levy et al. (2017) further developed CDC identification, removing the need for pre-selected relevant articles, by first deriving a *claim sentence query* to retrieve CDCs from a large unlabeled corpus. (Indeed, this retrieval task is increasingly becoming a distinct and challenging task in its own right, with applications such as *args.me* [Wachsmuth et al. 2017] and new shared tasks such as Touche[29] driving the area forward.) Such large volumes of CDCs can be used both as potential points to be made by the Debater system as well as to aid in the interpretation of spoken material containing breaks, repetitions, or other irregularities (Lavee et al. 2019). The method introduced by Levy et al. is used in Shnarch et al. (2018) to generate **weak labeled data** (data of low quality compared to manual annotation,

26 http://corpus.amidaproject.org/.
27 https://www.research.ibm.com/artificial-intelligence/project-debater/.
28 http://www.kurzweilai.net/introducing-a-new-feature-of-ibms-watson-the-debater.
29 http://touche.webis.de.

Table 1
Significant structured argumentation data sets available online.

Name	Description	Size	IAA	URL	Reference
AIFdb Corpora					
Argumentation Schemes	Examples of occurrences of Walton's argumentation schemes found in episodes of the BBC Moral Maze Radio 4 programme.	6,704 words	Single annotator	`http://corpora.aifdb.org/schemes`	Lawrence and Reed 2016
Digging By Debating	Collection of analyses of 19th century philosophical texts from the Hathi Trust collection.	35,789 words	Single annotator	`http://corpora.aifdb.org/dbyd`	Murdock et al. 2017
Dispute Mediation	Argument maps of mediation session transcripts.	26,923 words	$\kappa = 0.68$	`http://corpora.aifdb.org/mediation`	Janier and Reed 2016
MM2012	Analyses of all episodes from the 2012 summer season of the BBC Moral Maze Radio 4 programme.	29,068 words	$\kappa = 0.55$ (types), 0.61 (relations)	`http://corpora.aifdb.org/mm2012`	Budzynska et al. 2014
US2016	2016 US presidential elections: annotations of selected excerpts of primary and general election debates, combined with annotations of selected excerpts of corresponding Reddit comments.	87,064 words	$\kappa = 0.75$	`http://corpora.aifdb.org/US2016`	Visser et al. 2018
Imported into AIFdb					
AraucariaDB	An import of 661 argument analyses produced using Araucaria and stored in the Araucaria database.	62,881 words	Single annotator	`http://corpora.aifdb.org/araucaria`	Reed 2006
AraucariaDBpl	A selection of over 50 Polish language analyses created using the Polish version of Araucaria.	2,654 words	Single annotator	`http://corpora.aifdb.org/araucariapl`	Budzynzka 2011
Argument Annotated Essays	The corpus consists of argument annotated persuasive essays, including annotations of argument components and argumentative relations.	147,271 words	$\kappa = 0.64$–0.88 (types), 0.71–0.74 (relations)	`http://corpora.aifdb.org/AAECv2`	Stab and Gurevych 2017
eRulemaking	Argument maps of 67 comment threads from `regulationroom.org`.	26,083 words	$\kappa = 0.73$	`http://corpora.aifdb.org/RRD`	Park and Cardie 2014
Internet Argument Corpus (IAC)	Consisting of 11,000 discussions and developed for research in political debate on Internet forums. Subsets of the data have been annotated for topic, stance, agreement, sarcasm, and nastiness, among others.	1,031,398 words	$\kappa = 0.22$–0.60, $\bar{\kappa} \approx 0.47$	`http://corpora.aifdb.org/IAC`	Walker et al. 2012
Language of Opposition	Used in Rutgers for the SALTS project (`http://salts.rutgers.edu/`).	48,666 words	Not reported	`http://corpora.aifdb.org/looc1`	Ghosh et al. 2014
Microtext	112 manually created, short texts with explicit argumentation, and little argumentatively irrelevant material.	7,828 words	$\kappa = 0.83$	`http://corpora.aifdb.org/Microtext`	Peldszus 2014
Available elsewhere					
Argument Annotated User-Generated Web Discourse	User comments, forum posts, blogs and newspaper articles annotated with an argument scheme based on an extended Toulmin model.	84,673 words	$\alpha_U = 0.51$–0.80	`https://bit.ly/2vdkHOD`	Habernal and Gurevych 2017
Consumer Debt Collection Practices (CDCP)	User comments about rule proposals by the Consumer Financial Protection Bureau collected from an eRulemaking website.	∼88,000 words	$\alpha = 0.65$ (types), 0.44 (relations)	`http://joonsuk.org`	Niculae, Park, and Cardie 2017
Internet Argument Corpus (IAC) 2	Corpus for research in political debate on Internet forums. It includes topic annotations, response characterizations, and stance.	∼500,000 forum posts	Not reported	`https://nlds.soe.ucsc.edu/iac2`	Abbott et al. 2016
IBM Project Debater Data sets	Collection of annotated data sets developed as part of Project Debater to facilitate this research. Organized by research sub-fields.	Various	Various	`https://ibm.co/2OlqieA`	Rinott et al. 2015, Levy et al. 2017, etc.

but which can be automatically obtained in large quantities) and then combined with a smaller quantity of high quality, manually labeled data (**strong labeled data**). Using the combined strong and weak data set resulted in improved performance for topic-dependent evidence detection, suggesting that this kind of data gathering can be a valuable asset, particularly in data-hungry neural network systems. The annotated data sets used in this and other Project Debater work are all available online.[30]

Bosc, Cabrio, and Villata (2016) address another rich online data source, taking data from Twitter and defining guidelines to detect "tweet-arguments" among a stream of tweets about a certain topic, before then pairing the identified arguments, and finally, providing a methodology to identify which kind of relation holds between the arguments composing a pair (i.e., support or attack). Bosc, Cabrio, and Villata report agreement of $\alpha = 0.81$ for detecting argumentative tweets, and $\alpha = 0.67$ for argument linking, with the resulting DART (Data set of Arguments and their Relations on Twitter) data set containing 4,000 tweets annotated as argument/not-argument with 446 support and 122 attack relations.

Two of the major issues with the data currently available are the lack of a standardized methodology for annotation, and a central location for the storage and retrieval of consistently formatted annotated material. AIFdb Corpora[31] (Lawrence and Reed 2014) aims to address these issues, leveraging the ability for material in a range of formats to be converted to AIF and imported into AIFdb and providing simple interfaces to collect and share corpora. AIFdb Corpora already collects over 7,000 of the 12,000 analyses contained in AIFdb into a range of corpora that are publicly available in perpetuity at fixed permalinks. A list of the most significant corpora is given in Table 1, including those imported and available through AIFdb as well as those elsewhere.

5. Argument Mining: Automating Argument Analysis

In the preceding sections, we have looked first at a range of different techniques that are precursors to the task of argument mining, and at the manual analysis of the argumentative structure of a text, gaining an understanding of both the nature of argumentative structure as well as the process by which a human analyst understands and extracts this structure. We have then moved on to look at argument data, considering not just the corpora of data that are available, but also the automated methods used to extend these data. In this section we now break down the argument mining task into a range of individual challenges (see Figure 3). In Sections 6, 7, and 8, we will then look at each of these tasks in more detail, drawing together work targeted at varying domains, and using different approaches, to understand the challenges and progress made in each of these areas.

For the purposes of this review, we use these tasks as a framework to present and organize the work carried out in the field. In Section 6 we look at automatic approaches for identifying argument components and determining their boundaries. In Section 7 we move on to look at the automatic identification of properties that these clauses have, and in Section 8 we look at the identification of relations from simple premise/conclusion relations to argumentation scheme instances and dialogical properties. Where a piece of work offers a large contribution to several areas, we include these in multiple sections, grouping each part of their contribution with other works addressing the same tasks

30 http://www.research.ibm.com/haifa/dept/vst/debating_data.shtml.
31 http://corpora.aifdb.org/.

individually. For each task, we consider work carried out using a broad range of techniques, including statistical and linguistic methods.

We have seen in Section 3 how the steps in manual analysis increase in complexity, from segmenting argumentative components to identifying argumentation schemes and dialogical relations. These levels are also reflected in the automation of argument analysis. In some cases it is sufficient to know merely the range of argumentative types used in order to grade student essays (Ong, Litman, and Brusilovsky 2014), to know what stance an essay takes toward a proposition in order to check that it provides appropriate evidence to back-up its stance (Persing and Ng 2015), or whether a claim is verifiable in order to flag these in online discussions (Park and Cardie 2014). However, if the goal is to reconstruct enthymemes (Razuvayevskaya and Teufel 2017) (see also the discussion of Feng and Hirst [2011] in Section 8.2) or ask critical questions about support relations, we also need to extract the nature of the argumentation schemes being used.

In Figure 3, we show how these automatic tasks are inter-related. Starting from the identification of argument components by segmenting and classifying these as part of the argument being made or not (these tasks are sometimes performed simultaneously, sometimes separated, and sometimes the latter is omitted completely), we move down through levels of increasing complexity: First, considering the role of individual clauses (both intrinsic, such as whether the clause is reported speech, and contextual such as whether the clause is the conclusion to an argument); second, considering argumentative relations from simple premise/conclusion relationships; and third, considering whether a set of clauses forms a complex argumentative relation, such as an instance

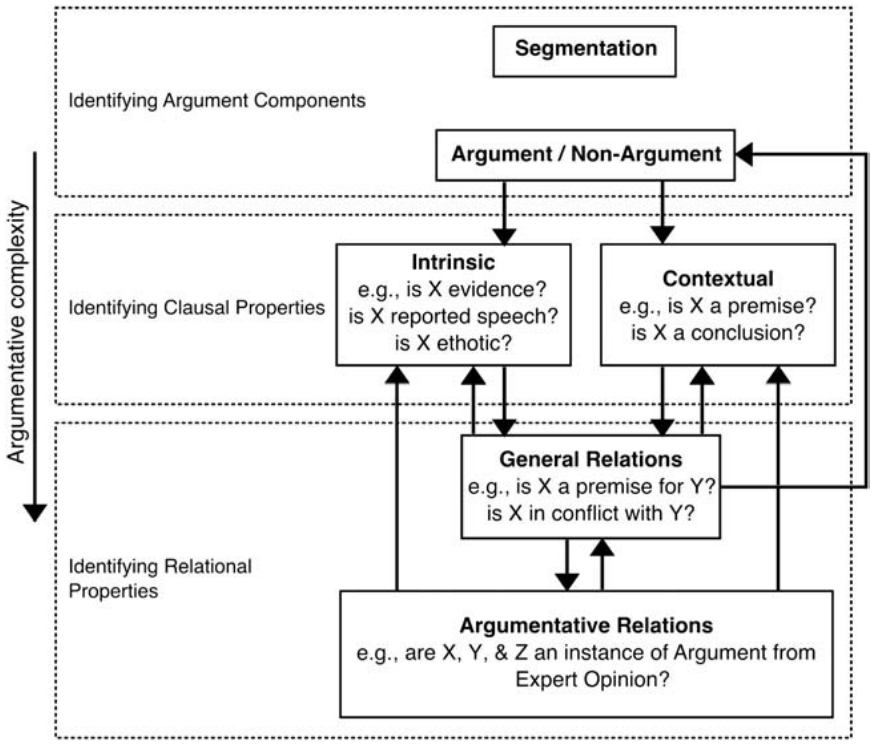


Figure 3
The tasks and levels of complexity in argument mining techniques.

of an argumentation scheme. A similar classification of argument mining tasks is given in Cabrio and Villata (2018), with Component Detection being split into the subtasks of Boundary Detection and Sentence Classification. Although this represents a robust starting point, it is also important to distinguish the types of classification (argument/non-argument and intrinsic/contextual). Cabrio and Villata also include the broad categorization of Relation Prediction, which again can be further broken down, looking at both general and argumentative relations.

