Discourse-driven argument mining in scientific abstracts

Pablo Accuosto and Horacio Saggion

LaSTUS/TALN Research Group, DTIC
Universitat Pompeu Fabra
C/Tànger 122-140, 08018 Barcelona, Spain
{name.surname}@upf.edu

Abstract. Argument mining consists in the automatic identification of argumentative structures in texts. In this work we address the open question of whether discourse-level annotations can contribute to facilitate the identification of argumentative components and relations in scientific literature. We conduct a pilot study by enriching a corpus of computational linguistics abstracts that contains discourse annotations with a new argumentative annotation level. The results obtained from preliminary experiments confirm the potential value of the proposed approach.

Keywords: Argument mining · RST · Scientific corpus.

1 Introduction

Argument mining [18, 16]–the automatic identification of arguments, its components and relations in texts–, has recently gained increased interest in natural language processing and computational linguistics research both in the academia [16] and the industry [1]. Being able to automatically extract not only what is being stated by the authors of a text but also the evidence that they provide to support their claims would enable multiple applications, including argumentative summarization, computer-assisted text quality assessment, information retrieval systems, reasoning engines and fact-checking tools. The identification of argumentative units and relations in scientific texts, in particular, would enable tools that could contribute to alleviate the information overload experienced by researchers, editors and students as a consequence of the accelerated pace at which scientific knowledge is being produced [3]. Argument mining in scientific texts has, however, proven as a highly challenging task. This is, mainly, due to the complexity of the underlying argumentative structures of the scientific discourse [9]. These difficulties are not only faced when trying to develop automated systems but also in the production of gold standards with which to train those systems. It has been observed [7] that even humans with expert domain knowledge can find it difficult to unambiguously identify premises, conclusions and argumentation schemes in scientific articles. The lack of annotated corpora, in turn, represents a major barrier for advancing argumentation mining research in the scientific domain.

In this work we investigate the potential exploitation of existing linguistic resources in order to facilitate the annotation of argumentative components and relations in the domain of computational linguistics. We propose a fine-grained annotation schema particularly tailored at scientific texts which we use to enrich a subset of abstracts from the SciDTB corpus [33], which have been previously annotated with discourse relations from the Rhetorical Structure Theory (RST) [17]. RST provides a set of coherence relations with which adjacent spans in a text can be linked together in a discourse analysis, resulting in a tree structure that covers the whole text. The minimal units that are joined together in RST are called *elementary discourse units* (EDUs). Let us consider the following example from [32], included in the SciDTB corpus, in which EDUs are numbered and identified by square brackets:

> [Text-based document geolocation is commonly rooted in language-based information retrieval techniques over geodesic grids.]$_1$ [These methods ignore the natural hierarchy of cells in such grids]$_2$ [and fall afoul of independence assumptions.]$_3$ [We demonstrate the effectiveness]$_4$ [of using logistic regression models on a hierarchy of nodes in the grid,]$_5$ [which improves upon the state of the art accuracy by several percent]$_6$ [and reduces mean error distances by hundreds of kilometers on data from Twitter, Wikipedia, and Flickr.]$_7$ [We also show]$_8$ [that logistic regression performs feature selection effectively,]$_9$ [assigning high weights to geocentric terms.]$_{10}$

From the argument mining perspective, we would like to identify, for instance, that the authors support their claim about the *effectiveness of using regression models* for text-based document geolocation (EDUs 4-5) by stating that this method *improves upon the state of the art accuracy* (EDU 6) and it *performs feature selection effectively* (EDU 9), which in turn is supported by the fact that it *assigns high weights to geocentric terms* (EDU 10).

