Transferring knowledge from discourse to arguments: A case study with scientific abstracts

Pablo Accuosto and Horacio Saggion

Large-Scale Text Understanding Systems Lab (LaSTUS) / TALN Group
Department of Information and Communication Technologies
Universitat Pompeu Fabra
C/Tànger 122-140, 08018 Barcelona, Spain
{name.surname}@upf.edu

Abstract

In this work we propose to leverage resources available with discourse-level annotations to facilitate the identification of argumentative components and relations in scientific texts, which has been recognized as a particularly challenging task. In particular, we implement and evaluate a transfer learning approach in which contextualized representations learned from discourse parsing tasks are used as input of argument mining models. As a pilot application, we explore the feasibility of using automatically identified argumentative components and relations to predict the acceptance of papers in computer science venues. In order to conduct our experiments, we propose an annotation scheme for argumentative units and relations and use it to enrich an existing corpus with an argumentation layer.[1]

1 Introduction

The growing number of scientific publications and the shortening of the research-publication cycles (Bornmann and Mutz, 2015) pose a challenge to authors, reviewers and editors. The development of automatic systems to support the quality assessment of scientific texts can facilitate the work of editors and referees of scientific publications and, at the same time, be of value for researchers to obtain feedback that can lead to improve the communication of their results.

The quality assessment of scientific texts has many dimensions, and each one involves different levels of difficulties. While the relevance of the problem at stake and the novelty of the solutions proposed by the authors are of great significance in terms of weighting the ultimate contributions of the work, aspects such as the argumentative structure of the text are key when analyzing its effectiveness with respect to its communication objectives (Walton and Walton, 1989). A fine-grained assessment of the contributions made in research articles requires to identify the main claims made by the authors and to determine if the evidence provided to support them is *strong* enough. Or, in other terms, if both the structure and the contents of the arguments proposed by the authors can persuade a potential reader of the validity of their contributions.

In addition to being useful for facilitating the assessment of some quality aspects of a text, the automatic identification of argumentative units and their relations–a set of related tasks known as *argument mining*–is a relevant problem in itself in the context of knowledge mining (Mochales and Moens, 2011). Being able to extract not only what is being stated by the authors of a text but also the reasons they provide to support it can be useful in multiple applications, ranging from a fine-grained analysis of opinions to the generation of abstractive summaries of texts. As an example of a potential application for argument mining, (Lippi and Torroni, 2016) suggest the possibility of developing an argumentative ranking component in a search engine so that it retrieves documents based on claims and evidence on a given topic extracted automatically from texts.

The tasks involved in the extraction of arguments from text–including the identification of argumentative sentences, the detection of argument component boundaries and the prediction of argument structures–are related to other text mining tasks–including sequence labeling, text segmentation, entity recognition and relation extraction–which are in general tackled by means of supervised learning methods (Lippi and Torroni, 2016). The lack of annotated data with argumentative information, however, presents a challenge when trying to apply these well-known approaches to argument mining (Stab and Gurevych, 2017). This is so, in part, due to the inherent difficulty of unambiguously identifying argumentative elements

[1] Available at http://scientmin.taln.upf.edu/argmin/scidtb_argmin_annotations.tgz.

Proceedings of the 6th Workshop on Argument Mining, pages 41–51
Florence, Italy, August 1, 2019.

in texts, which is reflected in the low levels of inter-annotator agreement reached in general for this task (Habernal et al., 2014). If this is true in several knowledge domains, it poses a more difficult problem in the case of scientific texts due to their inherent argumentative complexity (Kirschner et al., 2015; Green, 2015). We propose to address this challenge by leveraging data annotated with discourse relations, as previous works suggest potential benefits in linking discourse analysis and argument mining tasks (Peldszus and Stede, 2016; Stab et al., 2014; Cabrio et al., 2013; Biran and Rambow, 2011; Green, 2015).

1.1 Contributions

- We propose to tackle the limitations posed by the lack of annotated data for argument mining in the scientific domain by leveraging existing Rhetorical Structure Theory (RST) (Mann et al., 1992) annotations in a corpus of computational linguistics abstracts (SciDTB) (Yang and Li, 2018). In order to do so:
  1. We propose and test an annotation scheme that we use to conduct a pilot annotation experiment in which we enrich a subset of the SciDTB corpus with an additional layer of argumentative structures.
  2. We explore the potential of a transfer learning approach to improve the performance of an argument mining model trained with a small volume of data annotated with the proposed scheme.
- We report preliminary results on the prediction of acceptance or rejection of scientific papers in computer science conferences based on the automatic identification of argumentative components and relations in their abstracts.

