# Investigating the Role of Argumentation in the Rhetorical Analysis of Scientific Publications with Neural Multi-Task Learning Models

**Anne Lauscher,**[1,2] **Goran Glavaš,**[1] **Simone Paolo Ponzetto,**[1] and **Kai Eckert**[2]
[1]Data and Web Science Research Group
University of Mannheim, Germany
[2]Web-based Information Systems and Services
Stuttgart Media University, Germany
{anne, goran, simone}@informatik.uni-mannheim.de
{lauscher, eckert}@hdm-stuttgart.de

## Abstract

Exponential growth in the number of scientific publications yields the need for effective automatic analysis of rhetorical aspects of scientific writing. Acknowledging the argumentative nature of scientific text, in this work we investigate the link between the argumentative structure of scientific publications and rhetorical aspects such as discourse categories or citation contexts. To this end, we (1) augment a corpus of scientific publications annotated with four layers of rhetoric annotations with argumentation annotations and (2) investigate neural multi-task learning architectures combining argument extraction with a set of rhetorical classification tasks. By coupling rhetorical classifiers with the extraction of argumentative components in a joint multi-task learning setting, we obtain significant performance gains for different rhetorical analysis tasks.

## 1 Introduction

Scientific publications, as "tools of persuasion" in research (Gilbert, 1977), are carefully composed documents written to convince the reader of the validity and merit of the researchers' work. As such, they are inherently argumentative and often adhere to well-trodden rhetorical patterns and argumentation schemes of the respective research field. The accelerated growth of scientific literature (Bornmann and Mutz, 2015) makes exploration and analysis of relevant publications increasingly difficult. This yields the need for automatic analyses of these documents, including their argumentative and rhetorical structure.

Accordingly, computational models already support publication analysis tasks, e.g., classification of citation purpose and polarity (Jha et al., 2017; Lauscher et al., 2017b, *inter alia*) and classification of (sentential) discourse roles (Teufel et al., 1999; Liakata et al., 2010, *inter alia*). Further, rhetorical predictions at the (sub-)sentence level obtained using these models have been shown useful in higher-level downstream tasks such as publication classification (Teufel et al., 1999), (extractive) publication summarization (Cohan and Goharian, 2015), and research trend prediction (McKeown et al., 2016).

To allow for the holistic analysis of scientific publications with respect to the interactions between different rhetorical aspects of scientific text Fisas et al. (2016) created a corpus of scientific publications with manual annotations of several high-level rhetorical aspects of scientific writing (e.g., sentence-level discourse roles), but without annotations of the argumentative structure of publications. Despite (1) scientific texts being inherently argumentative (Gilbert, 1976), (2) the existence of theoretical argumentative frameworks (Toulmin, 2003; Kirschner et al., 2015), and (3) a wide range of argument extraction models in other domains (e.g., debates or essays, see Palau and Moens (2009); Habernal and Gurevych (2017), *inter alia*), there is still very little work on automatic argumentation mining from scientific literature. Consequently, there has been no work analyzing associations between argumentation and other rhetorical constructs in scientific writing, although such dependencies exist. Consider the following example:

> *"In general, our OMR preserves the high frequency content of the motion quite well [claim], since inverse rate control is directed by Jacobian values [data]."*

Here, the authors make a *claim* (underlined text) about their approach and support it with a technical fact (*data*) about the method (wave-underlined text). At the same time, regarding other rhetorical constructs, this sentence is stating the subjective aspect of *advantage* (of the proposed method), belongs to the discourse category of *outcome* (of the

authors' work), and may be considered *relevant* for the (extractive) summary of the publication. We argue that these rhetorical dimensions are interconnected and that fine-grained argumentation underpins other rhetorical layers in scientific text. For example, sentences stating an *advantage* of a method are likely to be argumentative and may contain *claims* that should be included in the summary.

Assuming that argumentation guides rhetorics in scientific text, we investigate neural multi-task learning (MTL) models which couple argument extraction with several other rhetorical analysis tasks. To this end, we augment the existing corpus of scientific publications (Fisas et al., 2016), containing several layers of rhetorical annotations, with an additional layer of argumentative components and relations. We then explore two neural MTL architectures based on shared recurrent encoders, intra-sentence attention, and private task-specific classifiers and couple the neural architectures with a joint MTL objective with uncertainty-based weighting of task-specific losses (Kendall et al., 2018). We validate our approach by testing that it outperforms traditional machine learning models in single-task settings. We finally show that coupling rhetorical analysis tasks with argument extraction using MTL models significantly improves the results for the rhetorical analysis tasks.

**Contributions.** We create the first corpus of scientific publications in English annotated with fine-grained argumentative structures and carry out the first study on dependencies between different rhetorical dimensions in scientific writing. Using MTL models, we show that argumentation informs other rhetorical analysis tasks. Finally, in the context of MTL research, our results indicate that the dynamic uncertainty-based loss weighting (Kendall et al., 2018) is beneficial for high-level natural language processing tasks.

## 2 Related Work

We provide an overview of (1) studies analyzing rhetorical aspects in scientific publications and (2) a large body of work on argumentation mining.