The arrows shown between tasks in the figure indicate ways in which the results from one task have been used to inform the execution of another. For example, the arrow from the Argument/Non-Argument task to the Contextual Clausal Properties task reflects much early argument mining work (e.g., Moens et al. 2007) that performed these tasks in sequence; deciding which parts of the text were argumentative and then assigning a role to them. This approach has been challenged, however, with Carstens and Toni (2015) being the first to point out that whether a sentence is argumentative or not often depends on the context in which it is used; and instead advocating classifying relations first and then considering sentences to be argumentative if they have a relation connecting them (reflected in the arrow from General Relations to Argument/Non-Argument).

Similarly, some tasks can inform each other, for example, whereas Feng and Hirst (2011) showed that argument scheme instances could be classified given general relations between ADUs, Lawrence and Reed (2015) showed that such general relations can be determined by classifying argument scheme components directly from segmented text. This inter-dependency between tasks has given rise to a growth in the application of multi-objective learning approaches (e.g., Eger, Daxenberger, and Gurevych 2017; Hou and Jochim 2017; Galassi, Lippi, and Torroni 2018; Morio and Fujita 2018), where all tasks are learned and performed at the same time. These examples highlight how the simple pipeline view of argument mining, which characterizes a lot of older research work, is increasingly being superceded by more sophisticated and interconnected techniques. We will explore a further way in which argument mining tasks can be interrelated and interdependent when we consider rhetorical figures in Section 8.2.2.

Developments in argument mining are both being informed by, and informing, the related areas discussed in Section 2. For example, the work of Ong, Litman, and Brusilovsky (2014) closely parallels both argumentative zoning and citation mining, offering the opportunity to link related elements automatically identified in scientific writing, such as how a claim may be supported by a nearby citation. Rumshisky et al. (2017) look at the dynamics of social or political conflict as it develops over time, automatically identifying controversial issues where such conflict is occurring. Accuosto and Saggion (2019) show how argumentation in certain sections of a publication (in this case abstracts) can be good indicators of the quality of the work as a whole.

6. Identifying Argument Components

The automatic identification of the argumentative sections of a text corresponds to the process of argument/non-argument classification discussed in Section 3.2. Although carrying out this task in isolation does not give us a detailed picture of the argument structure, it has found use in, for example, predicting the usefulness of online reviews based solely on the *amount* of argumentative text that they contain (Passon et al. 2018).

One of the first approaches to argument mining, and perhaps still the most developed, is the work carried out by Moens et al. (Moens et al. 2007; Palau and Moens 2009; Mochales and Moens 2011), which first attempts to detect the argumentative parts of a

text by first splitting the text into sentences and then using features of these sentences to classify each as either Argument or Non-Argument. By training a range of classifiers on manually annotated examples from the Araucaria corpus (Reed 2006), an accuracy of 0.74 is obtained using a multinomial naïve Bayes classifier trained on word couples, verbs, and text statistics.

Similarly, Goudas et al. (2014) look at extracting arguments from social media, proposing a two-step approach for argument extraction similar to that used by Moens et al., first using a statistical approach through the use of machine learning and, more specifically, the logistic regression classifier to classify sentences as being part of the argument being made or not. This approach is applied to a corpus obtained from social media, concerning renewable energy sources in the Greek language; and for identifying sentences that contain arguments, an increase in performance from an F-score of 0.21, for the base case, to 0.77 is achieved. This approach is further developed in Sardianos et al. (2015), where conditional random fields are used to identify those segments from similar Greek social Web texts that contain argumentative elements.

Although these results are encouraging, it is worth noting that the classification of sentences carried out refers only to features intrinsic to the sentence and as such the classification is not robust for sentences that may be part of an argument in one context, but not in a different context. Several examples of sentences that can be viewed as argumentative in some contexts, but not in others, can be seen in Carstens and Toni (2015), who instead advocate classifying pairs of sentences according to their argumentative relation and, if the relation is classified as support or attack, considering both sentences to be argumentative. In Section 8 we look at such techniques for identifying relations, and show that Carstens and Toni's approach is in many cases preferable to the pre-identification of argumentative components.

Saint-Dizier (2018) offers an example of a situation where domain knowledge is required in order to determine whether or not a proposition is argumentative. Given the issue "Vaccine against Ebola is necessary," it is argued that the proposition "7 people died during Ebola vaccine tests" is irrelevant or neutral with respect to the issue under a knowledge-based analysis, whereas a naïve reading would rather interpret it as an attack. The importance of contextual domain knowledge highlighted by this example was first explored by Saint-Dizier (2017) where, via the analysis of various corpora, the types of knowledge that are required to develop an efficient argument mining system are explored. This exploration shows that, in about 75% of cases, some contextual knowledge is required to accurately identify arguments with respect to a controversial issue.

The idea that the context in which a text span appears can determine whether it is part of an argument or not (Opitz and Frank [2019] have shown that context can be more important than content) can be problematic for the general application of the supervised machine learning approaches discussed so far. In cases where context is not adequately captured, a model trained on one set of data can struggle to classify spans in another set of data where the context is different. As a result, rule-based and unsupervised learning approaches have also been applied to this task. The application of an unsupervised extractive summarization algorithm, TextRank, for the identification of argumentative components is explored in Petasis and Karkaletsis (2016). The motivation is to examine whether there is any potential overlap between extractive summarization and argument mining, and whether approaches used in summarization (which typically model a document as a whole) can have a positive effect on tasks of argument mining. Evaluation is performed on two corpora containing user posts from an online debating forum and persuasive essays, with results suggesting that graph-based approaches

and approaches targeting extractive summarization can have a positive effect on tasks related to argument mining.

Similarly, Wachsmuth, Stein, and Ajjour (2017) propose a model for determining the relevance of arguments using PageRank (Brin and Page 1998). In this approach, the relevance of an argument's conclusion is decided by what other arguments reuse it as a premise. These results are compared with an argument relevance benchmark data set, manually annotated by seven experts. On this data set, the PageRank scores are found to beat several intuitive baselines and correlate with human judgments of relevance.

One of the first supervised learning approaches to segmentation was introduced by Soricut and Marcu (2003) as part of the SPADE system, which also operates on lexicalized syntactic trees. The authors compute the probability of inserting a discourse boundary between a child and parent node and attained an F-score of 0.83.

The current state-of-the-art results for EDU identification are obtained by the two-pass system of Feng and Hirst (2014), who use a sequence labeling approach. Similar to Soricut and Marcu (2003), the method makes predictions over pairs of tokens that are enriched with syntactic features. Feng and Hirst showed that predicting over token pairs and making these predictions in two passes improves the results, achieving a 0.93 F-score on the recognition of in-sentence boundaries.

ADU identification, however, is considerably more challenging than identifying EDUs, requiring an understanding of the argumentative function of each span. Madnani et al. (2012) aim to separate argumentative discourse into two categories; first, argumentative text, used to express claims and evidence, and second, language used to present and organize the claims and evidence ("shell"). In the example sentence "So I think the lesson to be drawn is that we should never hesitate to use military force...to keep the American people safe," the underlined text is identified as shell. Separating shell from argumentative text is attempted using three methods: a rule-based system, a supervised probabilistic sequence model, and a principled hybrid version of the two. The rule-based system gives an F-score of 0.44, with the hybrid version giving 0.61 compared with 0.74 for a human annotator and 0.21 for a baseline that labels words as shell if they appear frequently in persuasive writing. The rule-based system uses a set of 25 hand-written regular expression patterns, for example, "I [MODAL] [ADVERB] AGREEVERB with the AUTHORNOUN." The Supervised Sequence Model is based on conditional random fields using a small number of general features based on lexical frequencies, with the intuition behind these features being that shell language generally consists of chunks of words that occur frequently in persuasive language. It is important to note that, although the material identified as shell is not a part of the argument being made, this material contains valuable information about the argument structure, often indicating the occurrence of certain speech acts, or containing discourse markers (Hutchinson 2004).

Lawrence et al. (2014) present an alternative supervised learning approach to ADU segmentation, focusing specifically on identification of ADU boundaries. Two naïve Bayes classifiers are used to perform **Proposition Boundary Learning**, one to determine the first word of a proposition and one to determine the last. The classifiers are trained using a set of manually extracted propositions as training data. The text to be segmented is first split into words and a list of features is then determined for each of these words. The features used cover both intrinsic (the word itself, its length, and POS) and contextual (the word/punctuation before and the word/punctuation after). By looking at more general features (length and POS) and contextual features, this approach aims to overcome the variability in specific words that may start (or end) a proposition.

Having trained the classifiers, this same list of features is then determined for each word in the test data, enabling the classifiers to label each word as being "start" or "end." Once the classification has taken place, the individual starts and ends are matched to determine propositions, using their calculated probabilities to resolve situations where a start is not followed by an end (i.e., where the length of the proposition text to be segmented is ambiguous). Using this method, a 32% increase in accuracy is achieved over simply segmenting the text into sentences when compared to argumentative spans identified by a manual analysis process.

Ajjour et al. (2017) also find that considering the broader context of surrounding words, or even the document as a whole, aids in locating proposition boundaries. The approach in this case is framed as a sequence labeling task, with a neural network model utilizing structural, syntactic, lexical, and pragmatic features, as well as capturing long-distance dependencies. Capturing the entire text with this model provides the best results across all domains, with F-scores of up to 0.89.

Even reliably identifying ADU segment boundaries, however, is being recognized as insufficient for identifying ADUs simply because ADUs typically express propositions with a variety of linguistic surface phenomena obfuscating that propositional content. Mood, anaphora, ellipsis, deixis, reported speech, and more all introduce new challenges for ADU identification. Jo et al. (2019) have used a combination of techniques, some statistical, some rule-based, and some hybrid, organized in a cascade structure, in order to attempt to recover the propositional structure underlying ADUs, in order to improve the performance of other argument mining tasks.

7. Automatic Identification of Clausal Properties

In the previous section we explored a range of techniques for identifying the sections of a text that are argumentative; however, this does not yet tell us anything about the nature of these argumentative text spans, or how they work together. We now move on to look at techniques for automatically identifying properties of argumentative components. In this section, we look at identifying the function of each text span, first considering intrinsic properties (e.g., whether a text span is evidence for a claim) and then look at identifying how a text span is used in the argument as a whole (e.g., premise vs. conclusion). In Section 8, we move on to look at the identification of inter-clausal relations—for example, given a pair of text spans, identifying any support or conflict relationship between them.

7.1 Intrinsic Clausal Properties

The first type of clausal properties we look at are those that are intrinsic to the clause itself. Although these properties are limited in what they tell us about the overall argumentative structure, they provide valuable information about the role that a particular text span is playing in the argument as a whole. For example, knowing that a claim is verifiable suggests a link to a piece of evidence in the text supporting this claim (Park and Cardie 2014); knowing that a clause is increasing the author's ethos suggests that it is supporting a specific claim that they are making (Duthie, Budzynska, and Reed 2016); and knowing the type of evidence provided can be used to assign different weights to statements in clinical trials (Mayer, Cabrio, and Villata 2018), or help understand rulings in disability benefits claims (Walker et al. 2018).

Verifying the acceptability of propositions used as premises in an argument is a central issue in the linguistic and philosophical study of argumentation (Freeman 2000).

In the study of persuasive communication and rhetoric, this has led to a variety of typologies of evidence. For example, Reynolds and Reynolds (2002) distinguish between statistical, testimonial, anecdotal, and analogical evidence; Hoeken and Hustinx (2003) use a revised distinction between individual examples, statistical information, causal explanations, and expert opinions; and Fahnestock and Secor (1988) utilize the classical stasis issues of fact, definition, cause, value, and action.

This diversity is also evident in the computational classification of propositions and evidence. In Park and Cardie (2014), online user comments are examined for propositions that are UNVERIFIABLE, VERIFIABLE NON-EXPERIENTIAL, or VERIFIABLE EXPERIENTIAL, with associated supports of type *reason*, *evidence*, and *optional evidence*, respectively. A proposition is considered verifiable if it contains an objective assertion with a truth value that can be proved or disproved with objective evidence. Verifiable propositions are further split into experiential or non-experiential, depending on whether or not the proposition is about the writer's personal state. For example, "My son has hypoglycemia" is tagged as Verifiable Experiential, whereas "food allergies are seen in less than 20% of the population" is marked as Verifiable Non-Experiential. Following an annotation scheme developed on 100 randomly selected comments, manual annotation inter-coder reliability is moderate, yielding an unweighted Cohen's κ of 0.73, whereas SVM classifiers trained with a range of features including *n*-grams and features specific to each class exhibit statistically significant improvement over the unigram baseline, achieving a macro F-score of 0.69. These results show that identifying propositions of these types can be achieved with reasonable accuracy, although this would still need to be developed in order to identify the relations between these propositions and determine the argument structure. By having an indication of the required support for each proposition, this structure could then be used to identify areas where a proposition is not adequately supported.