In this work we aim at exploring if the information provided by the discourse layer of the corpus, which establishes that these elements are linked by chains of discourse relations[1] can contribute to facilitate this task. With this objective we conduct a set of experiments aimed at the identification of argumentative structures in the abstracts, including their argumentative components, functions and attachment. As described in Section 5, we propose to learn each of these subtasks separately as well as together, in a multi-task framework. Multi-task learning is a way of transferring information between machine learning processes, so they can positively influence each other. Caruana [5] describes multi-task learning as a way of improving generalization when training a machine learning model, by taking advantage of information contained in the training signals of related tasks. In order to do this, the tasks are trained in parallel while using a shared representation (such as the hidden layers of a neural newtork). We propose to compare the performance of training the argument mining subtasks in a multi-task architecture in order to explore to what degree the natural connections

[1] For instance, EDUs 4 and 9 are linked by an *evaluation* relation, which can provide a clue for the identification of a *support* relation from the argumentative perspective.

between them is actually captured in the training process and reflected in an improved performance of the resulting models. This idea is in line with current research across multiple natural language processing problems. In particular, those that include the identification of units and relations as related tasks, as is the case of argument mining. In [24], Ruder provides a thorough overview of the current state of multi-task learning in the context of deep learning architectures.

Contributions. Our main contributions can therefore be summarized as:

- The proposal of a new, fine-grained, schema for the annotation of arguments in scientific texts;
- The first iteration in the development of a corpus of computational linguistic abstracts containing a layer of argumentative annotations (in addition to a previously existing discourse annotation layer);
- New evidence for analyzing the interplay between argumentative and discoursive components and relations and how existing tools and resources for discourse analysis can effectively be exploited in argument mining;
- Experimental results obtained by neural and non-neural architectures for mining arguments in scientific short texts;
- Additional elements to feed ongoing investigations on scenarios in which multi-task architectures have a positive effect over the independent learning of related tasks.

The rest of the paper is organized as follows: in Section 2 we briefly report previous work in the area. In Section 3 we describe the SciDTB corpus and in Section 4 our proposed annotation schema for new the argumentative layer. In Section 5 we describe our experimental settings and in Section 6 we report and analyze the results. Finally, in Section 7, we summarize our main contributions and propose additional research avenues as follow-up to the current work.

2 Related work

The Argumentative Zoning (AZ) model [30, 31] is a key antecedent in the identification of the discoursive and rhetorical structure of scientific papers. It includes annotations for knowledge claims made by the authors of scientific articles. In turn, the CoreSC annotation scheme [14] adopts the view of a scientific paper as a readable representation of a scientific research by associating research components to the sentences describing them. Our proposal for annotating argumentative units (described in Section 4), lies between CoreSC and AZ: while the set of annotation labels resembles that of CoreSC, they are intended to express argumentative propositions, as in the case of AZ. Unlike our proposal, neither AZ nor CoreSC consider relations between rhetorical units.

Due to the challenges posed by the identification of arguments in scientific texts, most of the previous works in argument mining are targeted at other textual registers (news, product reviews, online discussions). Lippi and Torroni [16] provide a thorough summary of initiatives in these areas. The corpus created by Kirschner et al. [9] was one of the first intended for the analysis of the argumentative structure of scientific texts. The authors introduce an annotation

schema that represents arguments as graph structures with two argumentative relations (*support*, *attack*) and two discourse relations (*detail*, *sequence*). Recently, Lauscher et al. [12] enriched a corpus of scientific articles with argumentative components and relations and analyzed the information shared by the rhetorical and argumentative structure of the documents by means of normalized mutual information (NMI) [29]. They then used the enriched corpus to train a tool (*ArguminSci*[2]) aimed at the automatic analysis of scientific publications, including the identification of claims, and citation contexts and the classification of sentences according to their rhetorical role, subjective information and summarization relevance.[11] Stab et al. [27] conducted preliminary annotation studies to analyze the relation between argument identification and discourse analysis in scientific texts and persuasive essays. In line with previous work [4, 2], the authors acknowledge the differences between both tasks (in particular, as discourse schema are not specifically aimed at identifying argumentative relations), while they also affirm that work in automated discourse analysis is highly relevant for argumentation mining, leaving as an open question how can this relation be exploited in practice.