In this work we adopt a pragmatic perspective in relation to exploring the predictive potential of the argumentative structure of an abstract for the acceptance or rejection of the corresponding manuscript in a peer review process. We do not intend to imply that the ultimate quality of the papers–or even the abstracts–could be determined solely by considering this limited information.

The rest of the paper is organized as follows: in Section 2 we describe previous work, focusing, in particular, on works aimed at identifying arguments in scientific texts. In Section 3 we describe the dataset used in our experiments and our proposed annotation scheme for fine-grained scientific argument mining. In Section 4 we describe our transfer learning experiments, their experimental settings and results and, in Section 5, we do the same with the experiments aimed at predicting the acceptance or rejection of papers in conferences. Finally, in Section 6, we summarize our main contributions and propose additional research avenues as follow-up to the current work.

2 Related work

This work is informed by previous research in the areas of argument mining, argumentation quality assessment and the relationship between discourse and argumentative structures and, from the methodological perspective, to transfer learning approaches. Due to space restrictions, we cannot cover in detail all the relevant background work. We refer the reader to (Lippi and Torroni, 2016) for a thorough summary of argument mining initiatives in various domains and with different approaches. (Wachsmuth et al., 2017) provide a comprehensive survey of quality assessment approaches in the context of computational argumentation and categorize them in relation to how they address logical, rhetorical and dialectical dimensions of argumentation. (Pan and Yang, 2010) provide an in-depth review of current trends in transfer learning, including inductive, transductive and unsupervised approaches. Furthermore, they classify the different approaches based on *what is transferred*: instances, feature representations, parameters or relational knowledge. A more direct antecedent to our work is the research conducted by Peldszus and Stede (Peldszus and Stede, 2016, 2015a; Stede et al., 2016), who annotated 112 argumentatively rich texts using RST and argumentation schemes in order to study the relationship between discourse and argumentation structures. The texts were generated in an experiment in which several participants wrote short texts of controlled linguistic and rhetoric complexity discussing a controversial issue from a pre-defined list. Based on this corpus, the authors conducted experiments in order to derive argumentative components and relations from RST trees, comparing three approaches: a transformation model, an aligner based on sub-graph matching and an evidence graph model (Peldszus and Stede, 2015b).

Our work is one of few that deal with argument mining in scientific texts which, as mentioned in Section 1, is considered as a particularly challenging domain (Kirschner et al., 2015; Green, 2015). (Stab et al., 2014) and (Kirschner et al., 2015) carried out annotation studies with scientific articles in educational research with binary argumentative and discourse relations (*support*, *attack*, *detail*, and *sequence*). In order to calculate the agreement between the annotators that participated in the process they developed a novel graph-based agreement measure, which can identify different annotations with similar meaning, thus obtaining higher agreement than with standard measures. The evaluation of argument annotations is still an open issue. (Stab et al., 2014) suggest that it might be interesting to explore, for this task, evaluation schemes that are able to deal with multiple correct annotations such as those used in text summarization. (Lauscher et al., 2018b) analyze the information shared by rhetorical and argumentative structure of scientific documents. In order to do this, they add an argumentation layer to the DrInventor Scientific Corpus (Fisas et al., 2016), which includes 40 computer graphics papers annotated with four layers including citation contexts, rhetorical role of sentences, subjective information and summarization relevance. The enriched corpus is used to trained new models for the automatic identification of claims and evidence, which are made available as a web service (Lauscher et al., 2018a). Some of the first initiatives aimed at the automatic identification of rhetorical and argumentative components in scientific texts include the Argumentative Zoning (AZ) model (Teufel et al., 1999, 2009) and the CoreSC scheme (Liakata et al., 2012). While AZ considers annotations for knowledge claims made by the authors of scientific articles, CoreSC associates research components to the parts of the texts describing them, thus obtaining a readable representation of the research process described by the paper. Both of them are sentence-based schema that are focused on the identification of the components and do not consider the relations between them. (Feltrim et al., 2006) adapted the AZ model for the automatic annotation of scientific abstracts in Portuguese (AZPort). The AZPort model was integrated as a module of SciPo,[2] a web-based tool aimed at supporting novice writers of academic texts: given an abstract, the system classifies its sentences by means of AZPort and, based on a set of rules for well-formed rhetorical structures, it provides feedback for potential improvements (e.g., re-ordering the elements of the text or adding missing content). More recently, (Vargas-Campos and Alva-Manchego, 2016) adapted the AZPort model to Spanish (AZEsp), which was also integrated into a computer-assisted writing tool for computer science dissertations in Spanish (SciEsp).

