Mining arguments in scientific abstracts with discourse-level embeddings

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

Argument mining consists in the automatic identification of argumentative structures in texts. In this work we leverage existing discourse-level annotations to facilitate the identification of argumentative components and relations in scientific texts, which has been recognized as a particularly challenging task. We propose a new annotation schema and use it to augment a corpus of computational linguistics abstracts that had previously been annotated with discourse units and relations. Our initial experiments with the enriched corpus confirm the potential value of incorporating discourse information in argument mining tasks. In order to tackle the limitations posed by the lack of corpora containing both discourse and argumentative annotations we explore two transfer learning approaches in which discourse parsing is used as an auxiliary task when training argument mining models. In this case, as no discourse information is used as input, the resulting models could be used to predict the argumentative structure of unannotated texts.

1 Introduction

The growing number of scientific publications and the shortening of the research-publication cycles makes it increasingly harder for authors, reviewers and editors to stay up-to-date with the state-of-the-art in their research fields [5]. Language-based tools –including semantic indexing and text mining tools– have been recognized by the scientific publishing community as valuable technologies to facilitate the discovery of scientific knowledge [8] but the analysis of scholarly reading patterns [55] and peer review processes [39] show that additional resources are needed to support the assessment of research articles' contributions.

The 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 [62]. In order to assess the contributions made in research articles it is necessary to identify the main claims made by the authors and to determine whether the evidence provided is strong enough to support them. 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. The automatic identification of arguments, its components and relations in texts is known as *argument mining* or *argumentation mining* [22].

Argument mining has increasingly being recognized as a relevant and challenging research area in natural language processing (NLP) and computational linguistics (CL) both in academia [31, 27] and the industry [3]. This is evidenced by the inclusion of the topic in the calls for papers of the main venues in the area, including the Annual Conference of the Association for Computational Linguistics (ACL),[1] the International Conference on Computational Linguistics (COLING),[2] and the Conference on Empirical Methods in Natural Language Processing (EMNLP),[3] as well as by the growing participation in the Argument Mining Workshop series (ArgMining), the premier research forum in the area, which is held annually at major NLP/CL conferences [52].

The potential applications of argument mining are multiple and diverse. Being able to extract what is stated by the authors of a text, their stance towards a particular issue and also the reasons that they provide to back up their claims can support multiple applications, ranging from fine-grained analysis of opinions to the generation of abstractive summaries, including argumentative-aware conversational search systems and decision-making support systems [60]. For instance, IBM's Project Debater[4] [45], which has been in development since 2012 and has recently gained great media interest, is aimed at developing a system that can debate humans on complex topics to help people build persuasive arguments and make well-informed decisions, with the ultimate goal of counteracting the rise of one-sided narratives, misinformation and superficial thinking.

The tasks involved in the extraction of arguments from texts -–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 -–e.g.: sequence labeling, text segmentation, entity recognition and relation extraction-– for which supervised learning methods have proven successful [27]. The lack of annotated data with argumentative information, however, presents a challenge when trying to apply these well-known approaches to argument mining [48]. This is so, in part, due to the inherent difficulty of unambiguously identifying argumentative elements in texts, which is reflected in the low levels of inter-annotator agreement reached in general for this task [14].

In this work we investigate the potential exploitation of existing linguistic resources to facilitate the identification of argumentative components and relations in the domain of computational linguistics. In particular, we propose to leverage existing discourse-level annotations, as previous works suggest potential benefits in linking discourse analysis and argument mining tasks [36, 49, 6, 4, 13].

[1] https://acl2020.org/calls/papers/

[2] https://coling2020.org/pages/call_for_papers

[3] https://https://www.emnlp-ijcnlp2019.org/calls/papers

[4] https://www.research.ibm.com/artificial-intelligence/project-debater/

We propose a fine-grained annotation schema particularly tailored at scientific texts which we use to enrich a subset of the SciDTB corpus [64], which includes abstracts that have previously been annotated with discourse relations based on the Rhetorical Structure Theory (RST) framework [28]. 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 [63], 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[5] can contribute to facilitate this task.

We conduct two sets of experiments aimed at the identification of argumentative structures in abstracts, including their argumentative units, functions and attachment between units. In the first set of experiments we consider the existing gold discourse annotations as features of both neural and non-neural machine learning systems and compare the results obtained in both scenarios. In the second set of experiments we explore the potentials offered by two inductive transfer learning approaches to embed discourse knowledge into argument mining models, so they can then be used to predict the argumentative structure of texts in which no discourse-level annotations are available.

1.1 Contributions

Our main contributions can be summarized as:

1. 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) annotations in a corpus of computational linguistics abstracts (SciDTB).

[5]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.

2. We propose and test a fine-grained annotation schema 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.

3. We compare the results obtained with three different sequence labeling algorithms trained with and without discourse-level information.

4. We explore the potential of two transfer learning approaches (multi-task learning and sequential learning) leverage all available discourse-level annotations in order to improve the performance of argument mining models trained with a small volume of annotated data.

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 Sections 3 and 4 we describe the dataset used in our experiments and our proposed annotation schema for fine-grained scientific argument mining, respectively. In Section 5 we present the argument mining subtasks considered in this work and in Section 6 we describe our first set of experiments and analyze the results obtained when considering discourse features in argument mining models. In Section 7 we describe the experimental settings and results of our second set of experiments based on two transfer learning approaches. Finally, in Section 8, we present our conclusions and suggest additional research avenues as follow-up to the current work.

2 Related work

The annotation schemas used in argument mining corpora are derived from theoretical proposals intended to formalize argumentative reasoning, such as Toulmin's model of arguments [58]. Toulmin's model, which describes the different parts necessary in a well-formed argument (*claim*, *data*, *warrant*, *qualifier*, *rebuttal*, *backing*), has been adapted in several ways according to needs emerging in different areas. Particularly relevant for the computational analysis of arguments is the work of Walton et al. [61], in which different types of argumentation schemas are proposed. These schemas were put into use in the *Araucaria* system [41], aimed at the identification and visualization of the structure of arguments in terms of their constituents and the relationships between them, thus facilitating the development of argument mining corpora. Various annotation efforts have been done since then to identify argumentative components in texts from different domains. Lawrence and Reed [22] and Lippi and Torroni [27] provide thorough analyses of argument mining initiatives in various domains, including legal documents [30], online discussions [12], Wikipedia articles [3], newspapers [11], student essays [47] and television debates [59]. Some works even consider sources from several registers [15] and domains [44]. But very few initiatives are aimed at the identification of arguments in scientific texts. The inherent complexity and argumentative ambiguity of the scientific language has made this task particularly challenging, as illustrated by the results obtained by the small number of works in this area mentioned below in this section [49, 18, 13].