### 2.1 Rhetorical Analysis of Scientific Texts

Previous work has analyzed a number of rhetorical aspects of scientific publications. Teufel et al. (1999, 2009) analyzed the discourse structure of scientific publications. They annotated sentences with discourse categories named *argumentative zones*. Liakata et al. (2010) proposed a more general discourse scheme dubbed *core scientific concepts* and in subsequent work (Liakata et al., 2012) trained a conditional random fields (CRF) model to assign discourse labels to text spans. Several authors focused on tasks relating to citations: extraction of citation context (e.g., Abu-Jbara et al., 2013; Jha et al., 2017), classification of citation polarity (e.g., Athar, 2011) and purpose (e.g., Teufel et al., 2006; Jochim and Schütze, 2012), and the automatic detection of referenced parts of the cited publication (Jaidka et al., 2017). Both discourse and citation information have been exploited for summarizing scientific publications (Cohan and Goharian, 2015; Teufel and Moens, 2002; Abu-Jbara and Radev, 2011; Chen and Zhuge, 2014; Lauscher et al., 2017a). Intuitively, citation contexts may contain information relevant to the summary. Similarly, summaries commonly contain sentences with diversified discourse properties.

Fisas et al. (2016) provided different layers of rhetorical annotations on the same corpus of scientific text.Their Dr. Inventor Corpus is annotated with a combination of existing discourse annotation schemes (Teufel et al., 2009; Liakata et al., 2010) and citation-based annotations. Despite the argumentative nature of scientific texts, the Dr. Inventor Corpus contains no annotations of argumentative components such as claims. Several computational studies followed, addressing the rhetorical tasks corresponding to the layers of the Dr. Inventor Corpus (Ronzano and Saggion, 2015, 2016; Accuosto et al., 2017), but none of them investigated dependencies between different tasks.

The work of Kirschner et al. (2015) is the closest to ours, since they also annotated scientific publications with fine-grained argumentation. However, their corpus is in German and contains no annotations of other rhetorical dimensions. Moreover, their corpus is significantly smaller than the Dr. Inventor Corpus (Fisas et al., 2016). In contrast, we augment the Dr. Inventor Corpus with an argumentation layer, effectively allowing for combinations of argumentation extraction and other rhetorical analysis tasks in MTL settings.

### 2.2 Argumentation Mining

Argumentation mining (AM) refers to extracting (and ideally understanding) arguments from natural language text (Lippi and Torroni, 2015, 2016) and includes tasks like argument detection (Palau

and Moens, 2009), argument component identification (Daxenberger et al., 2017), and argument relation classification (Boltužić and Šnajder, 2014). In their pioneering work on automatic AM, Palau and Moens (2009) discriminated argumentative from non-argumentative sentences and proposed a rule-based approach for extracting argumentative structures in documents. Habernal and Gurevych (2016, 2017) extracted argumentative components from online discussions. They framed the argumentative component extraction as a sequence labeling task and applied structured SVMs as a learning model.

Recent work started exploiting dependencies between AM tasks using global optimization (Peldszus and Stede, 2015; Persing and Ng, 2016; Stab et al., 2014) and MTL models (Eger et al., 2017; Niculae et al., 2017). Peldszus and Stede (2015) used decoding based on minimum spanning trees to jointly predict argumentative segments and their types as well as argumentative relations, to generate an argumentation graph from text. Persing and Ng (2016) and Stab and Gurevych (2017) similarly produced argumentative structures by globally optimizing local predictions of argumentative components and relations. Potash et al. (2017) proposed a neural architecture based on a pointer network for jointly predicting types of argumentative components and identifying argumentative relations. In a similar effort, Eger et al. (2017) combined the AM tasks using the MTL framework of Søgaard and Goldberg (2016). Remedying for data sparsity, Schulz et al. (2018) treated different argumentation formalisms as different tasks and combined respective extraction tasks and datasets in a MTL setting. In contrast to these efforts that combine several AM subtasks or formalisms with joint optimization and MTL models, in this work we examine the dependencies between argumentative components and other rhetorical aspects of scientific writing.

## 3 Data Annotation

We first briefly describe the Dr. Inventor Corpus (Fisas et al., 2016), which we augment with argumentative annotations. We then explain in more detail our argumentation annotation scheme and the annotation process.

### 3.1 Dr. Inventor Corpus

We chose the Dr. Inventor Corpus (Fisas et al., 2015, 2016) as a starting point for two reasons. First, containing 40 publications with a total of 10,789 sentences, it is one of the largest corpora of scientific text manually labeled with rhetorical information. Secondly, it contains *four* different layers of rhetorical annotations: (1) a *discourse* layer, specifying discourse roles of sentences, (2) a *citation context* layer, specifying the textual context of citations, (3) a layer with *subjective aspect* categories assigned to sentences, and (4) a *summarization relevance* layer, indicating how relevant sentences are for the summary. The overview of labels for all annotation layers with the distribution of instances across labels is shown in Table 1. For more details on the original Dr. Inventor Corpus we refer the reader to (Fisas et al., 2015, 2016).

| Annotation Layer | Labels | % |
|---|---|---|
| Discourse Role | *Background* | 20 |
| | *Challenge* | 5 |
| | *Approach* | 57 |
| | *Outcome* | 16 |
| | *Future Work* | 2 |
| Citation Purpose | *Criticism* | 23 |
| | *Comparison* | 9 |
| | *Use* | 11 |
| | *Substantiation* | 1 |
| | *Basis* | 5 |
| | *Neutral* | 53 |
| Subjective Aspect | *Advantage* | 33 |
| | *Disadvantage* | 16 |
| | *Adv.-Disadv.* | 3 |
| | *Disadv.-Adv.* | 1 |
| | *Novelty* | 13 |
| | *Common Practice* | 32 |
| | *Limitation* | 2 |
| Summarization Relevance | *Totally irrelevant* | 66 |
| | *Should not appear* | 6 |
| | *May appear* | 14 |
| | *Relevant* | 6 |
| | *Very relevant* | 8 |

Table 1: Annotation layers of the Dr. Inventor Corpus (Fisas et al., 2016) with label distributions .