3 Annotated data

In order to explore the possibility of leveraging discourse information for the identification of argumentative components and relations we add a new annotation layer to the Discourse Dependency TreeBank for Scientific Abstracts (SciDTB) (Yang and Li, 2018). SciDTB contains 798 abstracts from the ACL Anthology (Radev et al., 2013) annotated with elementary discourse units (EDUs) and relations from the RST Framework. Polynary discourse relations in RST are binarized in SciDTB following a criteria similar to the "right-heavy" transformation used in other works that represent discourse structures as dependency trees (Morey et al., 2017; Stede et al., 2016; Li et al., 2014).

We consider a subset of the SciDTB corpus consisting of 60 abstracts from the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) and transformed them into a format suitable for the GraPAT graph annotation tool (Sonntag and Stede, 2014)[3], which had been previously tailored to the specificities of our proposed annotation scheme, described in Section 3.1.

The corpus enriched with the argumentation[4] level contains a total of 327 sentences, 8012 tokens, 862 discourse units and 352 argumentative units linked by 292 argumentative relations.

3.1 Annotation scheme

Several argumentation mining works (Lippi and Torroni, 2016) use *claims* and *premises* as basic argumentative units. In the case of scientific discourse, however, it is frequent to find that claims

[2] http://www.nilc.icmc.usp.br/scipo/

[3] http://angcl.ling.uni-potsdam.de/resources/grapat.html

[4] The annotations are made available to download at http://scientmin.taln.upf.edu/argmin/scidtb_argmin_annotations.tgz

are not explicitly stated in an argumentative writing style but are instead left implicit (Hyland, 1998). The description of the problem addressed in the paper, for instance, usually conveys implicit claims in relation to the relevance of the problem at stake and/or the adequacy of the proposed approach. We introduce a fine-grained annotation scheme aimed at capturing information that accounts for the specificities of the scientific discourse, including the type of evidence that is offered to support a statement (e.g., background information, experimental data or interpretation of results). This can provide relevant information, for instance, to assess the *argumentative strength* of a text. The types of proposed units in our scheme were considered so they can be mapped–even if with a different level of granularity–to concepts in CoreSC (Liakata et al., 2010) and AZ categories, which would enable additional research on the potential of using existing annotated corpora for argument mining tasks. Like (Peldszus and Stede, 2016)–and in contrast with CoreSC and AZ–we consider EDUs as the minimal spans that can be annotated. Argumentative units can, in turn, cover multiple sentences.

The proposed units include:

- ***proposal*** (problem or approach)
- ***assertion*** (conclusion or known fact)
- ***result*** (interpretation of data)
- ***observation*** (data)
- ***means*** (implementation)
- ***description*** (definitions/other information)

In line with (Kirschner et al., 2015), we adopt in our annotation scheme the classic ***support*** and ***attack*** argumentative relations and the two discourse relations ***detail*** and ***sequence***.

Figure 1 shows a subset of the argumentative components and relations annotated in an abstract from (Zhang and Wang, 2014),[5] including a *proposal* and two supporting units: an *assertion* and a *result*. Figure 2 shows the original discourse units and relations as annotated in SciDTB.

In the subset of SciDTB annotated for our experiments, the types of argumentative units are distributed as follows: 31% of the units are of type *proposal*, 25% *assertion*, 21% *result*, 18% *means*, 3% *observation*, and 2% *description*. In turn, the relations are distributed: 45% of type *detail*, 42% *support*, 9% *additional*, and 4% *sequence*. No *attack* relations were identified in the set of currently annotated texts. When considering the distance[6] of the units to their parent unit in the argumentation tree, we observe that the majority (57%) are linked to a unit that occurs right before or after it in the text, while 19% are linked to a unit with a distance of 1 unit in-between, 12% to a unit with a distance of 2 units, 6% to a unit with a distance of 3, and 6% to a unit with a distance of 4 or more.[7]