The most relevant antecedents aimed at generating resources for the automatic identification of rhetorical and argumentative components in scientific texts include the Argumentative Zoning (AZ) model [57] and the CoreSC schema [24]. AZ, which was later

extended as AZ-II [56], was used to annotate a corpus of 61 chemistry articles. This schema includes annotations (such as *Novelty or advantage*) for knowledge claims made by the authors of the papers and others (such as *Support*) that allow to establish connections with previous works by the same or other authors. CoreSC, in turn, associates research components to the parts of the texts describing them, thus obtaining a readable representation of the research process described by the paper, including categories such as *Motivation* (to describe the reasons behind an investigation) or *Result* (to describe statements about the outputs of an investigation). The CoreSC annotation schema was used to construct a corpus of 265 annotated papers from physical chemistry and biochemistry [25]. Differences and similarities between AZ and CoreSC are studied in [26], where correlations between both annotation schemas are analyzed. The authors also discuss the benefits of combining rhetorical-based and content-based analysis of the papers by applying the two schemas to the same set of documents.

Both AZ and CoreSC are sentence-based schemas focused on the identification of the components and do not consider the relations between them. The corpus created by Kirschner et al. [18] was one of the first intended for the analysis of the argumentative structure of scientific texts, considering not only argumentative components but also how they are linked to each other. In this work, the authors introduce an annotation schema that represents arguments as graph structures with two argumentative relations (*support*, *attack*) and two discourse relations (*detail*, *sequence*), which is used to annotate the introduction and conclusion sections of 24 German scientific articles in the educational domain. Although the annotators were familiar with the domain, the inter-annotator agreement achieved in this experiment was low (a Cohen's *kappa* coefficient of 0.43), thus corroborating claims made by other authors in relation to the challenges involved in the annotation of arguments in scientific texts [49, 13]. It is relevant to note, nevertheless, that the evaluation of argument annotations is still an open issue and that traditional agreement scores might not properly reflect the reliability of the annotations. In fact, multiple annotations of the same text might reflect different interpretations of the authors' intentions and could therefore be considered as fully or partially correct. Stab et al. [49] suggest that it might be interesting to explore, to evaluate argument annotations, methods that are able to deal with multiple correct annotations such as those used in text summarization. In the case of the work reported in [18], a novel graph-based measure was developed that makes it possible to consider different annotations with similar meaning, thus obtaining higher agreement scores than those observed with standard measures.

As mentioned, few other initiatives have been aimed at the annotation of arguments in scientific texts. Lauscher et al. [21] carried on one of these initiatives in the area of computer graphics papers. In their work, they enriched the DrInventor Scientific Corpus [10] with an argumentation layer. The original DrInventor corpus contains 40 documents annotated with four layers, including citation contexts, rhetorical role of sentences, subjective information and summarization relevance. Lauscher et al. first performed a normalized mutual information (NMI) analysis [53] of the information shared by the rhetorical and argumentative annotation layers and then used the new annotations to train a model for the automatic identification of claims and evidence [20].

Previous works have postulated correspondences between argumentative relations and other types of discourse relations. Stab et al. [49] conducted preliminary annotation stud-

ies to analyze the relation between argument identification and discourse analysis in both scientific texts and persuasive essays. In line with previous work [6, 4], the authors acknowledge the differences between both tasks –in particular, as discourse schemas are not specifically aimed at identifying argumentative relations– but 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. Green [13] highlights differences between evidence-based arguments and discourse relations, as those in RST, or rhetorical roles of sentences labeled according to Teufel's model –indicating they have different objectives and one type of relationship is not subsumed under the other– but she also suggests that information provided by these frameworks can be useful in the automatic extraction of arguments from scientific texts. Our proposal to leverage discourse information in argument mining tasks is more directly inspired by research conducted by Peldszus and Stede [36, 34, 51], who annotated 112 argumentatively rich texts using RST and argumentation schemas 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 [35]. The argumentative components that they consider are *argumentative discourse units* (ADUs), which consist of one or more EDUs of the RST schema. They propose two basic argumentative relations: *support* and *attack*, further dividing attacks into *rebuttal*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.[6] We, instead, propose our analysis directly at the level of the argumentative units and can therefore use them in experiments without considering discourse information.[7]

This work is also informed by research in the areas of transfer learning and multi-task learning. Pan and Yan [33] provide a survey for transfer learning methods including inductive, transductive and unsupervised approaches, while Ruder [43] presents an in-depth exploration of the application of neural transfer learning to natural language processing.

This paper extends previous works [1, 2] in which we explored the potential advantages of including discourse knowledge when training argument mining tasks .

3 SciDTB Corpus

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) [64]. SciDTB contains 798 abstracts from the ACL Anthology [40] annotated with elementary discourse units (EDUs) and relations from the RST framework with minor adaptations to the sci-

[6] E.g.: determine if, given two EDUs, they are connected by an argumentative relation.

[7] We have not generated annotations with unsegmented text, so the implicit effect of considering already available EDUs as building blocks is not analyzed in this work.

entific domain. The SciDTB annotations use 17 coarse-grained relation types and 26 fine-grained relations. Poly-nary 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 [32, 51, 23], which makes it particularly suitable as input of sequence tagging algorithms.

Fig. 1 shows an example of discourse units and relations annotated in an abstract[8] included in SciDTB and their corresponding argumentative units and relations. While the original annotation contained seven EDUs, at the argumentative layer they are combined into three ADUs: one *proposal* and two supporting units: an *assertion* and a *result*.