Our work is inspired by that of Peldszus and Stede [20]. In this work, an annotation study of 112 argumentatively rich short texts using RST and argumentation schemes is produced. The authors provide a qualitative analysis of commonalities and differences between the two levels of representation in the corpus and report on experiments in automatically mapping RST trees to argumentation structures. The argumentative components that they consider are argumentative discourse units (ADUs), which consist of one or more EDUs of the RST scheme. They propose two basic argumentative relations: *support* and *attack*, further dividing attacks between *rebutal*s (denying the validity of a claim) and *undercut*s (denying the relevance of a premise for a claim). They also include a non-argumentative meta-relation (*join*) to link together EDUs that are part of the same argumentative unit. In their case the experiments are conducted at the discourse units level.[3] We, instead, propose our analysis at the level of the argumentative units (which can be formed by more than one EDU) and can therefore compare the results obtained with and without including explicit[4] discourse information in argument mining tasks.

3 SciDTB Corpus

The Discourse Dependency TreeBank for Scientific Abstracts (SciDTB) [33] is a corpus containing 798 abstracts from the ACL Anthology [22] annotated with elementary discourse units and relations from the RST Framework with minor adaptations to the scientific domain.

[2] `http://lelystad.informatik.uni-mannheim.de/`

[3] For instance, if given two EDUs they are connected by an argumentative relation.

[4] We have not generated annotations without previously segmented text, so the implicit effect of considering already available EDUs as building blocks is not analyzed in this work.

The SciDTB annotations use 17 coarse-grained relation types and 26 fine-grained relations. 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 [19, 28, 13], which makes it particularly suitable as input of sequence tagging algorithms.

4 Argumentation annotations

We propose an annotation schema for scientific argument mining and test it in a pilot study with 60 abstracts.[5] The annotation are made by means of an adapted version of the GraPAT [26][6] tool for graph annotations.

4.1 Relations

In line with [9], we adopt in our annotation scheme the classic ***support*** and ***attack*** argumentative relations and the two discourse relations ***detail*** and ***sequence***. In order to simplify both the creation and processing of the annotations we restrict the accepted argumentative structures to dependency trees.[7] To account for cases in which two or more units are mutually needed to justify an argumentative relation, we introduce the ***additional*** meta-relation. In this case the annotator chooses one premise to explicitly link to the supported or attacked unit while the rest are chained together by *additional* links. We observed that this restriction does not limit the expressiveness of the schema but, on the contrary, contributes to hierarchically organize the arguments according to their relevance or logical sequence.

4.2 Argumentative units

Previous works in argument mining [16] frequently use *claims* and *premises* as basic argumentative units. Due to the specificity of the scientific discourse in general [8], and abstracts, in particular, we consider this schema to be too limiting, as it does not account for essential aspects such as the degree of assertiveness and subjectivity of a given statement. We therefore propose a finer-grained annotation schema that includes the following set of classes for argumentative components: ***proposal*** (problem or approach), ***assertion*** (conclusion or known fact), ***result*** (interpretation of data), ***observation*** (data), ***means*** (implementation), and ***description*** (definitions/other information). While *proposal* could broadly be associated with claims, *result* and *observation* are in general used to provide supporting evidence. The units labeled as *assertion* can have a dual role of claim and premise and *means* and *description* are, in general, used to provide non-argumentative information.

In line with [20], and unlike previous works that consider sentences as annotation units [15, 30], we consider EDUs as the minimal spans that can be

[5] All of the abstracts are from papers included in the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

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

[7] Each argumentative unit can only have one argumentative function and is attached to one parent.

annotated, while there is not a pre-established maximum span. Argumentative units can cover multiple sentences.