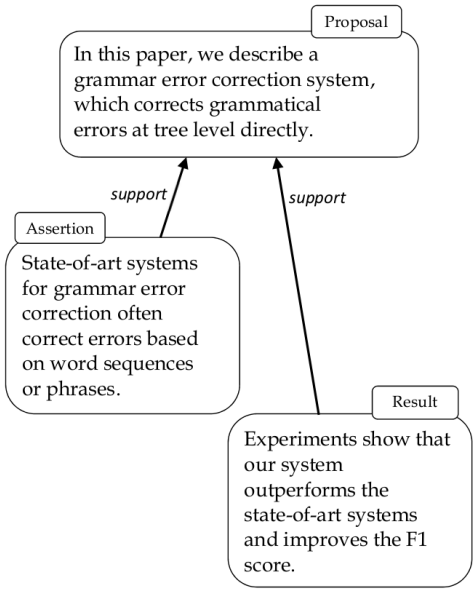

Figure 1: Partial argumentative structure

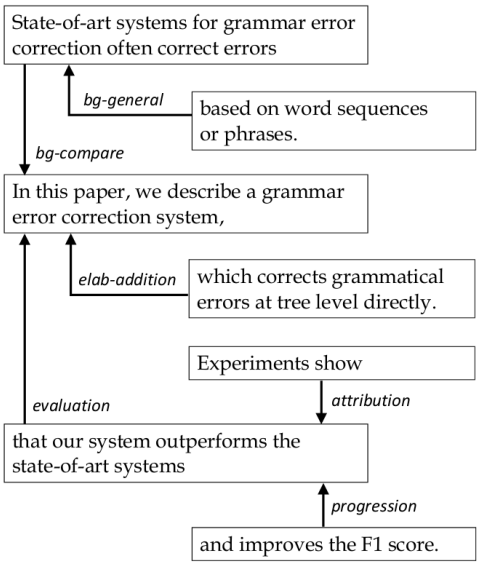

Figure 2: Partial discourse structure

[5] http://aclweb.org/anthology/D14-1033

[6] By *distance* we refer to the number of argumentative units that occur between two units in the text.

[7] According to the position of the parent unit, there are 200 relations pointing forward and 92 in which the parent occurs

4 Transfer learning experiment

The first set of experiments, described in this section, are aimed at exploring the potential of applying a transfer learning method to improve the performance of argument mining tasks trained with a small corpus of 60 abstracts by leveraging the discourse annotations available in the full SciDTB corpus.

4.1 Tasks

We define the following set of argument mining tasks:

- **AFu (argumentative function)**: Identify the boundaries and argumentative functions of the components. In the example in Fig. 1, it would imply to identify the boundaries of the three nodes and the two *support* relations that link them.
- **ATy (argumentative unit)**: Identify the boundaries and types of the components. In the example, the *proposal*, *assertion* and a *results* units.
- **APa (argumentative attachment)**: Identify the boundaries of the components and the relative position of the parent argumentative unit. For instance, the *assertion* unit in Fig. 1 is attached to the *proposal* unit with a relative distance of one unit in the forward direction (as the assertion occurs right before the proposal in the text). The *result* unit, in turn, is attached to the *proposal* with a distance of four units in the background direction (the units that occur between these two nodes are omitted in the figure).

4.2 Experimental setups

We train each of the tasks described in 4.1 separately and compare the results obtained with those obtained by an inductive transfer learning method in which we use encoders trained with the RST annotations available in the SciDTB corpus. These encoders are then used to produce contextualized representations of the input tokens that are fed to the argument mining learning processes.

The discourse parsing tasks considered to train the specialized encoders are:

- **DFu (discourse function)**: Identify the boundaries and discourse roles of the EDUs (*attribution*, *evaluation*, *progression*, etc.).[8]
- **DPa (discourse attachment)**: Identify the boundaries of the EDUs and the relative position of the parent units in the RST tree.

The discourse tasks (DFu and DPa) are trained with the 738 abstracts left in the SciDTB corpus when excluding the 60 abstracts annotated with arguments. This is done in order to avoid introducing a bias that would not reflect the results obtained when no discourse annotations are available.

All the argument mining models (AFu, ATy, APa) are trained and evaluated in a 10-fold cross-validation setting.

In all cases the models are generated by means of bi-directional long short-term memory (BiLSTM) networks, as this type of architecture has proven to perform reasonably well in argument mining tasks across different classification scenarios (Eger et al., 2017). In order to simplify the experiments and the interpretation of their results we use the same architecture for all tasks: two layers of 100 recurrent units, Adam optimizer, naive dropout probability of 0.25 and a conditional random fields (CRF) classifier as the last layer of the network. We use, for the BiLSTMs, the implementation made available by the Ubiquitous Knowledge Processing Lab of the Technische Universität Darmstadt (Reimers and Gurevych, 2017).[9] As our intention is to compare the different approaches and not necessarily obtain the best possible models for these tasks, no hyperparameter optimization is done in these experiments and, in all of the cases, the networks are trained for 100 epochs.