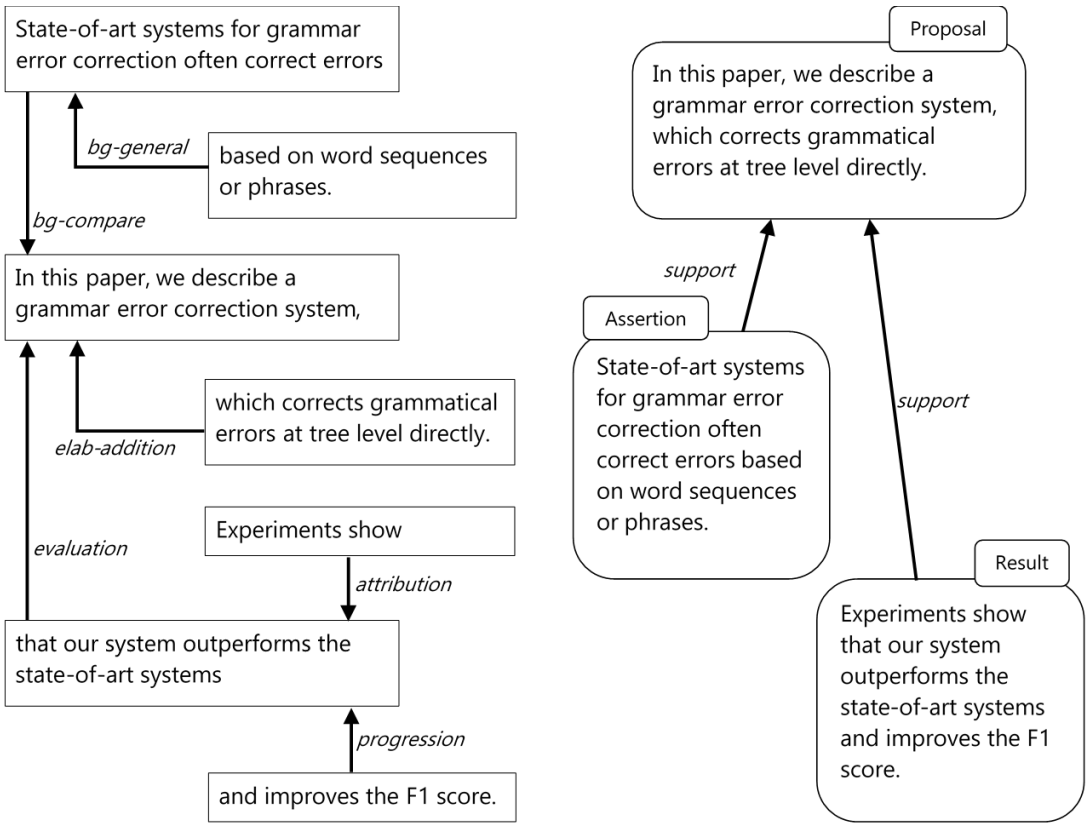

Figure 1: Discursive and argumentative structures.

4 New argumentation layer

In this section we propose a new annotation schema for scientific argument mining, which we tested in a pilot study with 60 abstracts from SciDTB.[9] The annotations of the pilot experiment were produced by means of an adapted version of the GraPAT [46][10] tool for graph annotations. The corpus enriched with the argumentation layer is made available to download.[11]

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

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

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

[11] http://scientmin.taln.upf.edu/argmin/scidtb_argmin_annotations.tgz

4.1 Argumentative units

Previous works in argument mining [27] frequently use *claims* and *premises* as basic argumentative units. Due to the specificity of the scientific discourse in general, 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. In the case of scientific discourse it is frequent to find that claims are not explicitly stated in an argumentative writing style but are instead left implicit [16]. The description of the problem addressed in the paper, for instance, can be considered to convey an implicit claim in relation to the relevance of the problem at stake and/or the adequacy of the proposed approach. We introduce a fine-grained annotation schema 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 considered in our schema can be mapped –even if with a different level of granularity– to concepts in CoreSC [26] and AZ categories, which would enable additional research on the potential of leveraging annotated corpora for argument mining tasks.

Like [36] –and in contrast with CoreSC and AZ– we consider EDUs as the minimal spans that can be annotated as an argumentative unit, while there is not a pre-established maximum span. Argumentative units can, in turn, cover multiple sentences.

Our schema 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)
- ***description*** (definitions/other information)

While *proposal* could broadly be associated with claims, units of type *result*, *observation* are, in general, used to provide supporting evidence. The units labeled as *means* and *assertion* can be used to introduce claims or premises, depending on the unit to which they are attached (e.g., a *means* unit attached to a *proposal* could provide additional information about the proposed solution and its implementation while, when attached to a unit of type *result*, it can provide additional evidence for the validity of the results). Units of type *description* are used to provide non-argumentative information.

It can be considered that our annotation schema for argumentative components 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. As mentioned, unlike our proposal, neither AZ nor CoreSC consider relations between units.

4.2 Relations

In order to simplify both the creation and processing of the annotations we restrict the form of valid argumentative graphs to trees. This means that each argumentative unit can only have one argumentative function and is linked to another one by a directed relation. We say that the child unit is *attached to* the parent unit and refer to the relation as the child unit's *argumentative function*. 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 the relevance and logical sequence of its constituents.

In line with previous work in the area [18], we consider in our annotation schema ***support*** and ***attack*** argumentative relations. We also include the relations ***detail*** and ***additional***, which are used to link a child unit to a parent unit for which it provides optional and required background information, respectively. These relations are therefore used to complement and/or contextualize the parent unit. Finally, the discursive ***sequence*** relation is also included in the schema (with the same meaning as in RST).

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.[12] Table 1 shows the distribution of the argumentative units in relation to their type, argumentative function (relation) and distance to their parents.[13] Even if not enforced by the annotation schema, argumentative unit boundaries coincide with sentences in 93% of the cases.

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.

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 keep 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.

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

[13]According to the position of the parent unit, the units that occur after its parent are aproximately the double than the units in which the parent occurs after in the text.

5 Argument mining tasks

In the following section we describe the experiments conducted to assess the potential of discourse annotations for the extraction of argumentative structures (units and relations) in computational linguistics abstracts.

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:

- **ATy (argumentative unit)**: Identify the boundaries and types of the units. In the example in Fig. 1, it would imply to identify the first and last token of each of the three units and their types: *proposal*, *assertion* and *result*.
- **AFu (argumentative function)**: Identify the boundaries[14] and argumentative functions of the components. In the example, the two *support* relations that link the children to the parent unit. Root nodes are assigned a dummy *root* relation.
- **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 one unit in the background direction. In our annotations these relations are therefore assigned attachment values of *1* and *-1*, respectively, while the root node is assigned a value of *0* to represent that it has no parent.

In Section 6 we analyze the performances obtained when human discourse annotations are incorporated as features into the models and, in Section 7, when the discourse information is transferred to the argument mining models by means of two different inductive transfer learning methods: multi-task learning and sequential learning.

6 Discourse features experiments

6.1 Experimental setups

The three argument mining tasks (ATy, AFu, APa) are modeled independently as sequence labeling problems at the token level where the argumentative units are encoded with the BIO tagging scheme. A post-processing filter is run in order to ensure that all the BIO-encoded identified units are well-formed.[15]

For this set of experiments we consider positional features (**Pos**)[16] and syntactic features (**Syn**),[17] and compare the results obtained for each of the tasks with and without

[14] In all cases the identification of the boundaries is considered as part of the task, as the three tasks are trained and evaluated independently.

[15] In particular, if an I tag is found in which the label does not match the one in the preceding B tag, the label is changed to the most frequent one in the argumentative unit.

[16] Position of the token in the sentence.

[17] Tokens' lemmas and parts-of-speech, syntactic function and parent in the dependency tree.

including discourse features obtained from the gold annotations available in the RST layer of the corpus (**Disc**).[18]

6.2 Algorithms

In order to compare the impact of training the argument mining models with discourse information, we consider three different algorithms: **majority-class** classifiers, conditional random fields (**CRF**) and bi-directional long short-term memory neural networks (**BiLSTMs**).

The **majority-class classifiers** are based on the correlations between syntactic, positional and discourse-level features and the class to be predicted for each token. We do this by mapping values obtained by combining multiple features to the most frequent class in the training set. We compare results obtained when considering only syntactic and positional features to those in which features from the RST annotations are also included. When no RST information is taken into account, the most predictive combination of features is the concatenation of the lemma of the syntactic root of the sentence and the position of the token being considered. On the other hand, when rhetorical information is considered, the concatenation of the discourse function in which the token participates and the position of the token in the sentence is the combination the better predicts the token's class. It is relevant to note that these are strong baselines to beat. For instance, the discourse function predicts correctly the argumentative parent (APa) for 57% of the argumentative units.[19] As mentioned, in addition to the type, function and parent we want to predict the units' boundaries. Given the high level of coincidence between argumentative units and sentences the majority classifiers are set to always predict the sentence boundaries as the boundaries of the argumentative units.