### 3.2 Argumentation Annotation Scheme

We considered several existing argumentation frameworks (e.g., Anscombre and Ducrot, 1983; Walton et al., 2008; Dung, 1995, *inter alia*) and selected the Toulmin's model (Toulmin, 2003) as a starting point for our study. We chose the Toulmin's model because: (1) it is a well-established in philosophy as well as in computer science (e.g, Freeman, 1991; Bench-Capon, 1998; Verheij, 2009, *inter alia*) and (2) it contains different types of argumentative components and relations between them into account, which is useful for fine-grained argu-

mentative analyses.

To test the applicability of the framework for our purposes,we first carried out a small preliminary annotation round with two expert annotators and adjusted the annotation scheme according to their observations.

**Argumentative components.** We devised an adapted version of the Toulmin model,[1] containing the following argumentative components:

- *Background claim*: An argumentative statement related to the work of other authors, state-of-the-art methods, or common practices;

  *"The range of breathtaking realistic 3D models is only limited by the creativity of artists and resolution of devices."*

- *Own claim*: An argumentative statement about own work, covered by the publication itself;

  *"Using our method, character authors may use any tool they like to author characters."*

- *Data*: A fact that the authors state as evidence that either supports or contradicts a claim.

  *"SSD is widely adopted in games, virtual reality, and other realtime applications due to its ease of implementation and low cost of computing."*

Argumentative components are annotated as arbitrary spans of text (in terms of length, annotated components ranged from a single token to multiple sentences). Annotators were instructed to annotate the shortest possible span of text that completely captures the argumentative component. Thus, we do not bind arguments to sentences, i.e., we allow for fine-grained argumentative components.

**Argumentative relations.** Authors connect argumentative components in order to form convincing reasoning chains. To allow for the detection of long argumentation chains, we also annotated relations between argumentative components. Following proposals from previous work (Dung, 1995; Bench-Capon, 1998), we distinguish between three relation types:

- *Supports*: indicates that a *claim* component is supported by a data component or another claim. The (assumed) validity of the *supporting* component (data or claim) contributes to the validity of the *supported* claim.

- *Contradicts*: indicates that the validity of a claim decreases with the validity of another argumentative component. If an argumentative component is assumed to be true, the claim it contradicts is assumed to be false, and vice versa.

- *Same claim*: connects different mentions of what is essentially the same claim. It is common to repeat important claims (e.g., the central claim) of the work several times in the publication (*claim coreference*).

Further details about the annotation scheme can be found in the annotation guidelines we provided to our annotators.[2]

### 3.3 Annotation Procedure and Results

**Annotation process.** We hired four annotators for the task, one of whom we considered to be an *expert* annotator[3] and executed the process in two phases. In the first phase, we calibrated the annotators for the task in five iterations, on five publications from the Dr. Inventor Corpus. After all annotators labeled one of the five documents, we met with them, discussed the disagreements, identified erroneous annotations, and, when required, revised the annotation guidelines. At the end of the calibration phase, the annotators re-annotated the five calibration publications and resolved the remaining disagreements by consensus.

In Figure 1 we show the IAA for both component identification and relation classification, in terms of averaged pairwise $F_1$ score,[4] after each of the five calibration iterations. It can be seen that the discussions in the calibration phase helped to get a common understanding of the task among the annotators. However, we note that when considering argumentative relations in addition to the components only, the agreement decreases. Apart from the increased complexity compared to the component identification only this is due to the high ambiguity of argumentative structures, which is one of the main challenges in argument mining,

[1] We omitted some of Toulmin's component types (e.g., *Backing*) due to very rare occurrence in the corpus.

[2] http://data.dws.informatik.uni-mannheim.de/sci-arg/annotation_guidelines.pdf

[3] A researcher in computer science, albeit not in computer graphics, which is the domain of the corpus.

[4] We measured the agreement in terms of the $F_1$ measure because (1) it is straight-forward to compute, (2) it is directly interpretable, and (3) it can account for spans of varying length, allowing for computing relaxed agreements in terms of partial overlaps, and (4) the chance-corrected measures, e.g., Cohen's Kappa, approach $F_1$-measure when the number of negative instances grows (Hripcsak and Rothschild, 2005).

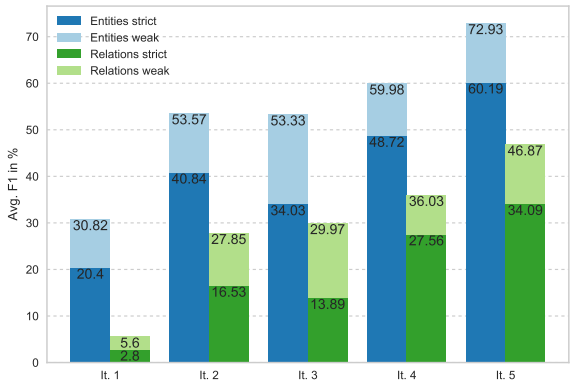



Figure 1: IAA evolution over calibration phases (*blue* for argumentative components; *green* for relations). We report both *strict* (annotated components match in span and type; relations match in type and components at both ends match strictly) and *relaxed* agreement scores (components match in type and overlap in span; relations match in type and their components at both ends match according to the relaxed criterion).

as suggested by Stab et al. (2014). Moreover, disagreements in the argumentative component identification are propagated and cause disagreements in relation annotations, since relation annotations match only when the agreement criterion for the components at both ends is met. Interestingly, the average agreement of our *expert* annotator with *non-expert* annotators was similar to the average agreement between non-expert annotators. This is encouraging, because it suggests that annotating argumentative structures in scientific text does not require expert knowledge of the domain. In the second phase, we evenly split the remaining 35 documents of the Dr. Inventor Corpus among the four annotators, without any overlaps.