These classifications are revised in Park and Cardie (2018) to propositions of non-experiential fact (*fact*), propositions of experiential fact (*testimony*), propositions of value (*value*), propositions of policy (*policy*), and reference to a resource (*reference*). With these revised proposition categories and their associated supports of type *reason* and *evidence*, a further annotation study was carried out, resulting in the Consumer Debt Collection Practices (CDCP) corpus. This corpus consists of 731 user comments on the CDCP ruling, with 4,931 elementary units (of which the majority were propositions of value—45%), and 1,221 support relations (1,174 reason, and only 46 evidence). On this data set, Niculae (2018) achieved a maximum F1-score of 0.74 for proposition classification using linear structured SVMs.

Egawa, Morio, and Fujita (2019) adjust the annotation scheme of Park and Cardie slightly, replacing *reference* with *rhetorical statement* (which implicitly states the subjective value judgment by expressing figurative phrases, emotions, or rhetorical questions) and replacing the relations with the more standard *attack* and *support*. This scheme was then used to annotate 345 posts from the ChangeMyView sub-Reddit,[32] resulting in 4,612 proposition classifications and 2,713 relations that were then used in analyzing the semantic role of persuasive arguments.

The value of being able to identify verifiable propositions is highlighted by the classification of evidence types presented in Addawood and Bashir (2016), where Twitter posts are automatically identified as either a news media account (NEWS), blog post (BLOG), or no evidence (NO EVIDENCE). The data for this study are taken from

32 https://www.reddit.com/r/changemyview/.

tweets on the FBI and Apple encryption debate, with 3,000 tweets annotated. SVM classifiers trained with *n*-grams and other features capture the different types of evidence used in social media and demonstrate significant improvement over the unigram baseline, achieving a macro-averaged F-score of 0.83. Similarly, Dusmanu, Cabrio, and Villata (2017) look at argumentative tweets, classifying them as either fact or opinion with an F-score of 0.80 and the source of their information (e.g., CNN) with an F-score of 0.67.

The classification of factual statements for critical evaluation has gained prominence as part of fact-checking. Hassan, Li, and Tremayne (2015) classify sentences as non-factual, unimportant factual, and check-worthy factual. Similarly, Patwari, Goldwasser, and Bagchi (2017) and Jaradat et al. (2018) automatically determine the fact-check-worthiness of factual claims in political debates. Naderi and Hirst (2018a) automatically distinguish between true, false, stretch, and dodge statements in parliamentary proceedings.

Anand et al. (2011) consider a different level of intrinsic clausal properties than those discussed so far, looking not at the structural nature of propositions, but at their function. This work describes the development of a corpus of blog posts where attempts to persuade and the corresponding tactics used in this persuasion are annotated. Persuasion involves the change in mental state of the other party classed as either Belief Revision, Attitude Change, or Compliance Gaining. The methods that can be used to achieve these changes in mental state are considered in Marwell and Schmitt (1967), who offer 12 strategy types for securing behavioral compliance. A further six non-logical "principles of influence" are covered in Cialdini (2001). By combining these with argumentative patterns inspired by Walton, Reed, and Macagno (2008), and removing overlapping tactics, Anand et al. produce a list of 16 types of rhetorical tactic for persuasive acts. By using a naïve Bayes classifier for seven possible combinations of three feature sets to perform this classification, Anand et al. report a best result with an F-score of 0.58. However, rhetorical relations are often implicit and not clearly indicated in the text, and, as such, their discovery requires a richer set of features.

Duthie, Budzynska, and Reed (2016) consider another facet of persuasion, using a pipeline of techniques to extract positive and negative ethotic statements (Aristotle 1991) from parliamentary records. Although this work differs from many other argument mining approaches (which despite often looking at persuasion, nonetheless typically focus exclusively on *logos* rather than *ethos* or *pathos*), there is a clear link, with ethotic relations often following the same logotic structures, but with the character of a person as their target. In this work, those statements in which the speaker refers to another person (referred to as Ethotic Sentiment Expressions, **ESE**s) and those in which they do not (**non-ESE**s) are first extracted using a combination of named entity recognition, POS tagging, and a set of domain specific rules to locate statements referring to another person, organization, or agentive entity. These are then passed to the anaphora layer where both source-person and target-person of the statement are retrieved from the original text. Finally, a sentiment layer consisting of a sentiment classifier combined with sentiment and ethotic word lexicons classifies ESEs as positive and negative. The resulting pipeline achieves an F-score of 0.70 for ESE/non-ESE classification, compared with 0.45 for a baseline classifier that predicts only the target class (ESE); and 0.78 for +/−ESE classification, compared with a baseline of 0.67. A similar corpus of statements aimed at defending against ethotic attacks, or defending the speaker's reputation, is presented in Naderi and Hirst (2018b), and extracted from various issues in Canadian parliamentary proceedings.

In Villalba and Saint-Dizier (2012), an approach to the identification and analysis of arguments as they appear in opinion texts is developed. Examples are given that show that arguments are either incorporated into evaluative expressions with a heavy semantic load (e.g., evaluative adjectives such as 'repas familial' means a meal that has properties such as casual, home-made, good, and abundant), or composed of an evaluation and one or more discourse structures such as justification, elaboration, or illustration, whose aim is to persuade the reader of the evaluation.

For example:

- **Justification**: The hotel is 2 stars [JUSTIFICATION due to the lack of bar and restaurant facilities].
- **Reformulation**: Could be improved [REFORMULATION in other words, not so good].
- **Elaboration by Illustration or Enumeration**: The bathrooms were in a bad condition: [ILLUSTRATION the showers leaked, and the plug mechanism in the bath jammed. . .] Breakfast selection is very good [ENUMERATION with a range of cereals, tea and coffee, cold meats and cheese, fresh and canned fruit, bread, rolls and croissants, and a selection of cooked items.]
- **Elaboration via Precision**: Friendly and helpful staff [PRECISION especially the service executives at the counter.]
- **Elaboration via Comparison**: These head phones are excellent [COMPARISON as if you are in a concert room.]
- **Elaboration via Consequence**: a high soundproofing [ELAB-CONSEQUENCE that allows you to have a rest after a long working day.]
- **Contrast**: The price is very reasonable [CONTRAST but comfort is rather poor.]
- **Concession**: Very quiet [CONCESSION in spite of its downtown location in a nightlife area.]

These relations are processed using TextCoop (Saint-Dizier 2012), a platform designed for discourse analysis, with a logic and linguistic perspective. The results compared to a manual annotation on a corpus of 50 texts range between precision (0.85–0.92), and recall (0.76–0.86) over the eight relations listed above.

The Automatic Argumentative Analysis (A3) algorithm described in Pallotta and Delmonte (2011) provides an alternative approach to classifying statements according to rhetorical roles. A3 is a module developed based on the GETARUNS system (Delmonte 2007) for interaction mining (the discovery and extraction of insightful information from digital conversations, namely, those human–human information exchanges mediated by digital network technology). The module takes as input the complete semantic representation produced by GETARUNS and produces argumentative annotation using the following 20 discourse relation labels: circumstance, narration, adverse, obligation, evaluation, statement, result, hypothesis, elaboration, permission, cause, motivation, explanation, agreement, contrast, question, inception, setting, evidence, and prohibition. These labels come partly from Rhetorical Structure Theory (Mann and Thompson 1987) and partly from other theories, including those reported by Hobbs (1993) and Dahlgren (1988).

Discourse relations are automatically extracted by GETARUNS and these are then mapped onto five Meeting Description Schema (MDS) (Pallotta et al. 2004) argumentative labels: ACCEPT, REJECT/DISAGREE, PROPOSE/SUGGEST, EXPLAIN/JUSTIFY, and REQUEST. In the training stage, the system was used to process the first ten dialogues of the International Computer Science Institute meetings corpus (Janin et al. 2003) containing a total number of 98,523 words and 13,803 turns. In the test stage, two different dialogues were randomly chosen to assess the performance of the A3 algorithm; and on a total of 2,304 turns, 2,247 received an automatic argumentative classification, yielding a recall value of 0.98 (precision 0.81, F-Score 0.89).

Having labeled text segments in this way, it is easy to visualize them using, for example, conversation graphs (Ailomaa and Rajman 2009). Conversation graphs are diagrams that summarize what topics were discussed, how long they were discussed, which participants were involved in the discussion, and what type of arguments they contributed (an example conversation graph can be seen in Figure 4). Conversation graphs can be built directly by looking at the MDS labels assigned to a conversation's turns.

The benefits of using even a simple linguistic analysis to study the argumentative structure of a document are illustrated in Ong, Litman, and Brusilovsky (2014), where a series of simple rules are used to tag sentences with their role (either Current Study, Hypothesis, Claim, or Citation), for example, if the sentence contains a four-digit number, then it is tagged as Citation, if the sentence contains string prefixes from {suggest, evidence, shows, essentially, indicate}, then it is tagged as Claim. This approach again highlights the similarities between AZ (Section 2.4) and the determination of argumentative role. The ability to determine these roles offers the opportunity to link related elements, for example, a Claim may be backed by a nearby Citation.

Wyner et al. (2012) also use simple linguistic cues, in this case to support manual analysis by providing a rule-based tool for supporting textual analysis by semi-automatic identification of argumentative sections in the text. The tool is aimed specifically at online product reviews, and highlights potential argumentative text in the review according to discourse indicators (explicitly stated linguistic expressions of the relationship between statements [Webber, Egg, and Kordoni 2011]) and terminology specific to the domain (e.g., product names and their properties). The tool uses a set of discourse indicators, sentiment terminology, a user model, and a domain model. Discourse indicators are used to locate premises (after, as, because, for, since, when, assuming,...), conclusions (therefore, in conclusion, consequently,...), and contrast (but, except, not, never, no,...), whereas sentiment terminology signals lexical semantic contrast. A comprehensive list of terms is classified according to a scale of sentiment

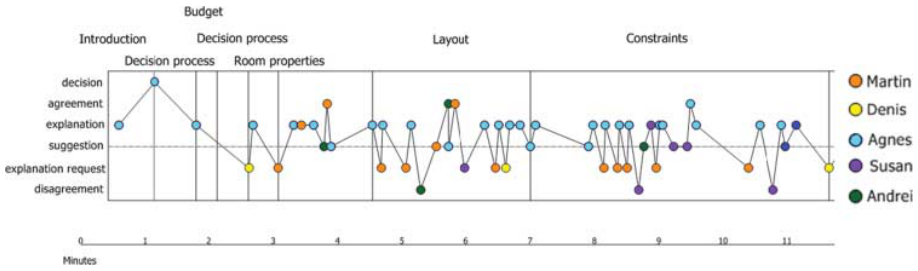


Figure 4
Conversation graph from Ailomaa and Rajman (2009).

ranging from highly negative to highly positive. The user model covers properties of user performing the review and, finally, the domain model specifies the objects and properties that are relevant to the users, for example, properties with binary values (such as has a flash), properties with ranges (such as the number of megapixels, scope of the zoom, or lens size), and multi-slotted properties (such as the warranty).

Wyner further develops the concept of using argument mining as a way to assist manual analysis in Wyner, Peters, and Price (2015), which describes the development of Argument Workbench, a tool designed to help the analyst reconstruct arguments from textual sources by highlighting a range of discourse indicators, topics used in the text, domain terminology, and speech act terminology. The tool integrates with the DebateGraph software,[33] to allow the user to produce detailed argument graphs.

7.2 Contextual Clausal Properties

Having considered the argumentative properties intrinsic to a text span, we now move on to look at identifying how a text span is used in the argument as a whole.