For the **CRF** classifiers we use Stanford's implementation [9] with un-weighted attributes, including positional, syntactic and discourse features for the current, previous and next tokens. The following parameters are set to true: *useClassFeature, useWord, useTags, useNGrams, noMidNGrams, useDisjunctive, usePrev, useNext, useSequences, usePrevSequences, , useTypeSeqs, useTypeSeqs2, useTypeySequences.* Additionally, we configured the classifier with: *maxNGramLeng=6, maxLeft=1, wordShape=chris2useLC.* This makes additional features for character n-grams and word shapes to be automatically generated and added when the algorithm is executed. Please refer to the documentation[20] for an explanation of each of these parameters.

For the experiments with **BiLSTM** networks, we use the implementation made available by the Ubiquitous Knowledge Processing Lab of the Technische Universität Darmstadt [42]. In order to simplify the experiments and the interpretation of their results we use the same architecture for the three tasks (AFu, ATy, APa): two layers of 100 recurrent units, Adam optimizer [17], a naive dropout probability of 0.25 and a CRF classifier the last layer of the network.

[18] Discourse relation (function), parent in the discourse tree, position of the token in the EDU.

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

[20] `https://nlp.stanford.edu/software/CRF-NER.shtml`

As with the previous algorithms, the tokens are tagged using the beginning-inside-outside (BIO) tagging scheme. Each token is encoded as the concatenation of 300-dimensional dependency-based word embeddings (DEmb)[21] ($\vec{k}$) [19] and 1024-dimensional contextualized word embeddings (ELMo)[22] ($\vec{e}$) [37].

For the experiments including discourse information the features obtained from the RST-level annotations (Disc) are also included as input, encoded as 40-dimensional dense vectors ($\vec{r}$).

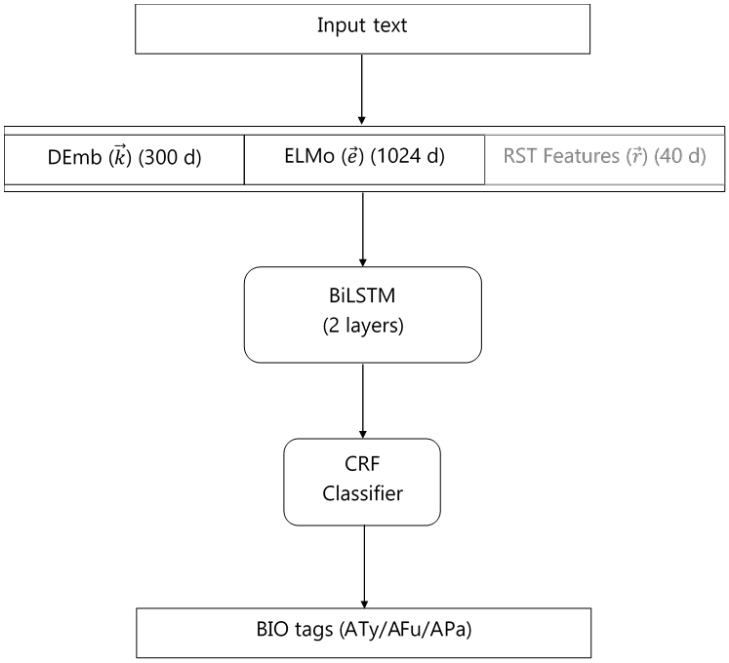

Figure 2: BiLSTM architecture with and without additional discourse features.

As our intention is not necessarily to obtain the best possible models for these tasks but, instead, to compare the different approaches and to analyze the potential impact of including discourse-level information when training argument mining models, no hyper-parameter optimization is done in these experiments and, in all of the cases, the networks are trained for 100 epochs.

6.3 Results

The experiments are evaluated with the CoNLL criteria for entity recognition: a true positive is considered when both the predicted boundaries and class (type, function, parent) match the gold annotations.

In all of the cases the classifiers are trained and evaluated in a 10-fold cross-validation setting.

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

[22] In these experiments we use the 5.5 billion-token version of ELMo trained with Wikipedia and monolingual news from the WMT 2008-2012 corpora, available from `https://allennlp.org/elmo`

Algorithm	Features	AFu		ATy		APa	
		Avg. F1	σ	**Avg. F1**	σ	**Avg. F1**	σ
Majority	Syn,Pos	46.52	3.54	43.84	10.69	31.26	6.29
Majority	Syn,Pos,Disc	**57.04**	7.87	**56.03**	8.14	**47.06**	9.89
CRF	Syn,Pos	53.33	17.53	61.62	12.21	39.81	15.42
CRF	Syn,Pos,Disc	**62.51**	10.54	**66.04**	15.42	**44.96**	7.61
BiLSTM	No Feat.	69.94	6.30	66.94	8.82	41.74	10.43
BiLSTM	Disc	**71.07**	8.51	**69.72**	8.70	43.23	10.17
BiLSTM	Syn,Pos	68.45	4.22	67.91	9.84	42.95	9.06
BiLSTM	Syn,Pos,Disc	70.02	5.40	69.67	9.07	**43.39**	10.66

Table 2: Average F1-measures and standard deviations obtained with and without discourse (Disc), syntactic (Syn) and positional (Pos) features for the different types of classifiers in 10-fold cross-validation settings.

Table 2 shows the average F1-measures obtained for each task in our cross-validation setting, with and without discourse information, respectively. In the case of the BiLSTM networks we consider the average F1-measures obtained in epochs 11 to 100 for each of the 10 partitions.[23].

The results show that, in all cases and independently of the type of classifier considered, explicitly incorporating discourse information significantly contributes to the identification of the argumentative functions and it has a positive effect in predicting the argumentative units' types and attachment.

Even if this set of experiments are not aimed at exploring differences in performance of the different architectures but, instead, to analyze the effect of explicitly incorporating discourse information in each type of classifier considered, it can be observed that the neural models seem to perform, in general, better than more traditional sequence labelling algorithms such as CRF –even with the limited amount of training data available and without optimizing their hyper-parameters. This is particularly clear in the prediction of the argumentative units' functions. In the case of the argumentative parents, the larger number of potential categories and limited training data do not allow the more complex algorithms to beat the majority classifiers which, as said, establish ambitious baselines. The small difference in the averaged results with respect to their standard deviations, nevertheless, does not allow us to draw definitive conclusions with respect to the preferred algorithms for the prediction of argumentative types and parent attachments.

It is also interesting to observe that, in the case of the BiLSTMs, the explicit incorporation of syntactic (Syn) and positional (Pos) features does not seem, in general, to contribute to improve the performance of the models. This is in line with research that shows that contextualized embeddings, such as ELMo, already capture rich linguistic knowledge. [38]

The experiments described in this section confirm our initial hypothesis in relation to the potential of leveraging RST annotations for argument mining in scientific texts.