Figure 1 shows a subset of the argumentative components and relations annotated in the abstract included in Section 1. The color of the units represent their type: yellow for units of type *result*, pink and red for *assertion*.[8]

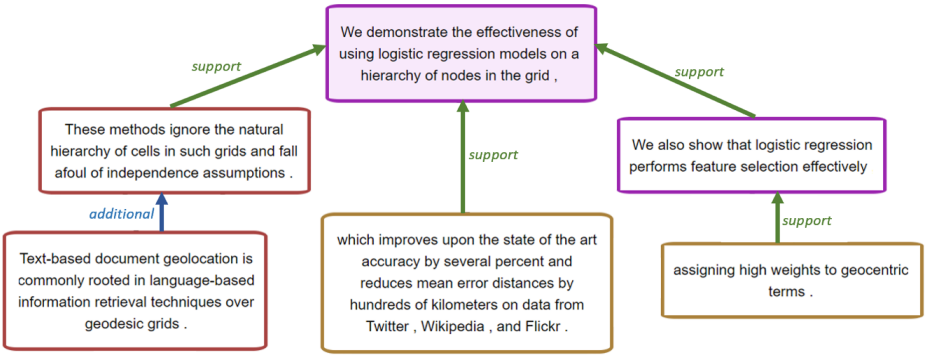

Fig. 1. Argumentative tree

The example shows how, in the case of scientific abstracts, claims and evidence provided to support them are frequently not stated explicitly in an argumentative writing style but are instead implicit.[9]

4.3 Argumentation corpus statistics

The corpus enriched with the argumentation level contains 60 documents with a total of 327 sentences, 8012 tokens, 862 discourse units and 352 argumentative units.[10] Even if not enforced by the annotation schema, argumentative unit boundaries coincide with sentences in 93% of the cases.

Table 1 shows the distribution of the argumentative units in relation to their type, argumentative function and distance to their parents.[11]

It is relevant to note that, while almost every document considered contains one or more *support* relations, there are no *attacks* identified in the set of documents currently annotated. We maintain the *attack* relation in our schema, nevertheless, as we plan to expand our work to longer scientific texts, where argumentative relations with different polarities are more likely to occur.

[8] Background assertions are displayed with a red border while assertions stated by the authors of the paper are displayed with a pink border.

[9] For instance, implicit claims in relation to the relevance of the problem at stake.

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

[11] According to the position of the parent unit, there are 200 relations pointing forward and 92 in which the parent appears before in the text.

Type		Function		Distance to parent	
proposal	110	*support*	124	*adjacent*	167
assertion	88	*attack*	0	*1 arg. unit*	55
result	73	*detail*	130	*2 arg. units*	36
observation	11	*additional*	27	*3 arg. units*	17
means	63	*sequence*	11	*4 arg. units*	11
description	7			*5 arg. units*	5
				6 arg. units	1

Table 1. Statistics of the corpus enriched with the argumentative layer.

5 Argument mining experiments

In this section we describe experiments conducted to assess the potential of discourse annotations for the extraction of argumentative structures (units and relations) in computational linguistics abstracts. We model all of these subtasks as sequence tagging problems, which allows us to compare the performance obtained by learning them separately as well as jointly, in a multi-task setting. We also compare the performance obtained when learning and evaluating these high-level tasks with neural and non-neural models.

5.1 Tasks

In order to capture the argumentative structure of a text it is necessary to identify its components and how they are linked to each other. The following set of interrelated tasks are aimed at this objective:

- **B (units)**: Identify the boundaries of the argumentative units.
- **Ty (types)**: Identify the types of the units (i.e.: *proposal*)
- **Fu (function)**: Identify the argumentative functions (i.e.: *support*).
- **Pa (attachment)**: Identify the position of the parent argumentative unit.

5.2 Experimental setups

We compare the results obtained for each of the tasks mentioned in Section 5.1 with and without considering rhetorical information available in the RST layer of the corpus. In each case, we run four different learning algorithms.

- **Baselines**: Basic classifiers, that exploit correlations between the abstracts' argumentative structure and rhetorical or syntactic information.
- **CRF**: Conditional random fields (CRF) tagger.
- **BiLSTM-ST**: A separate BiLSTM-CRF sequence tagger for each task.
- **BiLSTM-MT**: A BiLSTM-CRF sequence tagger with the four main tasks jointly trained in a mult-task setting.