All of the tasks are modeled as sequence labeling problems in which the tokens are tagged using the beginning-inside-outside (BIO) tagging scheme. The tokens are encoded as the concatenation of 300-dimensional dependency-based word embeddings (DEmb)[10] ($\vec{k}$) (Komninos and Manandhar, 2016) and 1024-dimensional contextualized word embeddings (ELMo) ($\vec{e}$) (Peters et al., 2018). In these experiments we use the 5.5 billion-token version of ELMo trained with Wikipedia and monolingual news from the WMT 2008-2012

before in the text.

[8] Please refer to (Yang and Li, 2018) for a description of the discourse roles used in SciDTB.

[9] https://github.com/UKPLab/elmo-bilstm-cnn-crf

[10] https://www.cs.york.ac.uk/nlp/extvec/

corpora.[11] For the experiments with the RST encoders we include the 200-dimensional embeddings obtained from the concatenation of the backward and forward hidden states of the top layers of the DFu or DPa models (RSTEnc) ($\vec{f}$ and $\vec{p}$, respectively). Table 1 summarizes the sets of embeddings used in these experiments and their dimensions.

Each argument mining task is paired with one discourse parsing task for the transfer learning experiments. While AFu and ATy are paired with DFu, APa is paired with DPa. This means that the input for the AFu and ATy tasks is obtained as the concatenation of the vectors $[\vec{k}, \vec{e}, \vec{f}]$, while in the case of APa the input is $[\vec{k}, \vec{e}, \vec{p}]$.

Abbreviation	Notation	Dimensions
DEmb	$\vec{k}$	300
ELMo	$\vec{e}$	1024
GloVe	$\vec{g}$	200
RSTEnc (DFu/DPa)	$\vec{f}$ / $\vec{p}$	200

Table 1: Word embeddings used in the experiments

4.3 Results

We adopt the ConNLL criteria for named-entity recognition[12] to evaluate the performances obtained in the identification of argumentative components and relations. Table 2 shows the average F1-measures obtained for each of the settings considering the epochs 10 to 100.[13] The argument mining models trained with the representations produced by the RST encoders (*DEmb+ELMo+RSTEnc*) yield better performances, with gains of 0.03, 0.04 and 0.02 F1 points for AFu, ATy and APa, respectively, over the models trained solely with the dependency-based and ELMo embeddings (*DEmb+ELMo*).

Setting	AFu	ATy	APa
DEmb+ELMo	0.66	0.63	0.38
DEmb+ELMo+GloVe	0.65	0.65	0.38
DEmb+ELMo+RSTEnc	**0.69**	**0.67**	**0.40**

Table 2: Average F1-measures in epochs 10-100

[11] https://allennlp.org/elmo

[12] A true positive is considered when both the boundary and the type of the entity match.

[13] The epochs before the 10th are not significant as the models have not had enough time to learn anything.