[23]We exclude epochs 1 to 10 since in the first iterations the algorithms have not had time to learn anything and therefore the results are not significant

But, as mentioned, to implicitly include discourse annotations as input features poses two important limitations. The first one is faced when training the models, as only the subset of texts containing both levels of annotations can be used. In our case this means that only a small fraction of the SciDTB corpus-the one containing argumentation annotations-can be leveraged. The second limitation refers to the applicability of the resulting models, as they could only be used to identify argumentative structures in texts for which RST annotations are available. Even if it would be possible to use a discourse parser as the first step of a pipeline, the resulting system would be dragging errors made in the final steps of the classification of the discourse relations. In Section 7 we explore alternative approaches to tackle these limitations.

6.4 Error analysis

In this section we analyze the most frequent sources of errors observed for each task. We consider, in each case, the results obtained for the best setting (BiLSTM with discourse features in the case of AFu and including also syntactic and positional features in the case of ATy and APa), but the same patterns of errors are observed in all the experimental settings, with numbers varying according to the respective performances of the systems.

Once the boundaries are corrected, the mis-classification of a B token with an I tag or an I token with a B tag are not significant (approximately 0.03% of the cases). For the sake of clarity and brevity we therefore report here only the errors produced in the classification of B tokens. Tables 3, 4 and 5 show the percentage of errors observed for the labels considered in tasks AFu, ATy and APa, respectively.

Relation	*additional*	*detail*	*none*	*sequence*	*support*
additional	-	1	0	0	6
detail	3	-	6	0	**25**
none	0	1	-	0	0
sequence	0	**14**	0	-	0
support	6	**36**	1	0	-

Table 3: Percentage of errors produced for pairs of AFu classes over total.

In the case of the identification of the argumentative function (AFu), the main source of errors are due to the mis-classification between the classes *support* and *detail*, which accounts for 61% of all the errors. Also significant is the mis-classification of relations of type *sequence* as *detail*, which happens systematically and gives raise to 14% of all the errors.

For the ATy task, the highest rate of errors are due to mis-classifying units of type *means* as *proposal*, which accounts for 23% of all the errors. The mis-classification in the other direction: units of type *proposal* being mis-classified as *means* is also significant, as it accounts for 15% of all the errors.

In Table 5 we show the errors that occur in the most frequent cases for the parent attachment task (APa).[24] In general, the root node (represented with a value of 0) is correctly identified (as can be observed also in Table 3). The most frequent errors in are

[24]As mentioned in Section 4.3, longer distances between a node and its parent are very infrequent.

Type	*assertion*	*description*	*means*	*observation*	*proposal*	*result*
assertion	-	0	4	0	5	7
description	5	-	0	0	3	1
means	5	0	-	0	**23**	3
observation	0	0	3	-	0	4
proposal	4	0	**15**	0	-	4
result	5	0	7	1	0	-

Table 4: Percentage of errors produced for pairs of ATy classes over total.

Rel. distance	*-4*	*-3*	*-2*	*-1*	*0*	*1*	*2*	*3*	*4*
-4	-	0	1	5	1	3	1	0	0
-3	0	-	3	**11**	0	1	1	0	0
-2	0	3	-	**14**	1	2	0	0	0
-1	0	3	8	-	0	9	1	0	0
0	0	1	1	0	-	0	0	0	0
1	0	1	3	8	0	-	0	0	0
2	0	0	1	1	0	0	-	0	0
3	0	1	1	2	1	1	0	-	0
4	0	0	1	1	0	0	0	0	-

Table 5: Percentage of errors produced for pairs of APa distances up to 4 over total.

due to the mis-classification of units with one or two units between it and its parent (-2 and -3 in the table). As mentioned in Section 4.3, the relations in the forward direction are less frequent (aproximately half) than relations pointing backwards. Therefore, it is not surprising that a greater number of errors are observed in these cases.

7 Transfer learning experiments

In this section we describe two inductive transfer learning [33, 43] approaches in which the prediction of discourse functions and parents are used as auxiliary tasks to learn intermediate representations used in the argument mining models. In particular, we describe, in section 7.1, a multi-task architecture in which each argument mining task is learned in parallel with a discourse parsing task and, in section 7.2 we present a sequential learning approach in which we use, to train argument mining models, representations obtained from encoders trained with discourse annotations.

The discourse parsing tasks considered are:

- **DFu (discourse function)**: Identify the boundaries and discourse roles of the EDUs (*attribution*, *evaluation*, *progression*, etc.).[25]
- **DPa (discourse attachment)**: Identify the boundaries of the EDUs and the relative position of the parent units in the RST tree.

In order to avoid the possibility of introducing an unintended bias in the discourse parsing models the discourse tasks (DFu and DPa) are trained with the 738 abstracts

[25]Please refer to [64] for a description of the discourse roles used to annotate the RST layer of SciDTB.

left in the SciDTB corpus when excluding the 60 abstracts that are used to train and evaluate the argument mining models.

The argument mining tasks considered are the same ones described in Section 5. They are each paired to a discourse parsing task for the transfer learning experiments. Previous works show that the task similarity is a key factor to consider when implementing transfer learning approaches [65, 43]. We therefore pair each argument mining task to the most similar discourse parsing one: while AFu and ATy are paired with DFu, APa is paired with DPa.

7.1 Multi-task learning

Multi-task learning is a particular way of transferring information between machine learning processes so they can positively influence each other. Caruana [7] 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. This can be useful when it is not possible or practical to use some features as inputs but they can be used, instead, as outputs of auxiliary tasks that are trained in parallel with the main task while using a shared representation. Neural networks are particularly well-suited architectures to do this as some layers of the networks can easily be shared by multiple tasks. This is the approach the we adopt. The resulting architecture is illustrated in Fig. 3.

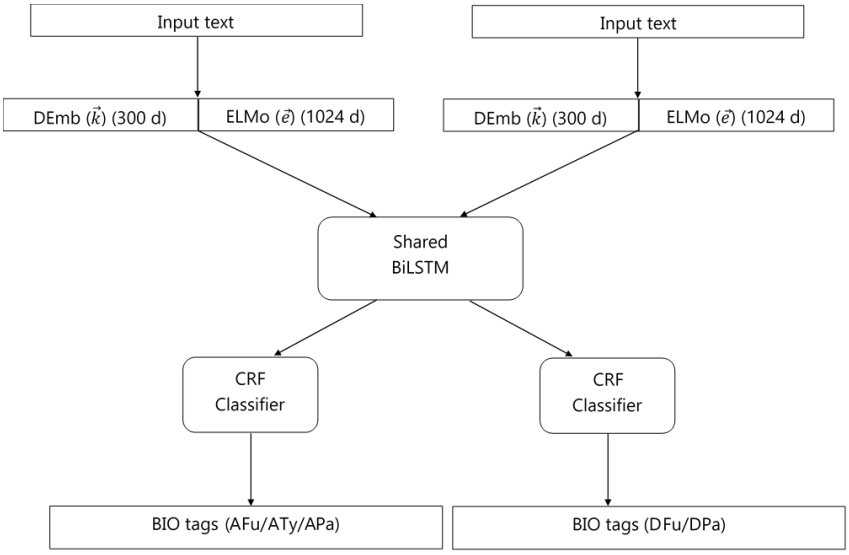

Figure 3: Multi-task argument mining / discourse parsing architecture.