The work of Moens et al. (2007) on classifying sentences as "argument" or "non-argument" is further developed in Palau and Moens (2009), where an additional machine learning technique was implemented to classify each argument sentence as either premise or conclusion, a method referred to as argument proposition classification. In this case, the examples considered are extended using material from the ECHR; and accuracy of classifying sentences as argument increases to 0.80 using the ECHR corpus. Argument proposition classification is carried out using a maximum entropy model and support vector machine, with F-scores of 0.68 for classification as premise and 0.74 for classification as conclusion. Again, this work inherits the shortcomings of the earlier research, as the same sentence can be a premise in one context and a conclusion in another.

Such contextual restrictions can, however, also be an advantage, allowing, for example, comments on an article to be related to the original article based on their relation to it. For example, the work of the IBM Debater project in *context dependent evidence detection*, which automatically detects evidence in Wikipedia articles supporting a given claim (Rinott et al. 2015).

Though it is obviously an oversimplification, it is also possible to reduce the complexity of the task of recognizing the stance of evidence toward claim into a binary classification.[34] This is the motivation behind the Same Side Stance shared task,[35] in which examples are tuples of $topic, argument_1, argument_2$, and the classification task one of determining whether or not two arguments on the same of a binary debate.

In Boltužić and Šnajder (2014) **argument-based opinion mining** is used to determine the arguments on which the users base their opinions. This builds upon previous work in opinion mining (as discussed in Section 2.1), to include not just the general opinion or stance toward a given topic, but also the arguments on which that stance is based. This is carried out on a specially created corpus of user comments, manually annotated with arguments, using a classifier to predict the correct label from the set of five possible labels (as shown in Table 2). The model uses textual entailment and

33 `debategraph.org`.

34 There are many classes of examples that do not fit the binary model well—situations, such as elections, with more than two candidates; political configurations in which factions within parties express extreme positions, etc.

35 `https://sameside.webis.de`.

Table 2
Labels for comment-argument pairs (Boltužić and Šnajder 2014).

Label	Description: Comment...
A	...explicitly **attacks** the argument
a	...vaguely/implicitly **attacks** the argument
N	...makes **no use** of the argument
s	...vaguely/implicitly **supports** the argument
S	...explicitly **supports** the argument

semantic textual similarity features with the best models outperforming the baselines and giving a 0.71 to 0.82 micro-averaged F-score. Although these results give a promising indication of the ability to determine how a comment relates to the argument being made, the topics studied are limited and the training data taken from `procon.org` and `idebate.org` would not extend to general topics.

The ability to identify even such basic contextual properties offers the opportunity to inform the user and aid in both writing and understanding text. This is again illustrated in Stab and Gurevych (2014b), who aim to identify argument in essays and works toward the long-term goal of integrating argumentation classifiers into writing environments. Two classifiers are described. First, for identifying argument components, a multiclass classification is carried out with each clause classified as major claim, claim, premise, or non-argumentative. This classifier is trained on a range of feature types, structural features (for example the location and punctuation of the argument component), lexical features (*n*-grams, verbs, adverbs, and modals), syntactic features, discourse indicators, and contextual features. Once the argument components have been identified, a second classifier is used to identify argumentative relations (support or non-support). The features used are similar to those for classifying the components, but look at the pairings of clauses. The presented approach achieves 88.1% of human performance for identifying argument components and 90.5% for identifying argumentative relations.

This work is further developed in Nguyen and Litman (2015), where the same methodology and data set are used, but a Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) topic model is first generated to separate argument and domain keywords. The output from the LDA algorithm is then post-processed using a minimal seeding of predefined argumentative words to determine argument and domain topics. The same features as Stab and Gurevych (2014b) are then used, replacing *n*-grams with unigrams of argument words, and numbers of argument and domain words. Using this updated feature set, the accuracy is improved for all of the argument component types: MajorClaim (from 0.48 to 0.59), Claim (from 0.49 to 0.56), and Premise (from 0.86 to 0.88). Although these results are promising, the relatively low numbers still highlight the difficulties in distinguishing between Claim and MajorClaim, due to the largely context dependent distinction between the two.

The categories from another theory of argumentation structure due to Toulmin (1958), of Data Claim and Warrant, are similarly difficult to distinguish. Indeed, the theoretical impossibility of completely acontextual identification was explored from first principles by Freeman (1991), who showed that under the appropriate circumstances, the difference between Data and Warrant dissolves. With appropriate context, however, the distinction becomes operationally important and was the driver for the first shard task in argument mining, conducted at SEMEVAL2018 by Habernal et al. (2018). The Argument Reasoning Comprehension Task required systems to use a given

premise and conclusion to distinguish between two given alternative potential warrants (there is further contextual information available too, with explicitly identified topic and background). For example:

> **Topic:** There She Is, Miss America
>
> **Additional info:** In 1968, feminists gathered in Atlantic City to protest the Miss America pageant, calling it racist and sexist. Is this beauty contest bad for women?)
>
> **Argument:** Miss America gives honors and education scholarships. And since . . . , Miss America is good for women.
>
> a) scholarships would give women a chance to study
> b) scholarships would take women from the home

The system should in this example choose option (a). Human performance (following brief training) on this task is at 0.91; system performance in the task varied, with a variety of techniques performing at between 0.50 and 0.70 F-score. Although these results seem extremely encouraging, Niven and Kao (2019) suggest that this result is entirely accounted for by exploitation of spurious statistical cues in the data set, and that by eliminating the major source of these cues, the maximum performance fell from just 3 points below the average untrained human baseline to essentially random. Niven and Kao counter these effects by the addition of adversarial examples, obtained by negating the claim and inverting the label for each datapoint.

Although the goal of argument mining is the extraction of argumentative structure from natural text, the availability of large quantities of appropriately annotated training data makes this challenging to carry out. An alternative starting point is presented in Peldszus (2014), where a corpus of "microtexts," short texts with explicit argumentation and little argumentatively irrelevant material, is created. The representation of the argument structure within these microtexts is based on Freeman's theory of argumentation structure (Freeman 1991, 2011), and is viewed as a hypothetical dialectical exchange between a proponent, who presents and defends their claims, and an opponent, who critically questions them. These moves can then be represented as an argument graph, with the nodes representing the propositions expressed in text segments and the edges between them representing different supporting and attacking moves. An agreement between untrained annotators is presented in Peldszus and Stede (2013b). The annotators achieved moderate agreement for certain aspects of the argument graph (e.g., $\kappa = 0.52$ in distinguishing proponent and opponent segments, or $\kappa = 0.58$ in distinguishing supporting and attacking segments) yet only a marginal agreement of $\kappa = 0.38$ on the full label set describing all aspects of the argument graph. A further study using expert annotators produced significantly higher agreement ($\kappa = 0.83$) on the full label set.

The annotation process assigns a list of labels to each segment based on different levels. The "role"-level specifies the dialectical role (proponent or opponent). The 'typegen'-level specifies the general type, namely, whether the segment presents the central claim (thesis) of the text, or supports/attacks another segment. The "type"-level additionally specifies the kind of support (normal or example) and the kind of attack (rebutter or undercutter). Peldszus tests a range of classifiers to automatically classify role, typegen, and type. The results show that an SVM classifier generally performs best on the most complex labels, suggesting that it deals well with the lower frequencies with which these occur. Meanwhile, the maximum entropy and naïve Bayes classifiers perform best on the simpler and more common labels.

The results on the microtext corpus are encouraging, but the artificial nature of its construction means that such results may not generalize well to unrestricted text. However, this corpus does provide a valuable resource for controlled "laboratory" testing of argument mining techniques.

8. Automatic Identification of Relational Properties

In this section we move on from looking at the identification of clausal properties to the identification of inter-clausal relations. We look first at general argumentative relations, for example, premise/conclusion relationships, and then move on to look at the more complex relationships involved in argumentation schemes and dialogical relations.

8.1 Identifying General Argumentative Relations

Identifying relations between pairs of propositions is a more complex and nuanced task than identifying the roles that an individual proposition may take. It is one thing to know, for example, that a given proposition is a premise; much more challenging is to determine also for which conclusion (or conclusions) it serves as premise. Approaches to identifying these relations either build upon the prior classification of individual clauses, or aim to extract relations directly.

Palau and Moens (2009) build upon their classification of each argument sentence as either premise, or conclusion using a context-free grammar, produced by grouping manually derived rules. This context-free grammar is used to determine the internal structure of each individual argument. The accuracy of classifying sentences as argument or non-argument is 0.80 and we find F-scores of 0.68 and 0.74 for classification as premise and conclusion, respectively; for the harder task of determining argument structure, however, the accuracy achieved is 0.60.

Peldszus (2014) also builds on the initial task of identifying roles of segments in the Microtext corpus by adding a "combined"-level, showing, for all types, whether a segment's function holds only in combination with that of another segment (combined) or not (simple). The target is specified by a position relative identifier with a numerical offset identifying the targeted segment relative from the position of the current segment. The prefix "n" states that the proposition of the node itself is the target, and the prefix "r" states that the relation coming from the node is the target. Again the results for identifying the target of a relation (maximum F-score of 0.45) are lower than for identifying the roles (maximum F-score of 0.85).

This same microtext corpus is used in Peldszus and Stede (2015), who look at identifying conflict relations by examining the texts for occurrences of counter-considerations (e.g., "Even though. . . ," or "It has been claimed that. . . however. . . "), which the author uses to introduce a potential criticism of their argument, before going on to address the issue and so strengthen their point. This identification is carried out by labeling the textual segments as either "proponent" or "opponent" using a linear log-loss model, resulting in an F-score of 0.64 for identifying opposition relations between segments.

Although the work discussed thus far in this section builds upon previous identification of component roles before identifying relations, Cabrio and Villata (2012) propose an approach to detect arguments and discover their relationships directly by building on existing work in textual entailment (TE) (Dagan, Glickman, and Magnini 2006). TE refers to a "directional relation between two textual fragments, termed *text (T)* and *hypothesis (H)*, respectively." The relation holds whenever the truth of one text fragment follows from another. In this case, the T-H pair is a pair of arguments expressed by two

different users in a dialogue on a certain topic and the TE system returns a judgment (entailment or contradiction) on the argument pair.

A data set of 300 T-H pairs is created using manually selected topics from Debatepedia,[36] which provides pre-annotated arguments (pro or con), and following the criteria defined and by the organizers of the Recognizing Textual Entailment challenge.[37] Of these 300 T-H pairs, 200 are used to train (100 entailment and 100 contradiction) and 100 to test (50 entailment and 50 contradiction). The pairs collected for the test set concern completely new topics, never seen by the system, and are provided in their unlabeled form as input.

TE recognition is carried out using EDITS (Edit Distance Textual Entailment Suite).[38] EDITS implements a distance-based framework that assumes that the probability of an entailment relation between a given T-H pair is inversely proportional to the distance between T and H. The system uses different approaches to distance computation, providing both edit distance algorithms (cost of the edit operations [insert, delete, etc.] to transform T into H) and similarity algorithms. Each algorithm returns a normalized distance score between 0 and 1. During training, distance scores are used to calculate a threshold that separates entailment from contradiction. Of the EDITS configurations that Cabrio and Villata tested, the highest accuracy is obtained using either Word Overlap or Cosine Similarity (0.66 in both cases), with Token Edit Distance performing significantly less well (accuracy= 0.53), suggesting that semantic similarity plays a more important role than syntactic similarity (a result backed up by the comparative analysis of Aker et al. [2017], who also found syntactic features to be the least informative in all of the experimental settings considered). Although these numbers are quite low, this is an interesting result, suggesting that the relationship between topics in an argument gives more of a clue as to how the components relate than does the way in which those components are expressed. This is carried through in several later works that look at relations between topics and semantic similarity between propositions.

Nguyen and Litman (2016) argue that looking at the content of such pairings to determine relationships does not make full use of the information available. They propose an approach that makes use of contextual features extracted from surrounding sentences of source and target components as well as from general topic information. Experimental results show that using both general topic information and features of surrounding sentences are effective, but that predicting an argumentative relation will benefit most from combining these two sets of features.