**The augmented corpus.** We make the Dr. Inventor Corpus augmented with argumentation annotations (together with the annotation guidelines) publicly available.[5] The final corpus contains 12,289 annotations of argumentative components and 6,530 relation annotations. We show the distributions of labels in Table 2.

The number of *own claims* doubles the number of *background claims*. This is not surprising considering that the Dr. Inventor Corpus contains only original research articles (i.e., no survey nor position articles), in which authors primarily emphasize the contributions of their own work. There are two main reasons for having a smaller number of *data* components compared to *claims*. On one hand, there are longer argumentative chains in which claims are supported by other claims (i.e., only the first claim is supported by the data component). On the other hand, there is also a non-negligible amount of standalone (i.e., unsupported and unchallenged) claims, implied also by having less annotated relations than claims.

[5] http://data.dws.informatik.uni-mannheim.de/sci-arg/compiled_corpus.zip

| Category | Label | Occurrences | % |
|---|---|---|---|
| Component | *Background claim* | 2,751 | 22.4 |
| | *Own claim* | 5,445 | 44.3 |
| | *Data* | 4,093 | 33.3 |
| Relation | *Supports* | 5,790 | 88.7 |
| | *Contradicts* | 696 | 10.7 |
| | *Semantically same* | 44 | 0.7 |

Table 2: Distributions of labels of argumentative components and relations in the corpus.

| | AC | DR | SA | SR |
|---|---|---|---|---|
| AC | – | – | – | – |
| DR | **0.22** | – | – | – |
| SA | 0.08 | 0.11 | – | – |
| SR | 0.04 | 0.10 | 0.13 | – |
| CC | **0.18** | 0.10 | 0.04 | 0.01 |

Table 3: Normalized mutual information between the label sets of the annotation layers indicating argument components (AC), discourse roles (DR), subjective aspects (SA), and citation contexts (CC) in the extended Dr. Inventor Corpus.

To obtain an initial insight on the interrelations between the different rhetorical aspects in scientific writing, we conduct an information-theoretic analysis and assess the amount of information shared among the annotation layers by computing the normalized mutual information (Strehl and Ghosh, 2003). Normalized mutual information is a variant of mutual information, which has been shown to correlate with the gains that can be obtained in multi-task learning settings (Bjerva, 2017). The results can be seen in Table 3. The strongest link is observed between argument components and discourse roles, followed by argument components and citation contexts.

## 4 Multi-task Learning for Rhetorical Analysis of Scientific Writing

We next exploit the augmented corpus to exploit the dependencies between argumentation and other rhetorical dimensions. To this end, we adopt neural MTL as a methodological framework.

### 4.1 Tasks

The following are the rhetorical analysis and argument extraction tasks we investigate.

**Argumentative Component Identification (ACI).** The task is to extract and classify argumentative components. We frame ACI as a token-level sequence labeling task: given a sequence of tokens $\mathbf{x} = (x_1, .., x_n)$ of length $n$, the task is to assign a sequence of tags $\mathbf{y}_{aci} = (y_1, .., y_n), y_i \in Y_{aci}$. The tagset $Y_{aci}$ contains seven token-level tags, obtained by combining the standard B-I-O annotation scheme with three types of argumentative components: *Own claim*, *Background claim*, and *Data*.

**Discourse Role Classification (DRC).** The multi-class classification task in which each sentence needs to be assigned one out of the set of discourse roles $Y_{drc} = \{$*Background*, *Unspecified*, *Challenge*, *FutureWork*, *Approach*, *Outcome*$\}$.

**Citation Context Identification (CCI).** The task is to identify the span of the publication text that introduces or explains a reference. It is also a token-level sequence-labeling task – a sequence of tags $\mathbf{y}_{cci} = (y_1, .., y_n)$ with $y_i \in Y_{cci} = \{B_{CC}, I_{CC}, O\}$ is assigned to a sequence of tokens $\mathbf{x} = (x_1, .., x_n)$.

**Subjective Aspect Classification (SAC).** Another sentence-level classification task in which each sentence is assigned one of the subjective aspect labels, $Y_{sac} = \{$*None*, *Limitation*, *Advantage*, *Disadvantage-Advantage*, *Disadvantage*, *Common Practice*, *Novelty*, *Advantage-Disadvantage*$\}$.

**Summary Relevance Classification (SRC).** The task is to predict the relevance of a sentence for the (extractive) summary of the publication. Each sentence needs to be assigned one of the labels $Y_{src} = \{$*Very relevant*, *Relevant*, *May appear*, *Should not appear*, *Totally irrelevant*$\}$.

ACI and CCI are token-level sequence labeling tasks. The remaining three tasks can be cast as either (1) plain sentence classification tasks or (2) sentence-level sequence labeling tasks (assuming that there are regularities in sequences of sentence-level labels that can be captured). We propose one MTL architecture for each of the two possibilities.