7.2 Sequential transfer learning

An alternative method of transferring knowledge learned from one task to another is to train the corresponding models sequentially, instead of in parallel as in the case of multi-task learning. This can be particularly useful in cases, such as ours, where there is significantly more data available to train the source task that the target task. We adopt for this method the term *sequential transfer learning* as proposed by Ruder [43].

In this case we implement one of the most extended inductive transfer learning methods: we first train models with the RST annotations available in the SciDTB corpus which are then used to produce contextualized representations of the input tokens that are fed to the argument mining models. We refer to the models trained with the discourse-level tasks as *RST encoders* (RSTEnc).

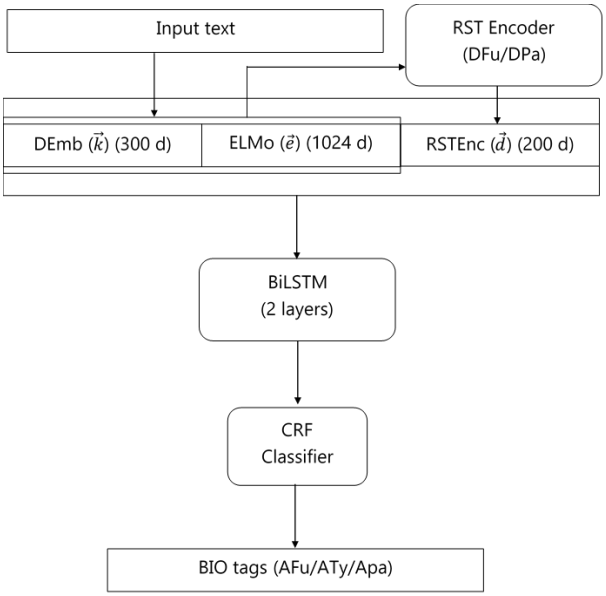

Figure 4: Sequential transfer learning architecture.

The RST encoders produce 200-dimensional embeddings ($\vec{d}$) obtained from the concatenation of the backward and forward hidden states of the top layers of the DFu or DPa models. These are the representations that we then use in the argument mining models as shown in Fig. 4.

7.3 Results

In order to evaluate the performances obtained in the identification of argumentative components and relations we use the same criteria adopted for the first set of experiments.

Table 6 shows the average F1-measures obtained for each of the settings: the argument mining models trained without discourse information (AM), in a multi-task setting with the discourse parsing tasks (AM+DP) and using the RST encoders in a sequential transfer setting (DP,AM).

Method	**Input**	**AFu**		**ATy**		**APa**	
		Avg. F1	σ	**Avg. F1**	σ	**Avg. F1**	σ
Single-task	$[\vec{k}, \vec{e}]$	69.94	6.30	66.94	8.82	41.74	10.43
Multi-task	$[\vec{k}, \vec{e}]$	67.38	6.90	65.65	12.39	40.69	9.98
Seq. transfer	$[\vec{k}, \vec{e}, \vec{d}]$	**70.98**	7.17	**70.38**	8.39	**43.44**	11.16

Table 6: Average F1-measures and standard deviations with and without information transferred from discourse parsing tasks.

The argument mining models trained with the representations produced by the RST encoders in a sequential transfer setting yield better performances, for all tasks, over the models trained solely with the dependency-based and ELMo embeddings. More relevant, these models perform at least as well as the neural models in which gold discourse annotations are used as input features (Table 2), with the advantage that it is possible to apply these models to predict the argumentative structure of texts for which gold RST annotations are not available.

These results support out initial hypothesis in the sense that transferring knowledge learned in discourse parsing tasks can in fact contribute to improve the performance of argument mining models trained with a rather small number of instances. In particular, this is the case when what is transferred are pre-trained representations in a sequential transfer setting.

When jointly training argument mining and discourse parsing tasks, in contrast, the results obtained for the argument mining models are worse than those obtained when the models are trained in single task settings. This effect, known as *negative transfer*, is not uncommon in multi-task settings [29]. In fact, multi-task learning architectures are known to be sensitive to a large number of parameters, including the distribution of the labels, the sizes of the respective datasets and the sampling strategies implemented in order to select the mini-batches when switching between tasks [54].

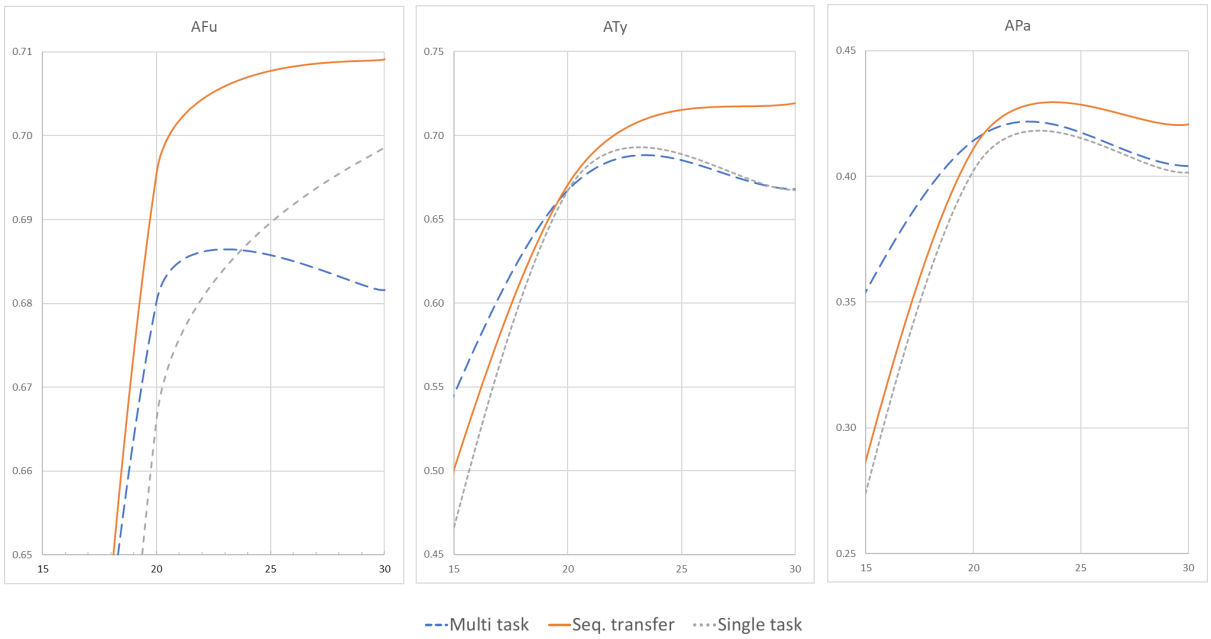

Figure 5: Trendlines of F1-scores in the first 30 epochs for AFu, ATy, APa, respectively

We understand that, due to the small size of our argument mining dataset, the regularization effect introduced by the auxiliary discourse parsing tasks is too strong and therefore affects the performance of the main tasks. This can be observed when analyzing the evolution of the performance of the argument mining tasks in the early training stages. Fig. 5 shows trendlines obtained considering average F1-scores obtained in the first 30 epochs for each argument mining task. It can be observed that the transferring of

knowledge is produced in the very early training stages. While in the sequential learning context the effect of including information conveyed by the RST encodings continues impacting positively the performance of the argument mining models in subsequent epochs (as evidenced by the better overall performances when considered epochs up to 100), in the multi-task setting this initial effect is rapidly counterbalanced by the excessive regularization introduced by the auxiliary tasks in the context of a very small training dataset.