In all of the settings the tasks are modeled as token classification problems where the argumentative units are encoded with the BIO tagging scheme. All the classifiers (including the baselines) are trained and evaluated in a 10-fold cross-validation setting.[12]

[12] For this pilot study the algorithms' hyperparameters–including the number of training epochs in the case of BiLSTM networks–were not optimized, as the main goal

Baselines. The result of the annotation process shows a correspondence of 93% between argumentative units and sentence boundaries. We therefore consider sentences as argumentative components for the implementation of the baseline algorithms. In order to predict argumentative functions, types and parents we generate simple classifiers based on the values of syntactic and/or discourse-level features and the classes to be predicted. We do this by mapping each value of the considered feature to its most frequent class in the training set. When no rhetorical information is included, we observed that the concatenation of the *lemma* of the syntactic root of the sentence and the sentence position is the best predictor in average (when all the tasks are considered). When rhetorical information is available, instead, the concatenation of the *discourse function* in which the token participates and the sentence BIO tag is the most predictive feature. It is relevant to note that this is a strong baseline to beat. For instance, the discourse relation predicts correctly the argumentative parent (Pa) for 57% of the argumentative units.[13]

CRF. For the CRF classifier we used Stanford's CRFClassifier [6] with unweighted attributes, including positional, syntactic and discourse features for the current, previous and next tokens.

- **Positional features**: sentence and EDU information: their position in the text, their boundaries and position of the token within them;
- **Syntactic features**: lemma, POS; dependency-tree parent; dependency-tree relation;
- **Discourse features**: discourse function and parent in the discourse tree.

BiLSTM. For the BiLSTM networks we used the implementation made by the Ubiquitous Knowledge Processing Lab of the Technische Universität Darmstadt [23].[14] We used one BiLSTM layer with 100 recurrent units with Adam optimizer and naive dropout probability of 0.25. For the parent attachment task (Pa) we included an additional BiLSTM layer of 100 recurrent units. We used a Softmax classifier as the last layer of the network, except in the case of the task B (boundary prediction), in which we used a CRF classifier. We used a batch size of 10 and trained the networks for 30 epochs for each task and each training-test split of our cross-validation setting. The tokens were encoded as the concatenation of 300 dimensional dependency-based word embeddings [10] and 1024-dimensional ELMo word embeddings [21]. In addition to the tokens, the BiLSTMs are fed with the same features used for the implementation of the baselines (sentence boundaries and position, lemma of the syntactic root and discourse function), which are encoded as 10-dimensional embeddings.

---

of this work was not to produce the best possible argument mining system but to obtain elements that would allow us to establish comparisons between the proposed approaches.

[13] In particular, for units that correspond to discourse roots–17% of all the units–the argumentative parent is predicted correctly 95% of the times.

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

6 Results and analysis

The experiments are evaluated with the ConNLL criteria for named-entity recognition. A true positive is considered when both the boundaries and class (type, function, parent) match. A post-processing filter is run in order to ensure that all the BIO-encoded identified units are well-formed. In the case of impossible sequences (e.g. an I tag without the preceding B), the labels are changed to the most frequent ones in the argumentative unit (the boundaries are considered to be correct). In the reported results the predicted boundaries are considered.

Algorithm	**Function (B+Fu)**		**Type (B+Ty)**		**Attachment (B+Pa)**	
	RST	**No RST**	**RST**	**No RST**	**RST**	**No RST**
Baseline	57.04	46.52	56.03	43.84	47.06	31.26
BiLSTM-MT	**71.88**	62.22	**68.78**	68.03	45.70	45.05
BiLSTM-ST	71.38	68.50	67.81	66.18	**47.90**	43.61
CRF	62.51	53.33	65.77	61.62	44.96	39.81

Table 2. F1-measures with and without discourse info.