### 4.2 Multi-Task Learning Models

We propose two different MTL architectures for the rhetorical and argumentative analysis of scientific publications. The *Simple model* treats sentence-level tasks (DRC, SAC, and SRC) as plain classification tasks (i.e., the prediction for each sentence ignores the content and labels of other, neighboring sentences). The *Hierarchical model* addresses sentence-level tasks as sequence labeling tasks. This model can be seen as a hierarchical sequence labeling model, in which the sentence-level recurrent network is stacked on top of the token-level sequence labeling network. Both architectures are illustrated in Figure 2.

**Token-level Predictions.** Given a sentence $s_i = (x_{i1}, .., x_{in})$ out of a sequence of sentences $d = (s_1, .., s_m)$ we first retrieve the pre-trained embedding vector for each token $x_{ij}$.We then obtain context-aware token representations $h_{ij}$ by applying a bidirectional recurrent network with long short-term memory cells (Hochreiter and Schmidhuber, 1997) on the sequence of pre-trained word embeddings:

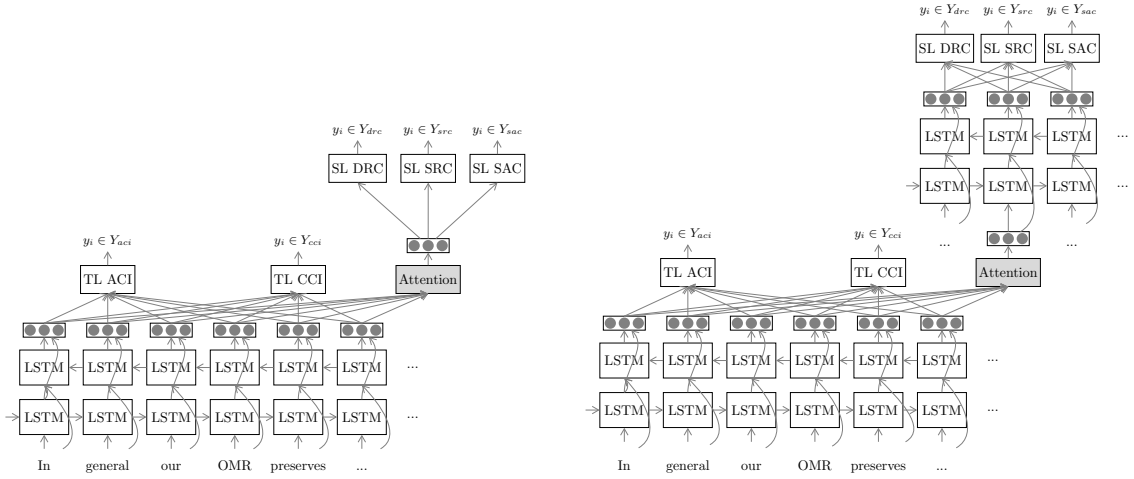
$$h_{ij} = [\overrightarrow{LSTM}(x_{i1}, \ldots, x_{ij}); \overleftarrow{LSTM}(x_{in}, \ldots, x_{ij})]\,. \quad (1)$$

This token-level Bi-LSTM encoder is shared between the tasks combined by the MTL models. Next, we define a separate classifier for each of the token-level (TL) tasks (i.e., ACI and CCI) and feed the contextualized token representations $h_{ij}$ to these classifiers. Each of the classifiers is defined as a feed-forward network with a single hidden layer. The label probability distribution is obtained by applying the *softmax* function on its output.

$$y_{ijt} = \text{softmax}(W_t h_{ij} + b_t)\,, \quad (2)$$

where $W_t \in \mathbb{R}^{2K \times |Y_t|}$ and $b_t \in \mathbb{R}^{|Y_t|}$ are the task-specific classification parameters for the task $t$, with $K$ being the size of the LSTM state and $|Y_t|$ the number of discrete labels of task $t$.

**Sentence-level Predictions.** We learn to aggregate a sentence representation $s_i$ from contextualized vectors of its tokens, $h_{ij}$ (produced by the




(a) Simple model.

(b) Hierarchical model.

Figure 2: Neural MTL architectures for rhetorical and argumentative analysis of scientific publications: (a) the *Simple model* addresses sentence-level tasks (DRC, SAC, SRC) as plain classification tasks, whereas (b) the *Hierarchical model* treats sentence-level tasks as sequence labeling tasks. Both models address ACI and CCI as token-level sequence labeling tasks.

token-level Bi-LSTM), using the intra-sentence attention mechanism (Yang et al., 2016):

$$s_i = \sum_j \alpha_{ij} h_{ij} \,, \tag{3}$$

with the weights $\alpha_i$ computed dynamically as:

$$\alpha_i = \text{softmax}(U_i \, u_{att}) \,, \tag{4}$$

where $u_{att}$ is the trainable attention head vector and $U_i$ is a matrix with non-linearly transformed token representations ($h_{ij}$) as rows:

$$U_{ij} = \tanh(W_{att} h_{ij} + b_{att}) \,. \tag{5}$$

In the *Simple* architecture, sentence representations $s_i$ are fed directly to the sentence-level task-specific classifiers, which are also feed-forward networks with a single hidden layer:

$$y_{it} = \text{softmax}(W_t s_i + b_t) \,. \tag{6}$$

Within the *Hierarchical* architecture, sentence representations are first contextualized with representations of other sentences via the sentence-level Bi-LSTM layer (denoted with the function Bi-LSTM$_S$) and then forwarded to the classifier:

$$y_{it} = \text{softmax}(W_t \text{Bi-LSTM}_S(s_i) + b_t) \,. \tag{7}$$

**Joint optimization and loss functions.** All of the tasks we consider are framed as multi-class classification tasks. Thus, we simply specify all task-specific losses to be L2-regularized cross-entropy errors. Let $y_{to}$ be the one-hot ground truth label vector for the prediction instance $o$[6] of the task $t$, and let $y'_{to}$ be the predicted probability distribution over the task labels for the same instance. With $Y_t$ as the set of labels for task $t$, the task-specific loss $L_t$ is computed as follows:

$$L_t = \lambda \|\Omega_t\|_2 - \sum_o \sum_{k=1}^{|Y_t|} y_{to}^{(k)} \cdot \ln\left(y_{to}^{\prime(k)}\right) \,, \tag{8}$$

where $\Omega_t$ is the set of model's parameters relevant for the task $t$[7] and $\lambda$ is the regularization factor. We train the MTL model jointly on different tasks by defining and minimizing the joint loss function $L$ that combines task-specific losses $L_t$. Instead of using constant weights, we opt for dynamic weighting of task-specific losses during the training process, based on the homoscedastic uncertainty of tasks, as proposed by Kendall et al. (2018):

$$L = \sum_t \frac{1}{2\sigma_t^2} L_t + \ln \sigma_t^2 \,, \tag{9}$$

where $\sigma_t$ is the variance of the task-specific loss over training instances, used to quantify the uncertainty of task $t$. Kendall et al. (2018) show that better MTL results can be obtained by dynamically

[6]The prediction instance is a token for ACI and CCI, and a sentence for DRC, SAC, and SRC.

[7]The set of relevant parameters differs across tasks: for token-level tasks (e.g., ARI) $\Omega_t$ denotes token-level Bi-LSTM parameters and the parameters $W_t$ and $b_t$ of task $t$'s classifier; for a sentence-level task (e.g., DRC) within the *Hierarchical* architecture, $\Omega_t$ includes all parameters of both token- and sentence-level Bi-LSTMs, intra-sentence attention parameters, and parameters of the task-specific classifier.

assigning less weight to the more uncertain tasks, as opposed to constant task weights throughout the whole training process.

## 5 Evaluation

We run two sets of experiments. First, we evaluate the performance of the *Simple* and the *Hierarchical* neural models on individual tasks (i.e., in single-task learning (STL) scenarios). We then evaluate the impact of the argumentative signal on other dimensions of rhetorical analysis by combining them in joint MTL settings.

### 5.1 Experimental Setup.

We randomly split the corpus on the document-level into train (roughly 70%, 28 documents containing 6,697 sentences) and test portions (roughly 30%; 12 documents with 2,874 sentences). We used roughly 20% of the train portion as the validation set for model selection.

**Model configuration and training.** We ran an initial grid search on the validation set with possible values for the hyperparameters learning rate $\nu \in \{10^{-4}, 10^{-5}\}$, L2 regularization factor $\lambda \in \{0.001, 0, 0001\}$, and LSTM states $K \in \{64, 128, 256\}$ and found the hyperparameter configuration $\nu = 10^{-4}$, $\lambda = 0.001$, and $K = 128$ to be optimal for the vast majority of the STL and MTL models. In all experiments, we represent tokens with pre-trained 300-dimensional GloVe embeddings (Pennington et al., 2014)[8] and optimize the model parameters using the Adam algorithm (Kingma and Ba, 2015). We initialize all model parameters using Xavier initialization (Glorot and Bengio, 2010), train the models in batches of $N = 16$ sentences and apply early stopping based on the validation set performance.

**Baselines.** As a type of "sanity check", we first compare the performance of the two neural architectures against traditional supervised machine learning algorithms on each of the tasks separately. For the token-level sequence labeling tasks (ACI and CCI) we use Hidden Markov Models (HMM) and Conditional Random Fields (CRF) (Lafferty et al., 2001) as baselines. The HMM works directly on the tokens, while we feed either the lexical representation or the embedding representation of the tokens as features for the CRF. For the sentence classification tasks (DRC, SAC, and SRC), we evaluate as baselines (1) the linear Support Vector Machines (SVM) with TF-IDF feature vectors and (2) SVM with RBF kernel and embedding features. In the latter case we obtain a sentence representation by averaging the pre-trained embeddings of sentence words. We tune the hyperparameter values of the SVM by conducting a grid search with possible penality parameter values $c \in \{0.1, 1.0, 10.0\}$ (linear SVM and SVM with RBF kernel) and the parameter of the radial basis function $\gamma \in \{0.01, 0.1, 1.0\}$ (SVM with RBF kernel). The possible hyperparameter values for the L1 regularization coefficient $c1$ and for L2 regularization coefficient $c2$ of the CRF are $c1, c2 \in \{0.1, 0.2, 0.001, 0.0001\}$.

In MTL experiments, we consider the respective task performances from single-task experiments and MTL with a joint loss function with equal weighting of the task losses as baselines.

[8] http://nlp.stanford.edu/data/glove.840B.300d.zip.

| | ACI | | | CCI | | |
|---|---|---|---|---|---|---|
| Model | P | R | F | P | R | F |
| HMM | 30.8 | 17.2 | 20.8 | 18.3 | 13.1 | 15.0 |
| $\text{CRF}_{lexical}$ | 38.8 | 29.1 | 31.7 | 15.3 | 17.8 | 16.4 |
| $\text{CRF}_{embeddings}$ | 37.9 | 23.3 | 26.1 | 12.8 | 1.4 | 2.5 |
| Neural: *Simple* | **47.0** | **44.5** | **44.7** | **48.7** | **43.8** | **46.1** |

Table 4: Single-task results for token-level tasks (macro-averaged $F_1$ performances).

| Model | DRC | SAC | SRC |
|---|---|---|---|
| $\text{SVM}_{tfidf}$ | 34.0 | 10.3 | 22.2 |
| $\text{SVM}_{embeddings}$ | 25.7 | 08.5 | 19.3 |
| Neural: *Simple* | **44.1** | **20.5** | 31.5 |
| Neural: *Hierarchical* | 42.6 | 19.1 | **33.2** |

Table 5: Single-task results for sentence-level tasks (macro-averaged $F_1$ scores).