7.4 Error analysis

In this section we present the errors observed in the best transfer learning scenario (sequential transfer) for ATy, AFu and APa in Tables 7, 8 and 9, respectively. As in Section 6.4, we report the mis-classification of beginning (B) tokens considering the percentage of errors for each pair of classes over the total.

Type	*assertion*	*description*	*means*	*observation*	*proposal*	*result*
assertion	-	1	6	0	9	**10**
description	4	-	0	1	3	0
means	1	3	-	0	**25**	4
observation	0	0	3	-	0	4
proposal	4	0	7	0	-	1
result	1	0	4	1	1	-

Table 7: Percentage of errors produced for pairs of ATy classes over total.

For the ATy task (Table 7), we observe that the most frequent errors occur due to the mis-classification of units of type *means* as *proposal*. This is also one of the most frequent sources of errors in the experiments using RST gold features. A difference between both results is that in this case the mis-classification in the reverse direction (units of type *proposal* being classified as *means*) has a lesser weight globally, while the mis-classification of units of type *assertion* as either *proposal* or *result* gains relevance.

Relation	*additional*	*detail*	*none*	*sequence*	*support*
additional	-	1	1	0	3
detail	6	-	1	0	**28**
none	0	3	-	0	3
sequence	0	**14**	0	-	0
support	6	**31**	1	1	-

Table 8: Percentage of errors produced for pairs of AFu classes over total.

In the case of the AFu task (Table 8), the percentage of errors when implementing the sequential transfer approach are very similar to those observed when using gold RST features. Here, again, the most frequent source of error is the mis-classification of units of types *support* and *detail* and the systematic mis-classification of units of type *sequence*.

The distribution of errors in the APa task for the sequence transfer experiment (Table 9) is also similar to the one observed when using RST features . The only difference that

Rel. distance	-4	-3	-2	-1	0	1	2	3	4
-4	-	0	1	5	0	3	0	0	0
-3	1	-	2	**10**	0	2	1	0	0
-2	2	2	-	**11**	1	2	0	1	0
-1	2	4	**10**	-	1	**11**	1	0	0
0	1	0	0	1	-	1	0	0	0
1	0	0	1	7	0	-	0	0	0
2	0	0	1	1	0	0	-	0	1
3	0	0	2	2	1	0	0	-	0
4	1	0	1	0	0	0	0	0	-

Table 9: Percentage of errors produced for pairs of APa distances up to 4 over total.

can be observed is that, percentually, the errors are slightly more distributed.

Tables 10 and 11 show examples of errors in the prediction of argumentative types and functions for the most frequent errors, while Table 12 show examples in which both argumentative types and functions are correctly predicted. The examples in Table 10 illustrate that the distinction between different types of units can be difficult even for humans. This is the case, for instance, for units of types *means* and *proposal*, as is is frequently ambiguous to interpret whether a unit that describes a method used for implementing a proposal should also be considered as part of the proposed solution. Similarly, the decision to annotate a unit as *assertion* or *result* depends on a subjective perception in relation to whether the authors' intention is to communicate a conclusion based on observations or a fact that should be accepted by the reader. Similar ambiguities arise when deciding between argumentative functions, as illustrated by the examples in Table 11.

Argumentative unit	**ATy**		**AFu**	
	Gold	**Pred.**	**Gold**	**Pred.**
ReNoun creates a seed set of training data by using specialized patterns and requiring that the facts mention an attribute in the ontology.	*means*	*proposal*	*detail*	*detail*
In addition, our approach easily scales to large data sets and is applicable to other data selection problems in natural language processing.	*assertion*	*result*	*detail*	*detail*
STIR uses information-theoretic measures from n-gram models as its principal decision features in a pipeline of classifiers detecting the different stages of repairs.	*proposal*	*means*	*detail*	*detail*

Table 10: Examples of errors in the prediction of argumentative types.

Argumentative unit	ATy		AFu	
	Gold	**Pred.**	**Gold**	**Pred.**
Our joint model with lexical normalization handles the orthographic diversity of microblog texts.	*assertion*	*proposal*	*support*	*detail*
For example, punctuation and entity tags in Wikipedia data define some word boundaries in a sentence.	*description*	*assertion*	*detail*	*support*
ReNoun then generalizes from this seed set to produce a much larger set of extractions that are then scored.	*means*	*means*	*sequence*	*detail*

Table 11: Examples of errors in the prediction of argumentative functions.

Argumentative unit	ATy		AFu	
	Gold	**Pred.**	**Gold**	**Pred.**
We propose the first probabilistic approach to modeling cross-lingual semantic similarity (CLSS) in context which requires only comparable data.	*proposal*	*proposal*	*none*	*none*
Search engines are increasingly relying on large knowledge bases of facts to provide direct answers to users' queries.	*assertion*	*assertion*	*additional*	*additional*
We show that finer resolution grounding helps coarser resolution grounding, and vice versa.	*result*	*result*	*support*	*support*

Table 12: Examples of correct predictions.

8 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 annotations, 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 independently of the learning algorithm. It is also relevant to confirm that similar or better results can

be obtained by argument mining models trained with word representations obtained by means of pre-trained encoders when no discourse annotations are available.

These results open several paths up for additional research, including the implementation of other transfer learning approaches –e.g.: the adaptation of discourse parsing pre-trained models to argument mining tasks– as well as other neural architectures – including attention-based ones, which have proven to achieve good results in argument mining tasks [50]. We will further explore the possibilities of multi-task learning strategies as more data with argumentative annotations becomes available. In particular, we plan to continue extending the coverage of the argumentative annotation layer of the SciDTB corpus. We expect this to become a valuable resource not only for our future experiments but also for the argument mining community in general.