The machine learning approaches to argument mining discussed so far in this section have all used supervised learning to perform classification; however, unsupervised learning has also been applied to the task. In Lawrence et al. (2014), a LDA topic model is used to determine the topical similarity of consecutive propositions in a piece of text. The intuition is that if a proposition is similar to its predecessor then there exists some argumentative link between them, whereas if there is low similarity between a proposition and its predecessor, the author is going back to address a previously made point and, in this case, the proposition is compared to all those preceding it to determine whether they should be connected. This assumes that the argument is built up as a tree structure in a depth-first manner, where an individual point is pursued fully before returning to address the previous issues. Although the assumption of a tree structure does not hold for all arguments, it is the case for around 95% of the argument

36 http://www.debatepedia.org.
37 http://www.nist.gov/tac/2010/RTE/.
38 http://edits.fbk.eu/.

analyses contained in AIFdb and 80% of arguments in the CDCP corpus, as reported by Niculae, Park, and Cardie (2017). No evidence is given by Niculae et al. supporting the hypothesis of topical relations with manual analysis of the data, but the automated results do support the hypothesis, with a precision of 0.72 and recall of 0.77 recorded when comparing the resulting structure to a manual analysis. It should also be noted that what is being identified here is merely that an inference relationship exists between two propositions, with no indication of the directionality of this inference.

This same approach is implemented in Lawrence and Reed (2015), where the use of LDA topic models is replaced by using WordNet[39] to determine the semantic similarity between propositions. This change is required to overcome the difficulties in generating a topic model when the text being considered is only a short span, such as an online comment or blog post. The results are comparable to those achieved using LDA, with precision of 0.82 and recall of 0.56. In this case the thresholds are adjusted to increase precision at the expense of recall, as the output from this method is combined with a range of other approaches to determine the final structure, and as such the failure of this approach to identify all of the connections can be compensated for by the other techniques.

A similar approach of assuming a relationship between argument components, if they refer to the same concepts or entities, is used by AFAlpha (Carstens, Toni, and Evripidou 2014), which represents customer reviews as trees of arguments, where a child–parent relationship between two sentences is determined if they refer to the same concepts, with the child being the sentence that has been posted later. A sentence is represented as a set of features, including its semantic characteristics such as metadata about the review in which the sentence appears, as well as features based on the sentence's syntactic and lexical nature such as occurrences of certain words and phrase types. A feature vector thus represents each pair of sentences and is classified using a model trained on a data set comprising data taken from the Q&A debating platform, Quaestio-it,[40] and IMDB.[41]

Carstens and Toni continue this line of work in Carstens and Toni (2015), focusing on the determination of argumentative relations, and foregoing the decision on whether an isolated piece of text is an argument or not. This focus is based on the observation that the relation to other text is exactly what describes the argumentative function of a particular text span. The paper mentions a number of use cases, describing a method of evaluating claims, by giving a gauge of what proportion of a text argues for or against them. Additionally a preliminary corpus of 854 annotated sentence pairs[42] is provided, with each sentence pair labeled with $L \in \{A, S, N\}$, where A = Attack, S = Support, or N = Neither (including both cases where the two sentences are unrelated and those where they are related, but not in an argumentative manner).

The important role played by similarity is also exploited by Gemechu and Reed (2019), who borrow notions of aspect, target concept, and opinion from opinion mining, and use these to decompose ADUs down into finer-grained components, and then use similarity measures between these components to identify argument relations. Such **decompositional argument mining** not only performs well on diverse single-author arguments (outperforming the techniques of Peldszus and Stede on their Microtext corpus, and of Stab and Gurevych on their AAEC corpus) but also on arguments

39 http://wordnet.princeton.edu/.
40 http://www.quaestio-it.com.
41 http://www.imdb.com.
42 Available at www.doc.ic.ac.uk/Ëœlc1310/.

situated in dialogue (albeit at lower levels of performance: F1 ranging from 0.74 to 0.77 on both Microtext and AAEC, and 0.63 on US2016).

Finally, Wachsmuth, Syed, and Stein (2018) highlight an interesting link between similarity and argumentative relations. The work presented aims to determine the best counterargument to any argument without prior knowledge of the argument's topic. The best performing model tested rewards a high overall similarity between a potential counterargument and the given argument's conclusion and premises, while punishing those counterarguments that are too similar to either of them. To some extent, this result captures the intuition that argumentative relations occur where something different is being said about the same topic.

8.2 Identifying Complex Argumentative Relations

The ability to successfully extract premises and conclusions is built upon in Feng and Hirst (2011), which presents the first step in the long-term goal of a method to reconstruct enthymemes, by first, classifying to an argumentation scheme (Walton, Reed, and Macagno 2008), then fitting the propositions to the template, and finally, inferring the enthymemes. For the first step of fitting one of the top five most commonly occurring argumentation schemes to a predetermined argument structure, accuracies of 0.63–0.91 are recorded in one-against-others classification and 0.80–0.94 in pairwise classification. As in Moens et al. (2007), the Araucaria corpus is used with complex Argument Units (AUs) first broken into simple AUs (with no embedded AUs). The AUs using the top five most common argumentation schemes are then selected, and a classifier is trained on both features specific to each individual scheme and a range of general linguistic features in order to obtain the scheme. Although these results are promising, and suggest that identifying scheme instances is an achievable task, they do rely on the prior identification of premises and conclusions, as well as the basic structure that they represent. Although this approach does not identify the roles of individual propositions in the scheme, knowing what type of scheme links a set of propositions is both a useful task in its own right and offers potential for subsequent processing to determine proposition types for each scheme component. This is a substantially easier task once the scheme type is known.

Another approach to identifying the occurrence of schemes is given in Lawrence and Reed (2015), where, rather than considering features of the schemes as a whole, the individual scheme components are identified and then grouped together into a scheme instance. In this case, only two schemes (Expert Opinion and Positive Consequences) are considered and classifiers trained to identify their individual component premises and conclusion. By considering the features of the individual types of these components, F-scores between 0.75 and 0.93 are given for identifying at least one component part of a scheme.

The approach followed by Feng and Hirst (2011) is similar in nature to the first steps suggested by Walton (2011), where a six-stage approach to identifying arguments and their schemes is proposed. The first of these stages is the identification of the arguments occurring in a piece of text; this is followed by identification of specific known argumentation schemes. Walton, however, points out that beyond this initial identification there are likely to be issues differentiating between similar schemes and suggests the development of a corpus of borderline cases to address the issue.

As Walton points out, the automatic identification of argumentation schemes remains a major challenge. As discussed in Section 3.4, a large number of scheme classifications exist, with additional domain specific schemes utilized in specific areas. For example, as part of the rule-based tool for semi-automatic identification of argumentative sections

in text presented in Wyner et al. (2012), a consumer argumentation scheme (Figure 5) is described and the structure of this scheme used to guide the argument identification process.

Similarly, Green (2015) lists ten custom argumentation schemes targeted at genetics research articles. For example, one of the schemes presented, Failed to Observe Effect of Hypothesized Cause, looks for situations where specific properties were not observed, and where it is assumed that a specific condition that would result in those properties is present, leading to the conclusion that the condition may not be present. Green (2018a) further argues for schemes expressed in terms of domain concepts rather than by generic definitions as in those of Walton, Reed, and Macagno (2008), carrying out a pilot annotation study of schemes for 15 arguments in the Results/Discussion section of biological/biomedical journal articles. Green (2018b) then explores how argumentation schemes in this domain can be implemented as logic programs in Prolog and used to extract individual arguments. In this case, the schemes are formulated in terms of semantic predicates obtained from a text by use of BioNLP (biomedical/biological natural language processing) tools.

Regardless of the theoretical backdrop, schemes generally introduce as much complexity as they do opportunity from annotation through to automated analysis. To pick an example from a substantially different theoretical approach, Musi, Ghosh, and Muresan (2016) present a novel set of guidelines for the annotation of argument schemes based on the Argumentum Model of Topics (Rigotti and Morasso 2010). This framework offers a hierarchical taxonomy of argument schemes based on linguistic criteria that are distinctive and applicable to a broad range of contexts, aiming to overcome the challenges in annotating a broad range of schemes.

With the data currently available, the ontologically rich information available in argumentation schemes has been demonstrated to be a powerful component of a robust approach to argument mining. Collaboration among analysts as well as the further development of tools supporting argumentation schemes is essential to growing the data sets required to improve on these techniques. Clear annotation guidelines and the development of custom argumentation schemes for specific domains will hopefully result in a rapid growth in the material available and further increase the effectiveness of schematic classification.

8.2.1 Dialogical Relations. Whereas some of the previously mentioned argument mining techniques have worked with data that is dialogical in nature, such as user comments and online discussion forums, none of these have focused on using the unique features of dialogue to aid in the automatic analysis process, producing an analysis that captures both the argumentative and dialogical structure. For example, although Pallotta et al. (2004) and Rienks, Heylen, and Weijden (2005) consider dialogical data, in both cases they do not consider the specific dialogical relations between utterances.

Similarly, there is a large body of work studying the nature of dialogue both in terms of dialogue modeling, which captures the nature and rules of a dialogue, and dialogue management, which takes a more participant-oriented viewpoint in determining what dialogical moves to make (Traum 2017). However, there is currently little work that puts these models to work enhancing argument mining techniques. It seems clear that

```
Premise: Camera X has property P
Premise: Property P promotes value V for agent A
Conclusion: Agent A should Action1 camera X
```

Figure 5
Consumer argumentation scheme.

by modeling a dialogue and understanding that the next move a participant is likely to make will be "disagreeing," for example, we would be able to obtain the argumentative structure easily. In this section we discuss formalizations of dialogue protocols and then move on to cover the work that has been done to apply this knowledge to argument mining.

In the case of more formally structured dialogues, a protocol for the dialogue can be described, and specified in a language such as the FIPA Agent Coordination Language (McBurney and Parsons 2009), the Dialogue Game Description Language (Bex, Lawrence, and Reed 2014), or the Lightweight Coordination Calculus (Robertson 2004). Such dialogue games have been developed to capture a range of more structured conversations, for example, to facilitate the generation of mathematical proofs (Pease et al. 2017) or help reach agreement on which course of action to take in specific circumstances (Atkinson, Bench-Capon, and McBurney 2005). In these cases, software such as Arvina (Lawrence, Bex, and Reed 2012) or D-BAS (Krauthoff et al. 2018) can be used to both run the dialogue according to the specified rules and automatically capture the argumentative structure generated as the dialogue progresses. These structures can then be used to allow for mixed initiative argumentation (Snaith, Lawrence, and Reed 2010), where a combination of human users and software agents representing the arguments made by other people can take part in the same conversation, using retrieval-based methods to select the most relevant response (Le, Nguyen, and Nguyen 2018). In such scenarios, the contributions of human participants can be interpreted by virtue of their dialogical connections to the discourse, allowing a small step toward mining argument structure from natural language.

Although formally structured dialogues can be captured and exploited in this way, many real world dialogues follow only very limited rules and the challenge of identifying the argumentative structure in free form discussion is complex. However, even very informal dialogues nevertheless provide additional data beyond that available in monologue, which can be used to help constrain the task.

Among other such features, Budzynska et al. (2014) identify illocutionary forces and dialogue transitions. Illocutionary forces are the speech act type of utterances. Their automatic recognition in Illocutionary Structure Parsing (Budzynska et al. 2016) is similar to Dialogue Act Annotation (Bunt et al. 2010), though often more specific. Automatic distinction between rhetorical, pure, and "assertive" questioning, for example, is nuanced and challenging. The preliminary results reported in Budzynska et al. (2016) point to accuracy of 78% on this task, but the data sets used are very small ($n = 153$).