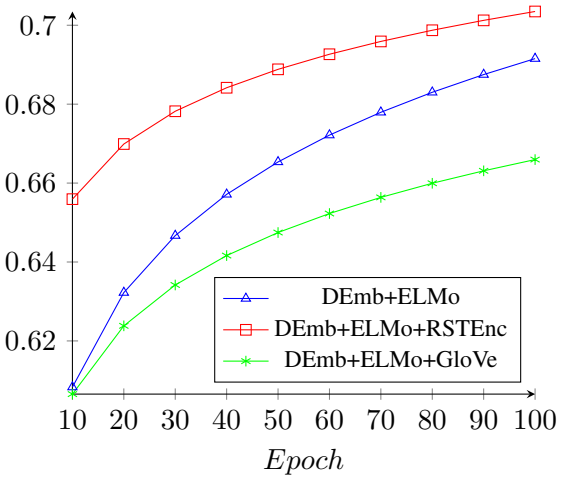

Figure 3: Trend lines for F1-measures in epochs 10-100 for AFu

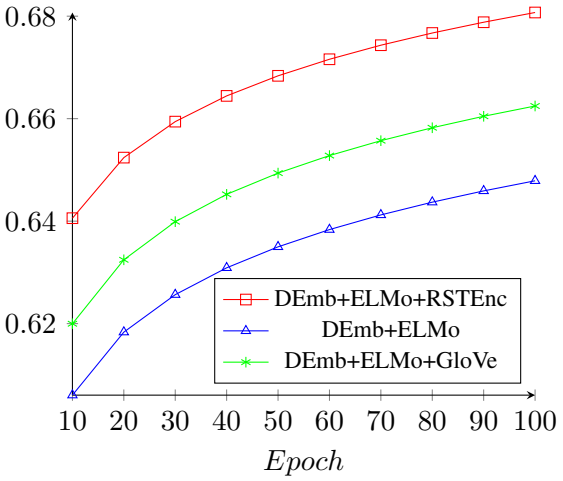

Figure 4: Trend lines for F1-measures in epochs 10-100 for ATy

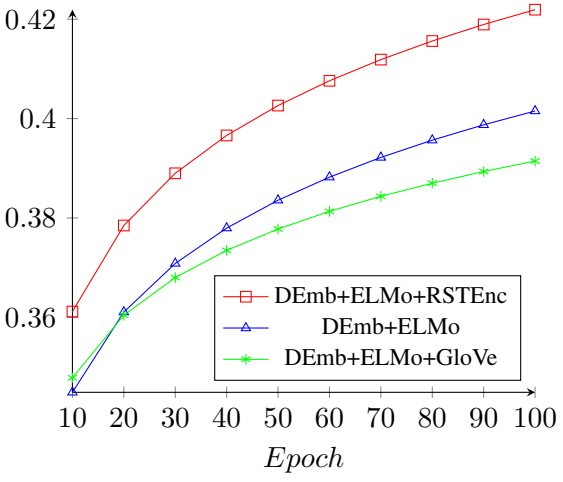

Figure 5: Trend lines for F1-measures in epochs 10-100 for APa

In order to determine whether the better performance of the RST encoders is due to the knowledge conveyed by the task-specific representations we conducted an additional experiment in which we concatenated 200-dimensional GloVe embeddings[14] (Pennington et al., 2014) ($\vec{g}$) obtaining 1524-dimension embeddings $[\vec{k}, \vec{e}, \vec{g}]$ used as input of each of the argument mining models. In this case, the results obtained are mixed, with an increase in performance of 0.02 F1 points in average–for the epochs 10 to 100–for ATy, a worse performance of 0.01 F1 points for AFu and no difference in performance for APa. The models with the GloVe embeddings (*DEmb+ELMo+GloVe*) have, therefore, worse performances in average of 0.04, 0.02 and 0.02 F1 points for AFu, ATy and APa with respect to the models that include the embeddings obtained by means of the RST encoders.

Figures 3, 4 and 5 show the trend lines of F1-measures obtained with the different models for the epochs 10 to 100 for the AFu, ATy and APa tasks, respectively. The graphs show that the models with information from the RST encoders not only learn better the argument mining tasks but they also do it in less time with respect to the other settings.

These results support out initial hypothesis in the sense that transferring discourse knowledge by means of representations learned in discourse parsing tasks can contribute to improve the performance of argument mining models trained with a rather small number of instances.

5 Acceptance prediction experiment

As a pilot application we explore the possibility of predicting the acceptance/rejection of papers in computer science conferences[15] based on the annotations generated by the best argument mining models of the experiments described in Section 4.

Quality assessment metrics that consider elements such as *clarity and simplicity, lack of redundancy and comprehensiveness* of scientific reporting have been developed for abstracts in other domains–in particular, in life sciences–(Timmer et al., 2003). These instruments were used in studies that show that abstracts with higher formal quality scores–as measured by human experts–are more frequently accepted for presentations in conferences (Timmer et al., 2001). We do not believe that these results can be directly extrapolated to the quality assessment of scientific abstracts in computer science, an area in which full manuscripts are most frequently considered for review and where abstracts have less fixed structures. Furthermore, clearer links between the formal quality of scientific reporting and the overall quality of research in computer science still need to be established. Considering all these limitations, we were interested in exploring whether the automatically identified argumentative structure of the abstracts could reflect some quality aspects of the full manuscripts and if this, in turn, could contribute to predict their acceptance in conferences in a specific research area in the field of computer science.

5.1 Dataset

As training set for the acceptance prediction experiment we use 117 abstracts of manuscripts submitted to the Compact Deep Neural Network Representation with Industrial Applications (CDNNRIA) and the Interpretability and Robustness for Audio, Speech and Language (IRASL) workshops held in the context of the Thirty-second Conference on Neural Information Processing Systems (NIPS 2018). As test set we use 30 abstracts of manuscripts submitted to the Sixth International Conference on Learning Representations (ICLR 2018). All of the abstracts were collected from the OpenReviews website (Soergel et al., 2013).[16]

The distribution of accepted/rejected papers in the training and test sets is shown in Table 3

Set	Conference	Accepted	Rejected
Train	*CDNNRIA*	35	23
Train	*IRASL*	30	29
		55	52
Test	*ICLR*	15	15

Table 3: Accepted/rejected papers in training and test sets

5.2 Experimental setup

The CDNNRIA, IRASL and ICLR abstracts are used as input to the AFu, ATy and APa models

[14] We used the 6 billion tokens versions trained with Wikipedia 2014 and Gigaword 5 available at https://nlp.stanford.edu/projects/glove/

[15] In particular, in the areas of neural-based systems and its applications to speech and language.