**Single-Task Experiments.** We first report the model performances for individual tasks in STL settings. Results for token-level tasks are shown in Table 4, whereas Table 5 displays results for sentence-level tasks. The scores (precision, recall, and $F_1$ score) are reported as macro-averages over all task labels. Expectedly, our neural architectures substantially outperform the traditional machine learning baselines on all tasks. For the three sentence-level tasks, the *Hierarchical* architecture outperforms the *Simple* model only when classifying sentences by summary relevance (SRC). This result seems intuitive – a *Very relevant* sentence is likely to be surrounded with *Relevant* and *May*

| | CCI | DRC | SAC | SRC |
|---|---|---|---|---|
| **Single Task** | | | | |
| *Simp* | 46.1 | 44.1 | 20.5 | 31.5 |
| *Hier* | – | 42.6 | 19.1 | 33.2 |
| **Multi Task (w. ACI)** | | | | |
| $Simp_{0.5}$ | 43.8 (44.2) | 43.5 (41.6) | 18.0 (42.0) | 32.2 (41.9) |
| $Simp_{\text{uncert}}$ | **49.9** (40.5) | **45.2** (38.6) | **22.1** (39.4) | **34.8** (41.0) |
| $Hier_{0.5}$ | – | 41.6 (42.1) | 17.8 (42.9) | 30.3 (43.4) |
| $Hier_{\text{uncert}}$ | – | **43.9** (40.8) | 18.9 (41.6) | **34.8** (40.8) |

Table 6: MTL results: rhetorical analysis tasks coupled with argumentative component identification. We report the F1 score macro-averaged over the classes. The scores achieved for ACI are shown in parentheses.[9]

*appear* sentences (and an *Irrelevant* sentence with other *Irrelevant* and *Should not appear* sentences). The fact that we observe no gains from the additional sentence-level Bi-LSTM encoder for the DRC and SAC tasks suggests that the content of the sentence informs its discourse role and subjective aspect much more strongly than neighboring sentences. In other words, the DRC and SAC seem to be more localized classification tasks than SRC.

**Multi-Task Learning Results.** Our core research question relates to the effect that recognizing fine-grained argumentative components has on other rhetorical analysis tasks. This is why, in our central set of experiments, we evaluate MTL models with homoscedastic uncertainty weighting which combine the ACI (as an auxiliary task) with each of the four other tasks. In each multi-task learning model, the token-level Bi-LSTM encoder is shared between the two tasks. For sentence-level tasks (DRC, SAC, SRC), we evaluate both the *Simple* and *Hierarchical* architecture. In Table 6 we show the performance of the MTL models on rhetorical analysis tasks (these can be compared to the respective single-task model performances from Tables 4 and 5.

When coupled in MTL settings with argumentation component identification (ACI) using the joint loss formulation of Kendall et al. (2018), the results significantly[10] improve for all rhetorical analysis tasks and models (except for SAC with the *Hierarchical* model), in comparison with the respective single-task models. However, the performance for the argumentation component identification does not improve in MTL. In other words, the extraction of fine-grained argumentative components seems to inform higher-level rhetorical analysis tasks, but not vice-versa. This indeed supports the hypothesis that argumentation guides scientific writing and influences rhetorical structure of publications. Furthermore, our results support the findings of Schulz et al. (2018) who show that, opposed to initial results of Alonso and Plank (2017), MTL can yield performance gains for higher-level semantic tasks.

[9] In the multi-task settings, the early stopping criterion was based on the auxiliary task score.

[10] Differences significant at $p < 0.05$, tested using the non-parametric stratified shuffling test (Yeh, 2000).

## 6 Conclusion

Acknowledging the argumentative nature of scientific text, we investigated the role of argumentation in the rhetorical analysis of scientific publications. We first extended an existing corpus annotated with four different layers of rhetorical information with annotations of argumentative components and relations, creating the largest argumentation-labeled corpus of scientific text in English. We explored intuitive neural architectures with recurrent encoders for argument extraction and rhetorical analysis tasks and showed significant improvements over traditional machine learning models. We then coupled argument extraction with different rhetorical analysis tasks in MTL models with dynamic loss weighting and demonstrated that the argumentative signal has a positive impact on high-level rhetorical analysis tasks.

Admittedly, the corpus we used in this work is limited to the domain of computer graphics. Nonetheless, we believe that our findings relating to the argumentative nature of scientific text and links between argumentation and other rhetorical aspects generalize to other domains too. This is also supported by the comparable agreement observed between expert and non-expert annotators.

In the future work, we would like to extend the collection of scientific text to other fields. Next, we intend to explore a wider range of MTL models, especially those involving more than two tasks. Having annotated argumentative relations, we will work on models for their automated identification in scientific publications.

## Acknowledgments

This research was partly funded by the German Research Foundation (DFG), grant number EC 477/5-1 (LOC-DB). We thank our four annotators for their dedicated annotation effort and the anonymous reviewers for constructive and insightful comments.