Table 2 shows the F1-measure obtained in average for each task with and without discourse information, respectively. Explicitly incorporating discourse information significantly contributes to the identification of the argumentative function. It has also a more moderate but positive effect in predicting the argumentative units' types and attachment. It can be observed that, even with the limited amount of training data available and without optimizing their hyperparameters, the neural models perform considerably better than more traditional sequence labelling algorithms such as CRF. In particular, for the prediction of the argumentative units' functions. No strong conclusions can be drawn from these results with respect the advantage of training the tasks separately or jointly in a multi-task framework. Diverging results can be attributed to the different difficulty levels of the tasks and the small number of training examples. As more annotated data becomes available, we will be in a better position to explore how these tasks relate to each other and their mutual effect in a joint-learning setting. More experiments with different values for the hyperparameters and different regularization strategies need to be conducted in order to explore the effects of the inductive biases introduced by means of training several tasks in parallel. In this sense, we believe that it could be productive to explore other multi-task learning architectures, that account for the differences of difficulties of the various tasks, such as the hierarchical architecture proposed in [25].

6.1 Error analysis

In terms of the observed errors, we see the same patterns in all the experimental settings, with the numbers varying according to the respective performances of the systems. In particular, for the **Ty** task, the highest rate of errors are due to mis-classifying units of type *means* as either *proposal*, which accounts for 21%

of all the errors (in average), or *result*, which accounts for 11% of the errors. The mis-classification in the other direction: units of type *proposal* or *result* being mis-classified as *means* is less frequent but still significant, as it accounts for 9% and 5% of all the errors, respectively. Of significance are also the errors generated by the mis-classification of units of type *assertion* as either *proposal* or *result*, giving origin to 11% and 9% of all the errors produced in average by all the systems. In the case of the identification of the argumentative function (**Fu**), the main source of errors are due to the mis-classification between the classes *support* and *detail*, which accounts for 59% of all the errors (with roughly the same number of errors in both directions). In the case of the parent attachment task (**Pa**), the two most frequent errors are due to missing one argumentative unit (for instance, attaching a unit to the adjacent unit instead of the following one in the text), which accounts for 30% of all the errors and in assigning the wrong direction to the relation, which accounts for 35% of all the errors.

7 Conclusions

In this work we addressed the problem of identifying argumentative components and relations in scientific texts, a domain that has been recognized as particularly challenging for argument mining. We presented work aimed at assessing the potential value of exploiting existing discourse-annotated corpora for the extraction of argumentative units and relations in texts. Our motivation lies in the fact that discourse analysis, in general, and in the context of the RST framework, in particular, is a mature research area, with a large research community that have contributed a considerable number of tools and resources–including corpora and parsers–which could prove valuable for the advancement of the relatively newer area of argument mining. In order to test our hypothesis, we proposed and pilot-tested an annotation schema that we used to enrich, with a new layer of argumentative structures, a subset of an existing corpus that had previously been annotated with discourse-level information. The resulting corpus was then used to train and evaluate neural and non-neural models. Based on the obtained results, we conclude that the explicit inclusion of discourse data contributes to improve the performance of the argument mining models.

The results of this preliminary study are auspicious and motivate us to expand it. In particular, we aim at extending our argumentative layer of annotations to the full SciDTB corpus in an iterative process of semi-automatic annotation and evaluation. We believe that this enriched corpus would become a valuable resource to advance the investigation of argument mining in scientific texts. In order to identify arguments in un-annotated abstracts, we will also analyze the results obtained by training our models with discourse annotations obtained automatically, by means of existing RST parsers. In a complementary line, we will explore the potential offered by jointly learning to predict argumentative and discourse annotations in a multi-task environment. The models thus obtained could then be used to identify argumentative structures when no discourse annotations are available.