[16] https://openreview.net/

described in Section 4 obtaining sequences of argumentative units, types and parent attachments. These sequences are then used as features to train and evaluate a binary classifier aimed at predicting the acceptance or rejection of the corresponding papers. Table 4 shows sample training/test instances. As the number of argumentative units identified in each abstract might differ we use padding values (*nofunc*, *notype* and *100* for AFu, ATy and APa, respectively) to generate training and test instances with a fixed number of features (equal to three times the maximum number of argumentative units identified in the dataset).

x_1	x_2	...	x_n
none	*additional*	...	*support*
support	*support*	...	*none*
...	...	...	...
support	*nofunc*	...	*nofunc*
proposal	*assertion*	...	*assertion*
result	*assertion*	...	*proposal*
...	...	...	...
observation	*notype*	...	*notype*
0	*1*	...	*1*
1	*1*	...	*0*
...	...	...	...
-5	*100*	...	*100*

y_1	y_2	...	y_n
REJECT	ACCEPT		ACCEPT

Table 4: Example of input instances to the classifier

Considering that we are dealing with a small set of features with a reduced number of potential values for each one, we use a decision tree algorithm for our pilot classification experiment. In addition to the training and evaluation speed of the algorithm we consider that the higher interpretability of the results–by examining the decision points–can also contribute to assess to what degree the different elements of the predicted argumentative structure are used in the classification. We use Weka's implementation of the C4.5 algorithm (Quinlan, 1993) (J48) with default parameters with the exception of the confidence factor used for pruning the tree, which was selected evaluating the different models obtained against a random split of 20% of the test set used for validation.[17] As the training set is not perfectly balanced, we pre-process the data with Weka's ClassBalancer algorithm, which assigns weights to each instance so that each class has the same total weight.

5.3 Results

The classifier trained with the argumentative units and relations extracted from the CDNNRIA/IRASL abstracts has a performance of 0.67 F1-score when evaluated with the training set obtained from processing the ICLR abstracts,[18] 0.17 F1 points above a random binary classification in a balanced set. As expected, the main decision points in the tree correspond, broadly, to those attributes that are also ranked higher when measuring their contribution to reduce the entropy with respect to the class.[19] Observing these features, we can see that the most relevant decision elements are the parent attachment of first argumentative unit, the argumentative functions of the first two units and the argumentative type of the first unit. Also relevant are the features that mark the end of the sequences of argumentative types and functions for the majority of the instances. This means that the number of identified units also have a relevant role in the predictions. However, the number of units by themselves is not a good predictor of the abstract's class. In fact, executing the same experiment but replacing the non-padding values for function, type and attachment for fixed values we obtain an F1-measure of 0.59 due, in particular, to a higher number of false negatives (accepted papers classified as rejected).

Features	**P**	**R**	**F1**
Arg. units alone	0.67	0.53	0.59
Arg. units with types, functions and parents	0.67	0.67	**0.67**

Table 5: Precision, recall and F1-measures for the acceptance prediction classifiers with and without fine-grained argumentative information

6 Conclusions and future work

In this work we explored the potential of leveraging existing discourse-annotated corpora to im-

[17] `weka.classifiers.trees.J48 -C0.6 -M2`

[18] 20 of the abstracts were correctly classified and ten were mis-classified: five as false positives and five as false negatives

[19] As calculated by means of Weka's InfoGainAttributeEval algorithm.

prove the performance of fine-grained argument mining models trained with a limited number of examples. In order to test our hypothesis, we proposed an annotation scheme and used it to enrich, with a new layer of argumentative structures, a subset of a corpus previously annotated with discourse-level units and relations. Promising results are obtained by implementing an inductive transfer learning method in which contextualized representations obtained by means of encoders trained with discourse parsing tasks are used as input of argument mining models. As a potential application of the annotations produced by the argument mining models, we implemented a simple classifier aimed at predicting the potential acceptance/rejection of computer science papers according to the argumentative structure of their abstracts. The results of these preliminary experiments are auspicious and motivate us to continue working in this area. As a first step in this direction, we plan to extend the coverage of the argumentative layer of annotations to the full SciDTB corpus. We expect this to become a valuable resource in argument mining research in scientific texts which, as mentioned, has been identified as a particularly challenging domain.