Al Khatib et al. (2018) identify six distinct "discourse acts" (Socializing, Providing evidence, Enhancing the understanding, Recommending an act, Asking a question, and Finalizing the discussion) in deliberative discussions. As a first step toward determining the best possible move for a participant in a deliberative discussion, Al Khatib et al. train an SVM model to classify examples of these discourse acts from Wikipedia data. Although the classifier achieves low F-scores for Socializing, Recommending an act, and Asking a question, these are the categories with the smallest number of examples in the data set to draw from—83, 137, and 106 turns, respectively. Performance on those acts with more examples is much better: Providing evidence (781 turns, F-score = 0.69), Enhancing the understanding (671 turns, F-score = 0.58), and Finalizing the discussion (622 turns, F-score = 0.71). These results are encouraging and suggest that with more data, further improvements could be expected.

Dialogue transitions, on the other hand, connect together dialogical moves. In Inference Anchoring Theory (Budzynska et al. 2014), illocutionary connections are anchored in these transitions. This explicit connectivity can be used to handle complex phenomena such as indexicality (where the propositional content of one locution can only be

reconstructed by reference to another locution, for example: "Isn't that a source of injustice?"—"Definitely not"). Budzynska et al. suggest that the patterns provided by transitions can constrain the mining process by defining expectations (for example, if an assertive question is followed by a negative polarity indexical assertion, then such a transition anchors the illocutionary connection of disagreeing). There are no results yet reported testing this hypothesis.

8.2.2 Rhetorical Figures. In much the same way that argumentation schemes capture common patterns of reasoning, **rhetorical figures** capture common patterns of speech. Although not as implicitly related to argumentative structure as argument schemes, rhetorical figures and argumentation are closely linked. Fahnestock (1999) makes a compelling case for the conception of rhetorical figures as couplings of linguistic form and function. Drawing on a tradition that links figures to *topoi*, running back to Aristotle (Aristotle 1991), Fahnestock argues that figures "map function onto form or perfectly epitomize certain patterns of thought or argument" (page 26). She demonstrates this claim for a specific group of figures related to organization. To the extent that the claim is true—that there is, in Fahnestock's terms, a "figural logic" at work in language—the potential for argument mining and other computational explorations of language is promising. Just as study in rhetoric has emphasized the connection to argumentation, similarly there is an emergence of work in argument mining that is considering rhetorical moves. Alliheedi, Mercer, and Cohen (2019), for example, aim to develop a framework to analyze argumentation structure in biochemistry procedures by developing an automated rhetorical move analysis platform.

Harris et al. (2018) argue for the importance of rhetorical figures for argument mining in particular, and present an annotation scheme to make figure detection more tractable for computational approaches. It is claimed in this work that many figures are formal patterns that algorithms can detect through surface analysis, illustrating this with an example from John F. Kennedy's 1961 United States presidential inaugural address: "Ask not what your country can do for you. Ask what you can do for your country." This constitutes an instance of the figure *antimetabole*, the repetition of words in reverse order, and is relatively easy to identify with a straightforward lexical or rule-based approach. For computational purposes, patterns of form are much easier to detect than conceptual ones (Gawryjolek, Di Marco, and Harris 2009; Dubremetz and Nivre 2017). For example, a figure like *polyptoton* (repetition of stems with different affixes: "hate the sin but not the sinner") is easier to algorithmically detect than a trope like *metaphor* (a cross-domain mapping: "Juliet is the sun").

The first work to directly connect rhetorical figure detection to argument mining appears in Lawrence, Visser, and Reed (2017), where the connection between eight rhetorical figures, the forms of which are relatively easy to identify computationally, and their corresponding argumentation structure is explored. For example, instances of **epizeuxis** (the repetition of a word or small group of words, with no other words in between, such as "very, very" or "many, many") are shown to often be attacking a previous point and yet, perhaps due to their vehement nature, attract little conflict themselves. Although instances of epizeuxis can provide information about a particular proposition, figures such as **eutrepismus** (the numbering and ordering of parts under consideration) give indication of structure, commonly being used to provide a number of premises for a conclusion ("X because, first Y, second Z,…").

Such instances of figures complement multiple levels of argument mining tasks, reinforcing the move away from a traditional pipeline to a more holistic approach. An instance of epizeuxis informs: segmentation (the text spans between each numbering

almost certainly correspond to ADUs); premise/conclusion classification (these ADUs are almost always premises); and relation identification (the premises are likely connected to a conclusion preceding them). Although the work on connecting rhetorical figures to argument structure is still at an early stage, it is an example of a technique that works on multiple argumentative levels, complementing existing, more focused approaches.

The fusion of multiple techniques at multiple levels is an increasingly common theme in argument mining, evidenced for example by a task proposed for SEMEVAL 2020 focusing on the detection of "propaganda techniques" in news articles. The task requires identification of a wide range of phenomena, including logical fallacies,[43] techniques appealing to emotions, loaded language, and more (San Martino et al. 2019).

9. Conclusion

The recent rapid growth in argument mining shows that there is an increasing demand for the automated extraction of deeper meaning from the vast amounts of data that we currently produce. Although techniques in opinion mining are able to tell us *what* people are thinking, we also need to be able to say *why* they hold those opinions. There is substantial commercial opportunity here as businesses increasingly want to build on the data that they gather in order to know more about the thoughts and behaviors of their customers, and it is unsurprising that many of the large players in the field are engaging, most visibly to date, IBM.

One of the first challenges faced by argument mining is the lack of consistently annotated argument data. Much recent work has focused on producing annotation guidelines targeted at specific domains (e.g., Kirscgner, Eckle-Kohler, and Gurevych 2015; Walker, Vazirova, and Sanford 2014; Kiesel et al. 2015), and although this has shown that data from these fields can be consistently annotated, the use of specific annotation schemes aimed at individual areas means that any techniques developed using these data are limited to that domain. The volume of data, particularly data annotated at the most fine grained level, is still far below what would be required to apply many of the techniques previously discussed in a domain independent manner. Attempts are being made to overcome this lack of data, including the use of crowdsourced annotation (Ghosh et al. 2014; Skeppstedt, Peldszus, and Stede 2018) and automatic methods to extend the data currently annotated (Bilu, Hershcovich, and Slonim 2015). As these efforts combine with increasing attention to manual analysis, the volume of data available should increase rapidly. Schulz et al. (2018) also offer some solace in this regard, showing how multi-task learning (training models across data sets from different domains) can improve results in domains where limited domain specific annotated data is available.

Even in cases where there is a greater volume of data, conflicting notions of argument are often problematic. In a qualitative analysis of six different, widely used, argument data sets, Daxenberger et al. (2017) show that each data set appears to conceptualize claims quite differently. These results clearly highlight the need for greater effort in building a *framework* in which argument mining tasks are carried out, covering all aspects from agreement on the argument theoretical concepts being identified, through to uniform presentation of results and data.

43 Of course, referring to logical fallacies as propaganda techniques is highly controversial, not least because the boundary between fallacies and schemes is such a fine one (Walton 1996). Hamblin (1970) and Groarke, Tindale, and Fisher (1997) represent a good introduction to the literature on fallacies.

A related problem is verifiability and reproducibility of results: As a young field, argument mining does not yet benefit from uniformly publicly available algorithms and codebases that would encourage incremental advance. Initiatives such as CLARIN (Krauwer and Hinrichs 2014) and LAPPSGrid (Ide et al. 2015) are trying to tackle this challenge across NLP, and argument technologies might be expected to contribute to, and benefit from, these initiatives in much the same way that other specialities within NLP can.

Beyond these logistical and theoretical issues, there also remains the fact that argument mining is a difficult task; as Moens (2018) points out, "a lot of content is not expressed explicitly but resides in the mind of communicator and audience." It seems that to overcome this challenge we need to look at the broader picture in which argument occurs. In this regard, works that either take a more holistic "end-to-end" view (Stab and Gurevych 2017; Persing and Ng 2016; Potash, Romanov, and Rumshisky 2017) or that aim to harness external data sources (Rinott et al. 2015; Lawrence and Reed 2017) seem to point the way.

Argument mining techniques have been successfully developed to extract details of the argumentative structure expressed within a piece of text, focusing on different levels of argumentative complexity as the domain and task require. For each task, we have considered work carried out using a broad range of techniques, including statistical and linguistic methods. We have presented a hierarchy of task types based on increasing argumentative complexity. First looking at the identification of argument components and the determination of their boundaries, we have then moved on to consider the role of individual clauses (both intrinsic, such as whether the clause is reported speech, and contextual, such as whether the clause is the conclusion to an argument). Finally, we have considered the identification of a range of argumentative relations from simple premise/conclusion relationships, to whether a set of clauses form an instance of an argumentation scheme.

The success of these techniques and the development of techniques for analyzing dialogical argument offers hope that techniques can be developed for automatically identifying complex illocutionary structures and the argumentative structures they build. We have also seen how these techniques can be combined, tying together statistical identification of basic structure and linguistic markers and identifying scheme components. In so doing, the resulting argument structures offer a more complete analysis of the text than any of these methods provide on their own.

Argument mining remains profoundly challenging, and traditional methods on their own seem to need to be complemented by stronger, knowledge-driven analysis and processing. However, the pieces required to successfully automate the process of turning unstructured data into structured argument are starting to take shape. As the volume of analyzed argument continues to increase, and existing techniques are further developed and brought together, rapid progress can be expected.

Acknowledgments

This work was funded in part by EPSRC in the UK under grant EP/N014871/1.

References

Abbott, Rob, Brian Ecker, Pranav Anand, and Marilyn A. Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 4445–4452, Portoroz.

Accuosto, Pablo 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.

Addawood, Aseel and Masooda Bashir. 2016. What is your evidence? A study of controversial topics in social media. In *Proceedings of the 3rd Workshop on Argumentation Mining*, pages 1–11, Berlin.

Agarwal, Shashank and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Aharoni, Ehud, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, MD.

Ailomaa, Marita and Martin Rajman. 2009. Enhancing natural language search in meeting data with visual meeting overviews. In *Proceedings of the 10th Annual Conference of the NZ ACM Special Interest Group on Human-Computer Interaction (CHINZ 2009)*, pages 6–7, Auckland.

Ajjour, Yamen, 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.

Aker, Ahmet, 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.

Al Khatib, Khalid, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling deliberative argumentation strategies on wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555, Melbourne.

Al-Khatib, Khalid, Henning Wachsmuth, Matthias Stein, Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of NAACL-HLT*, pages 1395–1404, San Diego, CA.

Alliheedi, Mohammed, Robert E. Mercer, and Robin Cohen. 2019. Annotation of rhetorical moves in biochemistry articles. In *Proceedings of the 6th Workshop on Argument Mining*, pages 113–123, Florence.

Anand, Pranav, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Doug Oard, and Philip Resnik. 2011. Believe me—we can do this! Annotating persuasive acts in blog text. In *Proceedings of the 11th International Workshop on Computational Models of Natural Argument (CMNA 2011) at AAAI 2011*, pages 11–15, San Francisco, CA.

Aristotle. 1958. *Topics*. Oxford University Press.

Aristotle. 1991. *On Rhetoric*. Oxford University Press.

Athar, Awais. 2011. Sentiment analysis of citations using sentence structure-based features. In *HLT-SS '11 Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR.

Atkinson, Katie, Trevor Bench-Capon, and Peter Mcburney. 2005. A dialogue game protocol for multi-agent argument over proposals for action. *Autonomous Agents and Multi-Agent Systems*, 11(2):153–171.

Awadallah, Rawia, Maya Ramanath, and Gerhard Weikum. 2012. Harmony and dissonance: Organizing the people's voices on political controversies. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pages 523–532, Seattle, WA.

Bar-Haim, Roy, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. 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*, Volume 1, pages 251–261, Valencia.

Bar-Haim, Roy, Dalia Krieger, Orith Toledo-Ronen, Lilach Edelstein, Yonatan Bilus, Alon Halfon, Yoav Katz, Amir Menczel, Ranit Aharonov, and Noam Slonim. 2019. From surrogacy to adoption; from bitcoin to cryptocurrency: Debate topic expansion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 977–990, Florence.

Bex, Floris, Thomas F. Gordon, John Lawrence, and Chris Reed. 2012. Interchanging arguments between Carneades and AIF – Theory and practice. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 390–397, Vienna.

Bex, Floris, John Lawrence, and Chris Reed. 2014. Generalising argument dialogue with the dialogue game execution platform. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 141–152, Pitlochry.