## References


Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 596–606, Atlanta, Georgia. Association for Computational Linguistics.

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Volume 1*, pages 500–509, Portland, OR, USA. Association for Computational Linguistics.

Pablo Accuosto, Francesco Ronzano, Daniel Ferrés, and Horacio Saggion. 2017. Multi-level mining and visualization of scientific text collections: Exploring a bi-lingual scientific repository. In *Proceedings of the 6th International Workshop on Mining Scientific Publications*, pages 9–16, Toronto, ON, Canada. Association for Computing Machinery.

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? Semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 44–53, Valencia, Spain. Association for Computational Linguistics.

Jean-Claude Anscombre and Oswald Ducrot. 1983. *L'argumentation Dans La Langue*. Editions Mardaga.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA. Association for Computational Linguistics.

Trevor JM Bench-Capon. 1998. Specification and implementation of toulmin dialogue game. In *Proceedings of the 11th Conference on Legal Knowledge Based Systems*, pages 5–20, Groningen, Netherlands. Foundation for Legal Knowledge Based Systems.

Johannes Bjerva. 2017. Will my auxiliary tagging task help? estimating auxiliary tasks effectivity in multi-task learning. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 216–220, Gothenburg, Sweden. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the 1st Workshop on Argumentation Mining*, pages 49–58, Baltimore, MD, USA. Association for Computational Linguistics.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Jingqiang Chen and Hai Zhuge. 2014. Summarization of scientific documents by detecting common facts in citations. *Future Generation Computer Systems*, 32:246–252.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400, Lisbon, Portugal. Association for Computational Linguistics.

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

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

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

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 3081–3088, Portorož, Slovenia. European Language Resources Association.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 42–51, Denver, CO, USA. Association for Computational Linguistics.

James B. Freeman. 1991. Dialectics and the macrostructure of arguments: A theory of argument structure.

G Nigel Gilbert. 1976. The transformation of research findings into scientific knowledge. *Social Studies of Science*, 6(3-4):281–306.

G Nigel Gilbert. 1977. Referencing as persuasion. *Social Studies of Science*, 7(1):113–122.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*,

pages 249–256, Sardinia, Italy. Proceedings of Machine Learning Research.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1122, Berlin, Germany. Association for Computational Linguistics.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Kokil Jaidka, Muthu Kuma Chandrasekaran, Devanshu Jain, and Min-Yen Kan. 2017. Overview of the CL-SciSumm 2017 Shared Task. In *Proceedings of the Computational Linguistics Scientific Summarization Shared Task*, pages 1–15, Tokyo, Japan. CEUR-WS.

Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R. Radev. 2017. NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, 23(1):93–130.

Charles Jochim and Hinrich Schütze. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1343–1358, Mumbai, India. The COLING 2012 Organizing Committee.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA. Institute of Electrical and Electronics Engineers.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.

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

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, Williamstown, MA, USA. Morgan Kaufmann Publishers Inc.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2017a. Citation-based summarization of scientific articles using semantic textual similarity. In *Proceedings of the Computational Linguistics Scientific Summarization Shared Task*, pages 33–42, Tokyo, Japan. CEUR-WS.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2017b. Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. In *Proceedings of the 6th International Workshop on Mining Scientific Publications*, pages 24–28, Toronto, ON, Canada. Association for Computing Machinery.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin R Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2054–2061, Valetta, Malta. European Language Resources Association.

Marco Lippi and Paolo Torroni. 2015. Argument mining: A machine learning perspective. In *Theory and Applications of Formal Argumentation: Third International Workshop, TAFA 2015, Buenos Aires, Argentina, July 25-26, 2015, Revised Selected Papers*, Lecture Notes in Artifical Intelligence, pages 163–176. Springer International Publishing.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25.

Kathy McKeown, Hal Daume, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R. Fleischmann, Luis Gravano, Rahul Jha, Ben King, Kevin McInerney, Taesun Moon, Arvind Neelakantan, Diarmuid O'Seaghdha, Dragomir Radev, Clay Templeton, and Simone Teufel. 2016. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, 67(11):2684–2696.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107, Barcelona, Spain. Association for Computing Machinery.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, CA, USA. Association for Computational Linguistics.

Peter Potash, 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, Denmark. Association for Computational Linguistics.

Francesco Ronzano and Horacio Saggion. 2015. Dr. Inventor Framework: Extracting structured information from scientific publications. In *Discovery Science, 18th International Conference, DS 2015, Banff, AB, Canada, October 4-6, 2015. Proceedings*, Lecture Notes in Computer Science, pages 209–220, Banff, Canada. Springer, Cham.

Francesco Ronzano and Horacio Saggion. 2016. Knowledge extraction and modeling from scientific publications. In *Semantics, Analytics, Visualization. Enhancing Scholarly Data, Second International Workshop, SAVE-SD 2016, Montreal, QC, Canada, April 11, 2016, Revised Selected Papers*, Lecture Notes in Computer Science, pages 11–25, Montreal, Canada. Springer, Cham.

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

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

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

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 21–25. CEUR-WS.

Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Simone Teufel, 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: Volume 3*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Simone Teufel, 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, Australia. Association for Computational Linguistics.

Stephen E. Toulmin. 2003. *The Uses of Argument*, updated edition. Cambridge University Press.

Bart Verheij. 2009. The toulmin argument model in artificial intelligence. In Guillermo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 219–238. Springer US, Boston, MA.

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

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, CA, USA. Association for Computational Linguistics.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953, Saarbrücken, Germany. Association for Computational Linguistics.