The obtained results open several paths up for additional research, including the implementation of other transfer learning approaches–e.g., multi-task learning settings[20]–as well as other neural architectures–including attention-based architectures, which have proven to achieve good results in argument mining tasks (Stab et al., 2018). As mentioned in Section 3.1, we are also interested in exploring the possibility of leveraging other existing tools and resources to facilitate the automatic identification of argumentative structures and relations, such as corpora annotated with different schema–including variants of CoreSC and AZ. We also intend to expand our acceptance prediction experiments using the PeerRead dataset (Kang et al., 2018),[21] which has a greater coverage than the NIPS and ICLR subsets that we used in our experiments. This dataset contains, in addition to the acceptance/rejection decisions, scores for different aspects of the papers–including *substance* and *clarity*, among others–, which would allow us to explore in more depth whether the argumentative structure of the abstracts–and, potentially, other sections–relate to more specific quality aspects of the manuscripts.

[20] We conducted preliminary experiments in this area with mixed results, so we plan to continue investigating this approach in order to clarify its true potential.

[21] https://github.com/allenai/PeerRead

Acknowledgments

This work is (partly) supported by the Spanish Government under the María de Maeztu Units of Excellence Programme (MDM-2015-0502).

References

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 5(04):363–381.

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.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 1–17. Springer.

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 (ACL 2017)*, volume Volume 1: Long Papers, pages (11–22). Association for Computational Linguistics.

Valéria D. Feltrim, Simone Teufel, Maria Graças V. das Nunes, and Sandra M. Aluísio. 2006. *Argumentative Zoning Applied to Critiquing Novices' Scientific Abstracts*, pages 233–246. Springer Netherlands, Dordrecht.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *LREC*.

Nancy Green. 2015. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014*.

Ken Hyland. 1998. *Hedging in scientific research articles*, volume 54 of *Pragmatics Beyond New Series*. John Benjamins Publishing.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Alexandros Komninos and Suresh Manandhar. 2016. Dependency-based embeddings for sentence classification tasks. In *Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies (NAACL 2016)*, pages 1490–1500.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018a. ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining (ArgMining2018)*, pages 22–28.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018b. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining (ArgMining2018)*, pages 40–46.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014) (Volume 1: Long Papers)*, volume 1, pages 25–35.

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 Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers.

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

William C Mann, CMIM Matthiessen, and Sandra A Thompson. 1992. Rhetorical structure theory and text analysis. *Discourse description: Diverse linguistic analyses of a fund-raising text*, pages 39–78.

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

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Andreas Peldszus and Manfred Stede. 2015a. An annotated corpus of argumentative microtexts. In *Proceedings of the First Conference on Argumentation, Lisbon, Portugal*.

Andreas Peldszus and Manfred Stede. 2015b. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 938–948.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 103–112.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL Anthology Network corpus. *Language Resources and Evaluation*, 47(4):919–944.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. *arXiv preprint arXiv:1707.09861*.

David Soergel, Adam Saunders, and Andrew McCallum. 2013. Open scholarship and peer review: a time for experimentation. In *ICML Workshop on Peer Reviewing and Publishing Models (PEER)*.

Jonathan Sonntag and Manfred Stede. 2014. GraPAT: A tool for graph annotations. In *LREC*, pages 4147–4151.

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, Forlì-Cesena, Italy, July 21-25, 2014*, pages 21–25.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 3664–3674.

Manfred Stede, Stergos D Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *LREC*.

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 (EMNLP 2009) (Volume 3)*, pages 1493–1502. Association for Computational Linguistics.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Antje Timmer, Robert J Hilsden, and Lloyd R Sutherland. 2001. Determinants of abstract acceptance for the digestive diseases week–a cross sectional study. *BMC medical research methodology*, 1(1):13.

Antje Timmer, Lloyd R Sutherland, and Robert J Hilsden. 2003. Development and evaluation of a quality score for abstracts. *BMC medical research methodology*, 3(1):2.

Irvin Vargas-Campos and Fernando Alva-Manchego. 2016. Sciesp: Structural analysis of abstracts written in spanish. *Computación y Sistemas*, 20(3):551–558.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (ACL 2017) (Volume 1: Long Papers)*, pages 176–187.

Douglas N Walton and David N Walton. 1989. *Informal logic: A handbook for critical argument*. Cambridge University Press.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018) (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Longkai Zhang and Houfeng Wang. 2014. Go climb a dependency tree and correct the grammatical errors. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 266–277, Doha, Qatar. Association for Computational Linguistics.