Bex, Floris, John Lawrence, Mark Snaith, and Chris Reed. 2013. Implementing the argument Web. *Communications of the ACM*, 56(10):66–73.

Bilu, Yonatan, 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.

Bilu, Yonatan, Daniel Hershcovich, and Noam Slonim. 2015. Automatic claim negation: Why, how and when. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 84–93, Denver, CO.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boltužić, Filip and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, MD.

Boltužić, Filip and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, Denver, CO.

Bosc, Tom, Elena Cabrio, and Serena Villata. 2016. DART: A dataset of arguments and their relations on Twitter. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, pages 1258–1263, Portoroz.

Brin, Sergey and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Budzynska, Katarzyna. 2011. Araucaria-PL: Software for teaching argumentation theory. In *Proceedings of the Third International Congress on Tools for Teaching Logic (TICTTL 2011)*, pages 30–37, Salamanca.

Budzynska, Katarzyna, Mathilde Janier, Chris Reed, and Patrick Saint-Dizier. 2016. Theoretical foundations for illocutionary structure parsing. *Argument & Computation*, 7(1):91–108.

Budzynska, Katarzyna, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. A model for processing illocutionary structures and argumentation in debates. In *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference (LREC)*, pages 917–924, Reyjavik.

Budzynska, Katarzyna and Serena Villata. 2017. Processing argumentation in natural language texts. In Baroni, Pietro, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre, editors, *Handbook of Formal Argumentation*, College Publications.

Bunt, Harry, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an ISO standard for dialogue act annotation. In *Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta.

Cabrio, Elena and Serena Villata. 2012. Generating abstract arguments: A natural language approach. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 454–461, Vienna.

Cabrio, Elena and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433, Stockholm.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carlile, Winston, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne.

Carstens, Lucas and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO.

Carstens, Lucas, Francesca Toni, and Valentinos Evripidou. 2014. Argument mining and social debates. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 451–452, Pitlochry.

Choi, Yoonjung, Yuchul Jung, and Sung-Hyon Myaeng. 2010. Identifying controversial issues and their sub-topics in news articles. In H. Chen, M. Chau, S. Li, S. Urs, S. Srinivasa, and G. A. Wang, editors, *Intelligence and Security Informatics*. Springer, pages 140–153.

Cialdini, Robert B. 2001. *Influence: Science and Practice*, 4. Allyn and Bacon, Boston, MA.

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

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. In The PASCAL

recognising textual entailment challenge. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. dAlche Buc, editors, *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*. Springer, pages 177–190.

Dahlgren, Kathleen. 1988. *Naive Semantics for Natural Language Understanding*. Springer.

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, volume 35, pages 1–16, Bangkok.

Daxenberger, Johannes, 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.

De Marneffe, Marie Catherine and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *COLING 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester.

Delmonte, Roldolfo. 2007. *Computational Linguistic Text Processing: Logical Form, Semantic Interpretation, Discourse Relations and Question Answering*. Nova Publishers.

Dori-Hacohen, Shiri and James Allan. 2013. Detecting controversy on the Web. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 1845–1848, San Francisco, CA.

Dubremetz, Marie and Joakim Nivre. 2017. Machine learning for rhetorical figure detection: More chiasmus with less annotation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 37–45, Gothenburg.

Dusmanu, Mihai, 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.

Duthie, Rory, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *Proceedings of the Sixth International Conference on Computational Models of Argument (COMMA 2016)*, pages 299–310, Berlin.

Egawa, Ryo, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence.

Eger, Steffen, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of ACL*, pages 11–22, Vancouver.

Fahnestock, J. 1999. *Rhetorical Figures in Science*. Oxford University Press.

Fahnestock, Jeanne and Marie Secor. 1988. The stases in scientific and literary argument. *Written Communication*, 5(4):427–443.

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

Feng, Vanessa Wei and Graeme Hirst. 2014. Two-pass discourse segmentation with pairing and global features. *CoRR*, abs/1407.8215.

Freeman, James B. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*, volume 10. Walter de Gruyter.

Freeman, James B. 2000. What types of statements are there? *Argumentation*, 14(2):135–157.

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

Galassi, Andrea, 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.

Gawryjolek, Jakub, Chrysanne Di Marco, and Randy A. Harris. 2009. An annotation tool for automatically detecting rhetorical figures system demonstration. In *Proceedings of the IJCAI-09 Workshop on Computational Models of Natural Argument*, Pasadena, CA.

Gemechu, Debela 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.

Ghosh, Debanjan, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on*

Argumentation Mining, pages 39–48, Baltimore, MD.
Givón, Talmy. 1983. *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, volume 3. John Benjamins Publishing.
Gleize, Martin, 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.
Gordon, Thomas F., Henry Prakken, and Douglas Walton. 2007. The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10):875–896.
Goudas, Theodosis, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media, In *Artificial Intelligence: Methods and Applications*, Springer, pages 287–299.
Green, Nancy. 2014. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Workshop on Argumentation Mining*, pages 11–18, Baltimore, MD.
Green, Nancy. 2015. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21, Denver, CO.
Green, Nancy. 2018a. Proposed method for annotation of scientific arguments in terms of semantic relations and argument schemes. In *Proceedings of the 5th Workshop on Argument Mining*, Brussels.
Green, Nancy L. 2018b. Towards mining scientific discourse using argumentation schemes. *Argument & Computation*, 9(2):121–135.
Grennan, Wayne. 1997. *Informal Logic: Issues and Techniques*. McGill-Queen's Press-MQUP.
Grimes, Joseph Evans. 1975. *The Thread of Discourse*, volume 207. Walter de Gruyter.
Groarke, Leo, Christopher Tindale, and Linda Fisher. 1997. *Good Reasoning Matters!: A Constructive Approach to Critical Thinking*. Oxford University Press, Toronto.
Grosse, Kathrin, Carlos Iván Chesñevar, and Ana Gabriela Maguitman. 2012. An argument-based approach to mining opinions from Twitter. In *First International Conference on Agreement Technologies (AT 2012)*, pages 408–422, Dubrovnik.
Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
Habernal, Ivan and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated Web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2137, Lisbon.
Habernal, Ivan and Iryna Gurevych. 2017. Argumentation mining in user-generated Web discourse. *Computational Linguistics*, 43(1):125–179.
Habernal, Ivan, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. SemEval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 763–772, New Orleans, LA.
Hamblin, C. L. 1970. *Fallacies*. Methuen, London.
Harrell, Maralee. 2005. Using argument diagramming software in the classroom. *Teaching Philosophy*, 28(2):163–177.
Harris, Randy Allen, Chrysanne Di Marco, Sebastian Ruan, and Cliff O'Reilly. 2018. An annotation scheme for rhetorical figures. *Argument & Computation*, 9(2):155–175.
Hassan, Naeemul, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, CIKM '15, pages 1835–1838, Melbourne.
Hastings, Arthur C. 1963. A Reformulation of the Modes of Reasoning in Argumentation. Ph.D. thesis, Northwestern University.
Hidey, Christopher and Kathleen McKeown. 2018. Persuasive influence detection: The role of argument sequencing. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA.
Hirschberg, Julia and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.
Hitchcock, David. 1985. Enthymematic arguments. *Informal Logic*, 7(2):289–98.
Hobbs, Jerry R. 1993. Intention, information, and structure in discourse: A first draft. In *Burning Issues in Discourse, NATO Advanced Research Workshop*, pages 41–66, Maratea.
Hoeken, Hans and Lettica Hustinx. 2003. The relative persuasiveness of different types of evidence. In *Proceedings of the Fifth Conference of the International Society for the*

Study of Argumentation, pages 497–501, Amsterdam.

Hogenboom, Alexander, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska De Jong. 2010. Mining economic sentiment using argumentation structures. In Trujillo J. et al., editor, *Advances in Conceptual Modeling–Applications and Challenges*, Springer, pages 200–209.

Holmes, Geoffrey, Andrew Donkin, and Ian H. Witten. 1994. WEKA: A machine learning workbench. In *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems*, pages 357–361, Brisbane.

Hou, Yufang 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.

Houngbo, Hospice and Robert Mercer. 2014. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 19–23, Baltimore, MD.

Hua, Xinyu and Lu Wang. 2017. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 219–230, Vancouver.

Hutchinson, Ben. 2004. Acquiring the meaning of discourse markers. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, page 684–691, Barcelona.

Ide, Nancy, James Pustejovsky, Christopher Cieri, Eric Nyberg, Denise DiPersio, Chunqi Shi, Keith Suderman, Marc Verhagen, Di Wang, and Jonathan Wright. 2015. The language application grid. In *International Workshop on Worldwide Language Service Infrastructure*, pages 51–70, Kyoto.

Janier, Mathilde and Chris Reed. 2016. Corpus resources for dispute mediation discourse. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, pages 1014–1021, Portoroz.

Janin, Adam, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, volume 1, pages 364–367, Hong Kong.

Jaradat, Israa, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. Claimrank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, LA.

Jo, Yohan, 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.

Katzav, Joel and Chris Reed. 2004. On argumentation schemes and the natural classification of arguments. *Argumentation*, 18(2):239–259.

Ke, Zixuan, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4130–4136, Stockholm.

Kienpointner, Manfred. 1992. *Alltagslogik: struktur und funktion von argumentationsmustern*. Frommann-Holzboog.

Kiesel, Johannes, Khalid Al Khatib, Matthias Hagen, and Benno Stein. 2015. A shared task on argumentation mining in newspaper editorials. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 35–38, Denver, CO.

Kim, Soo Min and Eduard Hovy. 2006a. Automatic identification of pro and con reasons in online reviews. In *Proceedings of COLING/ACL 2006*, pages 483–490, Sydney.

Kim, Soo Min and Eduard Hovy. 2006b. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney.

Kirschner, Christian, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO.

Kirschner, Paul A., Simon J. Buckingham-Shum, and Chad S. Carr. 2003. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-making*. Springer.

Kittur, Aniket, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: Conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems, pages 453–462, San Jose, CA.
Krauthoff, Tobias, Christian Meter, Gregor Betz, Michael Baurmann, and Martin Mauve. 2018. D-BAS: A dialog-based online argumentation system. In *Computational Models of Argument (COMMA)*, pages 325–336, Warsaw.
Krauwer, Steven and Erhard Hinrichs. 2014. The Clarin research infrastructure: Resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531, Reykjavik.
Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 3:159–174.
Lasnik, Howard and Juan Uriagereka. 1988. *A course in GB Syntax: Lectures on Binding and Empty Categories*. MIT Press, Cambridge, MA.
Lauscher, Anne, Goran Glavaš, and Kai Eckert. 2018. Arguminsci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 22–28, Brussels.
Lauscher, Anne, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels.
Lavee, Tamar, Matan Orbach, Lili Kotlerman, Yoav Kantor, Shai Gretz, Lena Dankin, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2019. Towards effective rebuttal: Listening comprehension using corpus-wide claim mining. In *Proceedings of the 6th Workshop on Argument Mining*, pages 58–66, Florence.
Lawrence, John, Floris Bex, and Chris Reed. 2012. Dialogues on the argument Web: Mixed initiative argumentation with Arvina. In *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA 2012)*, pages 513–514, Vienna.
Lawrence, John, Floris Bex, Chris Reed, and Mark Snaith. 2012. AIFdb: Infrastructure for the argument Web. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 515–516, Vienna.
Lawrence, John and Chris Reed. 2014. AIFdb corpora. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 465–466, Pitlochry.
Lawrence, John and Chris Reed. 2015. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO.
Lawrence, John and Chris Reed. 2016. Argument mining using argumentation scheme structures. In *Proceedings of the Sixth International Conference on Computational Models of Argument (COMMA 2016)*, pages 379–390, Potsdam.
Lawrence, John and Chris Reed. 2017. 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.
Lawrence, John, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, Baltimore, MD.
Lawrence, John, Mark Snaith, Barbara Konat, Katarzyna Budzynska, and Chris Reed. 2017. Debating technology for dialogical argument: Sensemaking, engagement, and analytics. *ACM Transactions on Internet Technology (TOIT)*, 17(3):25.
Lawrence, John, Jacky Visser, and Chris Reed. 2017. Harnessing rhetorical figures for argument mining. *Argument & Computation*, 8(3):289–310.
Lawrence, John, Jacky Visser, and Chris Reed. 2019. An online annotation assistant for argument schemes. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 100–107, Florence.
Le, Dieu Thu, 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.
Levy, Ran, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 1489–1500, Dublin.
Levy, Ran, 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.
Lewis, David D. 1998. In Naive (Bayes) at forty: The independence assumption in