Acknowledgments

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

References

1. Aharoni, E., Dankin, L., Gutfreund, D., Lavee, T., Levy, R., Rinott, R., Slonim, N.: Context-dependent evidence detection (Jul 3 2018), US Patent App. 14/720,847
2. Biran, O., Rambow, O.: Identifying justifications in written dialogs by classifying text as argumentative. International Journal of Semantic Computing **5**(04), 363–381 (2011)
3. Bornmann, L., Mutz, R.: 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 (2015)
4. Cabrio, E., Tonelli, S., Villata, S.: From discourse analysis to argumentation schemes and back: Relations and differences. In: International Workshop on Computational Logic in Multi-Agent Systems. pp. 1–17. Springer (2013)
5. Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)
6. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. pp. 363–370. Association for Computational Linguistics (2005)
7. Green, N.: Identifying argumentation schemes in genetics research articles. In: Proceedings of the 2nd Workshop on Argumentation Mining. pp. 12–21 (2015)
8. Hyland, K.: Hedging in scientific research articles, vol. 54. John Benjamins Publishing (1998)
9. Kirschner, C., Eckle-Kohler, J., Gurevych, I.: Linking the thoughts: Analysis of argumentation structures in scientific publications. In: Proceedings of the 2nd Workshop on Argumentation Mining. pp. 1–11 (2015)
10. Komninos, A., Manandhar, S.: 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. pp. 1490–1500 (2016)
11. Lauscher, A., Glavaš, G., Eckert, K.: Arguminsci: a tool for analyzing argumentation and rhetorical aspects in scientific writing. Association for Computational Linguistics (2018)
12. Lauscher, A., Glavaš, G., Ponzetto, S.P.: An argument-annotated corpus of scientific publications. In: Proceedings of the 5th Workshop on Argument Mining. pp. 40–46 (2018)
13. Li, S., Wang, L., Cao, Z., Li, W.: Text-level discourse dependency parsing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 25–35 (2014)
14. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics **28**(7), 991–1000 (2012)
15. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.: Corpora for the conceptualisation and zoning of scientific papers (2010)
16. Lippi, M., Torroni, P.: Argumentation mining: State of the art and emerging trends. ACM Trans. Internet Technol. **16**(2), 10:1–10:25 (Mar 2016)

17. Mann, W.C., Matthiessen, C., Thompson, S.A.: Rhetorical Structure Theory and text analysis. Discourse Description: Diverse linguistic analyses of a fund-raising text **16**, 39–78 (1992)
18. Moens, M.F.: Argumentation mining: Where are we now, where do we want to be and how do we get there? In: Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation. pp. 2:1–2:6. FIRE '12 and '13, ACM, New York, NY, USA (2007)
19. Morey, M., Muller, P., Asher, N.: How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In: Conference on Empirical Methods on Natural Language Processing. pp. pp–1330 (2017)
20. Peldszus, A., Stede, M.: Rhetorical structure and argumentation structure in monologue text. In: Proceedings of the Third Workshop on Argument Mining. pp. 103–112 (2016)
21. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). vol. 1, pp. 2227–2237 (2018)
22. Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The ACL Anthology network corpus. Language Resources and Evaluation **47**(4), 919–944 (2013)
23. Reimers, N., Gurevych, I.: Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 338–348 (2017)
24. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
25. Sanh, V., Wolf, T., Ruder, S.: A hierarchical multi-task approach for learning embeddings from semantic tasks. arXiv preprint arXiv:1811.06031 (2019)
26. Sonntag, J., Stede, M.: Grapat: a tool for graph annotations. In: Proceedings of the 2014 The International Conference on Language Resources and Evaluation. pp. 4147–4151 (2014)
27. Stab, C., Kirschner, C., Eckle-Kohler, J., Gurevych, I.: Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In: ArgNLP. pp. 21–25 (2014)
28. Stede, M., Afantenos, S.D., Peldszus, A., Asher, N., Perret, J.: Parallel discourse annotations on a corpus of short texts. In: Proceedings of the 2016 The International Conference on Language Resources and Evaluation (2016)
29. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research **3**(Dec), 583–617 (2002)
30. Teufel, S.: Argumentative Zoning: Information extraction from scientific text. Ph.D. thesis, University of Edinburgh (1999)
31. Teufel, S., Siddharthan, A., Batchelor, C.: 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-Volume 3. pp. 1493–1502. Association for Computational Linguistics (2009)
32. Wing, B., Baldridge, J.: Hierarchical discriminative classification for text-based geolocation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar (Oct 2014)

33. Yang, A., Li, S.: Scidtb: Discourse dependency treebank for scientific abstracts. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 444–449 (2018)