Acknowledgments

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

References

[1] Accuosto, P., Saggion, H.: Discourse-driven argument mining in scientific abstracts. In: 24th International Conference on Applications of Natural Language to Information Systems. pp. 1—-13. Springer (2019)

[2] Accuosto, P., Saggion, H.: Transferring knowledge from discourse to arguments: A case study with scientific abstracts. In: Proceedings of the 6th Workshop on Argument Mining (ArgMining 2019). pp. 41–51. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-4505

[3] 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

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

[5] 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)

[6] 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)

[7] Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)

[8] Clark, J.: Text mining and scholarly publishing. Publishing Research Consortium **2013** (2013)

[9] 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 (ACL 2005). pp. 363–370. Association for Computational Linguistics (2005)

[10] Fisas, B., Ronzano, F., Saggion, H.: A multi-layered annotated corpus of scientific papers. In: Proceedings of the 2016 The International Conference on Language Resources and Evaluation (2016)

[11] Florou, E., Konstantopoulos, S., Koukourikos, A., Karampiperis, P.: Argument extraction for supporting public policy formulation. In: Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 49–54. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013)

[12] Goudas, T., Louizos, C., Petasis, G., Karkaletsis, V.: Argument extraction from news, blogs, and social media. In: Hellenic Conference on Artificial Intelligence. pp. 287–299. Springer (2014)

[13] Green, N.: Identifying argumentation schemes in genetics research articles. In: Proceedings of the 2nd Workshop on Argumentation Mining. pp. 12–21 (2015)

[14] Habernal, I., Eckle-Kohler, J., Gurevych, I.: 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 (2014)

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

[16] Hyland, K.: Hedging in scientific research articles, vol. 54. John Benjamins Publishing (1998)

[17] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015) (2015)

[18] 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)

[19] 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 (NAACL 2016). pp. 1490–1500 (2016)

[20] Lauscher, A., Glavaš, G., Eckert, K.: ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In: Proceedings of the 5th Workshop on Argument Mining (ArgMining 2018). pp. 22–28 (2018)

[21] Lauscher, A., Glavaš, G., Ponzetto, S.P.: An argument-annotated corpus of scientific publications. In: Proceedings of the 5th Workshop on Argument Mining (ArgMining 2018). pp. 40–46 (2018)

[22] Lawrence, J., Reed, C.: Argument mining: A survey. Computational Linguistics pp. 1–54 (2019)

[23] 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 (ACL 2014) (Volume 1: Long Papers). vol. 1, pp. 25–35 (2014)

[24] 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)

[25] Liakata, M., Soldatova, L.N., et al.: Semantic annotation of papers: Interface & enrichment tool (SAPIENT). In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. pp. 193–200. Association for Computational Linguistics (2009)

[26] Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.: Corpora for the conceptualisation and zoning of scientific papers. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (May 2010)

[27] 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)

[28] 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)

[29] Martínez Alonso, H., Plank, B.: 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. pp. 44–53. Association for Computational Linguistics, Valencia, Spain (Apr 2017)

[30] Mochales-Palau, R., Moens, M.F.: Argumentation mining: The detection, classification and structure of arguments in text. In: Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009). pp. 98–107. ACM (2009)

[31] 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)

[32] 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: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). pp. 1319–1324. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)

[33] Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010)

[34] Peldszus, A., Stede, M.: An annotated corpus of argumentative microtexts. In: Proceedings of the First Conference on Argumentation, Lisbon, Portugal (June 2015)

[35] Peldszus, A., Stede, M.: 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). pp. 938–948 (2015)

[36] Peldszus, A., Stede, M.: Rhetorical structure and argumentation structure in monologue text. In: Proceedings of the Third Workshop on Argument Mining (ArgMining 2016). pp. 103–112 (2016)

[37] 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)

[38] Peters, M., Neumann, M., Zettlemoyer, L., Yih, W.t.: Dissecting contextual word embeddings: Architecture and representation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1499–1509. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)

[39] Publons: 2018 Global State of Peer Review (2018)

[40] Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The ACL Anthology network corpus. Language Resources and Evaluation **47**(4), 919–944 (2013)

[41] Reed, C., Rowe, G.: Araucaria: Software for argument analysis, diagramming and representation. International Journal on Artificial Intelligence Tools **13**(04), 961–979 (2004)

[42] 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)

[43] Ruder, S.: Neural transfer learning for natural language processing. Ph.D. thesis, NUI Galway (2019)

[44] Schulz, C., Eger, S., Daxenberger, J., Kahse, T., Gurevych, I.: 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). pp. 35–41. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2006

[45] Slonim, N.: Project Debater. In: Computational Models of Argument - Proceedings of COMMA 2018, Warsaw, Poland, 12-14 September 2018. p. 4 (2018). https://doi.org/10.3233/978-1-61499-906-5-4

[46] 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)

[47] Stab, C., Gurevych, I.: Annotating argument components and relations in persuasive essays. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1501–1510. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (Aug 2014)

[48] Stab, C., Gurevych, I.: Parsing argumentation structures in persuasive essays. Computational Linguistics **43**(3), 619–659 (2017)

[49] Stab, C., Kirschner, C., Eckle-Kohler, J., Gurevych, I.: 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. pp. 21–25 (2014)

[50] Stab, C., Miller, T., Schiller, B., Rai, P., Gurevych, I.: Cross-topic argument mining from heterogeneous sources. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018). pp. 3664–3674 (2018)

[51] 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)

[52] Stein, B., Wachsmuth, H. (eds.): Proceedings of the 6th Workshop on Argument Mining. Association for Computational Linguistics, Florence, Italy (Aug 2019)

[53] Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research **3**(Dec), 583–617 (2002)

[54] Subramanian, S., Trischler, A., Bengio, Y., Pal, C.J.: Learning general purpose distributed sentence representations via large scale multi-task learning. In: International Conference on Learning Representations (2018), `https://openreview.net/forum?id=B18WgG-CZ`

[55] Tenopir, C., Christian, L., Kaufman, J.: Seeking, reading, and use of scholarly articles: An international study of perceptions and behavior of researchers. Publications **7**(1), 18 (2019)

[56] 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 (EMNLP 2009) (Volume 3). pp. 1493–1502. Association for Computational Linguistics (2009)

[57] Teufel, S., et al.: Argumentative zoning: Information extraction from scientific text. Ph.D. thesis, University of Edinburgh (1999)

[58] Toulmin, S.E.: The Uses of Argument. University Press (1958)

[59] Visser, J., Konat, B., Duthie, R., Koszowy, M., Budzynska, K., Reed, C.: Argumentation in the 2016 us presidential elections: Annotated corpora of television debates and social media reaction. Language Resources and Evaluation pp. 1–32 (2019)

[60] Wachsmuth, H.: Argumentation mining (2019)

[61] Walton, D., Reed, C., Macagno, F.: Argumentation schemes. Cambridge University Press (2008)

[62] Walton, D.N., Walton, D.N.: Informal logic: A handbook for critical argument. Cambridge University Press (1989)

[63] 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). pp. 336–348. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1039

[64] Yang, A., Li, S.: 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). pp. 444–449. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

[65] Zhang, W., Fang, Y., Ma, Z.: The effect of task similarity on deep transfer learning. In: International Conference on Neural Information Processing. pp. 256–265. Springer (2017)