information retrieval. In *Machine Learning: ECML-98*, Springer, pages 4–15.
Lindahl, Anna, Lars Borin, and Jacobo Rouces. 2019. Towards assessing argumentation annotation—A first step. In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186, Florence.
Lippi, Marco and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.
Liu, Bing. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2:627–666.
Madnani, Nitin, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28, Montreal.
Mann, William C. and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*. University of Southern California, Information Sciences Institute.
Marwell, Gerald and David R. Schmitt. 1967. Dimensions of compliance-gaining behavior: An empirical analysis. *Sociometry*, 30(4):350–364.
Mayer, Tobias, Elena Cabrio, and Serena Villata. 2018. Evidence type classification in randomized controlled trials. In *Proceedings of the 5th Workshop on Argument Mining*, pages 29–34, Brussels.
McBurney, Peter and Simon Parsons. 2009. Dialogue games for agent argumentation. In G. Simari and I. Rahwan, editors, *Argumentation in Artificial Intelligence*, Springer, pages 261–280.
Merity, Stephen, Tara Murphy, and James R. Curran. 2009. Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 19–26, Suntec.
Metzinger, Thomas. 1999. Teaching philosophy with argumentation maps: Review of Can Computers Think? The debate by Robert E. Horn. *PSYCHE*, 5.
Mochales, Raquel and Aagje Ieven. 2009. Creating an argumentation corpus: Do theories apply to real arguments? A case study on the legal argumentation of the ECHR. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 21–30, Barcelona.
Mochales, Raquel and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19:1–22.
Moens, Marie Francine. 2018. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument & Computation*, 9(1):1–14.
Moens, Marie Francine, Erik Boiy, Raquel M. Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 225–230, Stanford, CA.
Morio, Gaku 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.
Murdock, Jaimie, Colin Allen, Katy Borner, Robert Light, Simon McAlister, Andrew Ravenscroft, Robert Rose, Doori Rose, Jun Otsuka, David Bourget, John Lawrence, and Chris Reed. 2017. Multi-level computational methods for interdisciplinary research in the HathiTrust digital library. *PLOS ONE*, 12(9):1–21.
Musi, Elena, Debanjan Ghosh, and Smaranda Muresan. 2016. Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the 3rd Workshop on Argumentation Mining*, pages 82–93, Berlin.
Musi, Elena, Debanjan Ghosh, and Smaranda Muresan. 2018. ChangeMyView through concessions: Do concessions increase persuasion? *Dialogue & Discourse*, 9(1):107–127.
Naderi, Nona and Graeme Hirst. 2018a. Automated fact-checking of claims in argumentative parliamentary debates. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65, Brussels.
Naderi, Nona and Graeme Hirst. 2018b. Using context to identify the language of face-saving. In *Proceedings of the 5th Workshop on Argument Mining*, pages 111–120, Brussels.
Nguyen, Huy and Diane Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO.
Nguyen, Huy V. and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual*

Meeting of the Association for Computational Linguistics, Berlin.
Niculae, Vlad. 2018. *Learning Deep Models with Linguistically-Inspired Structure*. Ph.D. thesis, Cornell University.
Niculae, Vlad, 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.
Niven, Timothy 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.
Okada, Alexandra, Simon J. Buckingham Shum, and Tony Sherborne. 2008. *Knowledge Cartography: Software Tools and Mapping Techniques*. Springer.
Ong, Nathan, 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, MD.
Opitz, Juri 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.
Palau, Raquel M. and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107, Barcelona.
Pallotta, Vincenzo and Rodolfo Delmonte. 2011. Automatic argumentative analysis for interaction mining. *Argument & Computation*, 2(2–3):77–106.
Pallotta, Vincenzo, Hatem Ghorbel, Afzal Ballim, Agnes Lisowska, and Stéphane Marchand-Maillet. 2004. Towards meeting information systems. In *Proceedings of the 6th International Conference on Enterprise Information Systems (ICEIS)*, pages 464–469, Porto.
Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86, Pennsylvania, PA.
Pantel, Patrick Andre. 2003. *Clustering by Committee*. Ph.D. thesis, University of Alberta.
Park, Joonsuk and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, MD.
Park, Joonsuk and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1623–1628, Miyazaki.
Passon, Marco, 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.
Patwari, Ayush, Dan Goldwasser, and Saurabh Bagchi. 2017. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, CIKM '17, pages 2259–2262, Singapore.
Pease, Alison, John Lawrence, Katarzyna Budzynska, Joseph Corneli, and Chris Reed. 2017. Lakatos-style collaborative mathematics through dialectical, structured and abstract argumentation. *Artificial Intelligence*, 246(Supplement C):181–219.
Peldszus, Andreas. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97, Baltimore, MD.
Peldszus, Andreas and Manfred Stede. 2013a. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.
Peldszus, Andreas and Manfred Stede. 2013b. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia.
Peldszus, Andreas and Manfred Stede. 2015. Towards detecting counter-considerations in text. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 104–109, Denver, CO.
Perelman, Chaïm and Lucie Olbrechts-Tyteca. 1969. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press.

Persing, Isaac and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of ACL*, pages 543–552, Beijing.
Persing, Isaac and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of NAACL-HLT*, pages 1384–1394, San Diego, CA.
Petasis, Georgios and Vangelis Karkaletsis. 2016. Identifying argument components through textrank. In *Proceedings of the 3rd Workshop on Argumentation Mining*, pages 94–102 Berlin.
Piao, Scott, Sophia Ananiadou, Yoshimasa Tsuruoka, Yutaka Sasaki, and John McNaught. 2007. Mining opinion polarity relations of citations. In *International Workshop on Computational Semantics (IWCS)*, pages 366–371, Tilburg.
Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12(5):601–638.
Pollock, John. 1986. *Contemporary Theories of Knowledge*. Rowman And Littlefield, Towota, NJ.
Pollock, John L. 1987. Defeasible reasoning. *Cognitive Science*, 11(4):481–518.
Pollock, John L. 1995. *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press.
Potash, Peter, Alexey Romanov, and Anna Rumshisky. 2017. 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.
Prakken, Henry. 2005. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15(6):1009–1040.
Razuvayevskaya, Olesya and Simone Teufel. 2017. Finding enthymemes in real-world texts: A feasibility study. *Argument & Computation*, 8(2):113–129.
Reed, Chris. 2006. Preliminary results from an argument corpus. In Eloína Miyares Bermúdez, and Leonel Ruiz Miyares, editors, *Linguistics in the Twenty-first Century*, Cambridge Scholars Press, pages 185–196.
Reed, Chris and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(4):961–980.
Reynolds, Rodney A. and J. Lynn Reynolds. 2002. Evidence. In James Price Dillard and Michael Pfau, editors, *The Persuasion Handbook: Developments in Theory and Practice*, Sage, Thousand Oaks, CA, pages 427–444.
Rienks, Rutger, Dirk Heylen, and van der E. Weijden. 2005. Argument diagramming of meeting conversations. In *Multimodal Multiparty Meeting Processing, Workshop at the 7th International Conference on Multimodal Interfaces*, pages 85–92, Trento.
Rigotti, Eddo and Sara Greco Morasso. 2010. Comparing the argumentum model of topics to other contemporary approaches to argument schemes: The procedural and material components. *Argumentation*, 24(4):489–512.
Rinott, Ruty, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence—an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon.
Robertson, David. 2004. A lightweight coordination calculus for agent systems. In *International Workshop on Declarative Agent Languages and Technologies*, pages 183–197, New York, NY.
Rumshisky, Anna, Mikhail Gronas, Peter Potash, Mikhail Dubov, Alexey Romanov, Saurabh Kulshreshtha, and Alex Gribov. 2017. Combining network and language indicators for tracking conflict intensity. In *International Conference on Social Informatics*, pages 391–404, Oxford, UK.
Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.
Saint-Dizier, Patrick. 2012. Processing natural language arguments with the <TextCoop> platform. *Argument & Computation*, 3(1):49–82.
Saint-Dizier, Patrick. 2017. Knowledge-driven argument mining based on the Qualia structure. *Argument & Computation*, 8(2):193–210.
Saint-Dizier, Patrick. 2018. A two-level approach to generate synthetic argumentation reports. *Argument & Computation*, 9(2):137–154.
San Martino, Giovanni Da, Alberto Barron-Cedeno, Preslav Nakov, Henning Wachsmuth, and Rostislav Petrov. 2019. SemEval-2020 task 11: Detecting propaganda techniques in news articles.
Sardianos, Christos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction

from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO.

Scheuer, Oliver, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5(1):43–102.

Schulz, Claudia, 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, LA.

Shnarch, Eyal, 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*, pages 599–605, Melbourne.

Skeppstedt, Maria, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155–163, Brussels.

Snaith, Mark, John Lawrence, and Chris Reed. 2010. Mixed initiative argument in public deliberation. In *From e-Participation to Online Deliberation, Proceedings of the Fourth International Conference on Online Deliberation (OD2010)*, pages 2–13, Leeds, UK.

Sobhani, Parinaz, 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.

Soricut, Radu and Daniel Marcu. 2003. Sentence-level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 149–156, Edmonton.

Stab, Christian and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 1501–1510, Dublin.

Stab, Christian and Iryna Gurevych. 2014b. 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.

Stab, Christian and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Stede, Manfred and Jodi Schneider. 2018. *Argumentation Mining*: Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Tan, Chenhao, 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*, pages 613–624, Montreal.

Teufel, Simone, Jean Carletta, and Marie-Francine Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen.

Teufel, Simone, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-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, Singapore.

Teufel, Simone, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney.

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

Traum, David. 2017. Computational approaches to dialogue. In Edda Weigand, editor, *The Routledge Handbook of Language and Dialogue*. Taylor & Francis, pages 143–161.

van Eemeren, Frans H., Bart Garssen, Eric C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Handbook of Argumentation Theory*. Springer.

van Gelder, Tim. 2007. The rationale for rationale. *Law, Probability and Risk*, 6(1–4):23–42.

van Rijsbergen, Cornelis Joost. 1979. *Information Retrieval*. Butterworth.
Villalba, Maria Paz G. and Patrick Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 23–34, Vienna.
Visser, Jacky, Rory Duthie, John Lawrence, and Chris Reed. 2018. Intertextual correspondence for integrating corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3511–3517, Miyazaki.
Wachsmuth, Henning, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen.
Wachsmuth, Henning, 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, NM.
Wachsmuth, Henning, Benno Stein, and Yamen Ajjour. 2017. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 1117–1127, Valencia.
Wachsmuth, Henning, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1, pages 241–251, Melbourne.
Walker, Marilyn A., Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 812–817, Istanbul.
Walker, Vern, Karina Vazirova, and Cass Sanford. 2014. Annotating patterns of reasoning about medical theories of causation in vaccine cases: Toward a type system for arguments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 1–10, Baltimore, MD.
Walker, Vern R., Dina Foerster, Julia Monica Ponce, and Matthew Rosen. 2018. Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment. In *Proceedings of the 5th Workshop on Argument Mining*, pages 68–78, Brussels.
Walton, Douglas. 1996. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ.
Walton, Douglas. 2011. Argument mining by applying argumentation schemes. *Studies in Logic*, 4(1):38–64.
Walton, Douglas, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.
Webber, Bonnie, Markus Egg, and Valia Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.
Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver.
Wyner, Adam, Wim Peters, and David Price. 2015. Argument discovery and extraction with the argument workbench. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 78–83, Denver, CO.
Wyner Adam, Jodi Schneider, Katie Atkinson, and Trevor Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 43–50, Vienna.