Argumentation mining in scientific literature: From computational linguistics to biomedicine

Pablo Accuosto[a], Mariana Neves[b] and Horacio Saggion[a]

[a]*LaSTUS/TALN Group, Universitat Pompeu Fabra, Spain*
{name.surname}@upf.edu
[b]*German Federal Institute for Risk Assessment (BfR), Germany*
mariana.lara-neves@bfr.bund.de

Abstract

In this work we propose to tackle the limitations posed by the lack of annotated data for argument mining in scientific texts by annotating argumentative units and relations in research abstracts in two scientific domains. We evaluate our annotations by computing inter-annotator agreements, which range from moderate to substantial according to the difficulty level of the tasks and domains. We use our newly annotated corpus to fine-tune BERT-based models for argument mining in single and multi-task settings, finally exploring the adaptation of models trained in one scientific discipline (computational linguistics) to predict the argumentative structure of abstracts in a different one (biomedicine).

Keywords

argument mining, scientific corpora, domain adaptation, transformer models

1. Introduction

The accelerated pace at which scientific knowledge is produced makes its discovery and assessment a challenging task. Natural language processing (NLP) technologies, in general, and text-mining tools, in particular, have become increasingly essential to identify and characterize the most relevant information produced in a given scientific discipline.

In order to assess a research article it is necessary to consider its logic, rhetoric and dialectic quality dimensions [1]. It is therefore not enough to identify the claims made by its authors but also the evidence that they provide to support them. NLP tools that help to identify the main argumentative elements of a given text and how they are connected to each other can support the assessment of a given article. The automatic identification of arguments, its components and relations in texts is known as *argument mining* or *argumentation mining* [2]. The tasks involved in the automatic extraction of arguments from texts (claim/premise identification, prediction of argumentative structure) are not substantially different to other text mining tasks for which neural-based supervised learning methods produce state-of-the-art results (e.g.: text segmentation, sequence labelling and entity linking) [3]. These approaches, however, rely on large volumes of annotated data which are difficult to obtain for complex tasks such as argument mining. Scarcity of annotated corpora, therefore, limits the possibilities of using supervised machine learning algorithms for the identification of argumentative units and relations in texts

BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR 2021, April 1, 2021.

[4]. This obstacle is greater when dealing with scientific discourse: the inherent complexity of scientific texts makes it very difficult to carry out large annotation efforts with lay annotators.

1.1. Contributions

In previous works [5, 6] we proposed an annotation scheme for argumentative units and relations which considered the specificities of the scientific discourse and conducted a pilot annotation experiment with 60 abstracts from papers in the computational linguistics domain. These pilot annotations were done by one person as they were intended: i) to analyze the possibility of leveraging information contained in discourse-level annotations in order to improve the performance of argument mining models trained with a small number of abstracts and, ii) to explore the potential value of the trained models to predict the acceptance/rejection of the manuscripts in computational linguistics conferences, which was considered as a proxy for argumentative quality aspects of the abstracts. In those pilot experiments we trained BiLSTM models with CRF classifiers on top and used contextualized word embeddings obtained by means of pre-trained ELMo encoders [7]. In this work we:

1. Refine our previous annotation scheme to better account for the argumentative structure of scientific abstracts and to simplify the annotation process.
2. Make available SciARG, a corpus obtained by applying our new scheme to the annotation of 510 scientific abstracts in two domains: computational linguistics (CL) and biomedicine. Three annotators participated in the annotation process in the CL domain, while two annotators were involved in the annotation of biomedical abstracts.
3. Assess the consistency of SciARG annotations by analyzing inter-annotator agreement.
4. Use the SciARG corpus to fine-tune and evaluate BERT-based argument mining models, both in single and multi-task settings.
5. Analyze the potential of adapting models trained with CL abstracts -the original discipline for which the annotation schema was developed- to the biomedical domain.

The SciARG corpus and the code used in the experiments described in this work are made publicly available as a contribution to the research community.[1]

The rest of the paper is organized as follows: in Section 2 we describe previous work aimed at identifying arguments in scientific texts. In Section 3 we describe the data used to generate the corpus, our proposed annotation scheme and the annotation process. In Section 4 we describe the experiments conducted with the generated corpus and, in Section 5, we analyze the results obtained. In Section 6, we present our conclusions and suggest potential follow-ups.

2. Related Work

The inherent complexity and ambiguity of the scientific language makes the identification of arguments in scientific texts a particularly challenging task [8, 9, 10]. The Argumentative Zoning (AZ) model [11, 12] and the CoreSC scheme [13, 14] provide relevant antecedents in this area. AZ includes categories used to annotate knowledge claims made by the authors of the

[1] SciARG is available at https://github.com/LaSTUS-TALN-UPF/SciARG

papers and to establish connections with previous works. CoreSC, in turn, provides a readable representation of the research process described by the paper. Differences and similarities between the two schemes are studied in [15]. AZ was originally applied to the annotation of computational linguistics texts and CoreSC in physical chemistry and bio-chemistry articles. Dernoncourt and Lee [16] released in 2017 the PubMed 200k RCT dataset as a resource to train sentence classifiers for unstructured abstracts. The dataset was constructed by retrieving 195,654 structured abstracts of randomized controlled trials from the 2016 MEDLINE/PubMed Baseline Database[2] and labelling each sentence with the name of the section it belongs to. It is relevant to note that the aforementioned corpora and datasets are aimed at the classification of the rhetorical role of sentences but not the discourse relations between them. In this work we intend to establish a bridge between these two annotation levels. Lawrence and Reed [2] and Lippi and Torroni [3] provide thorough analyses of argument mining initiatives in various types of texts and domains, including legal documents [17], online discussions [18], Wikipedia articles [19], newspapers [20], student essays [21] and television debates [22], while Habernal and Gurevych [23] and Schulz et al. [24] explore argument mining in collections of texts from multiple, diverse sources. Few annotation efforts have focused on the analysis of arguments in scientific articles when compared to the number of works aimed at identifying argumentative components and relations in other textual genres. The annotation of 24 German scientific articles in the educational domain by Kirschner et al. [9] is one of the first works intended for the analysis of the whole argumentative structure of scientific texts, considering not only argumentative components but also how they are linked to each other. Lauscher et al. [25, 26] carried out experiments in which they enriched, with an argumentation layer, 40 papers in the area of computer graphics included in the DrInventor Scientific Corpus [27]. As mentioned in Section 1, we have previously conducted experiments with 60 computational linguistic abstracts aimed at analyzing the potential benefits obtained by enriching argument mining models with discourse-level knowledge [6].

3. SciARG Corpus

In this section we describe the source data used as a basis of the SciARG corpus as well as the annotation schema that we propose. We describe the annotation process and assess the quality of the produced annotations by considering inter-annotator agreement measures.

3.1. Data

The SciARG corpus covers two knowledge areas: computational linguistics and biomedicine. We refer to these sub-corpora as CL and BIO, respectively.

- **CL corpus**. Includes 225 computational linguistics abstracts from the ACL Anthology [28].[3] These abstracts are a subset of the 798 abstracts annotated with discourse relations in the

[2]The MEDLINE database of life sciences and biomedical information (www.nlm.nih.gov/bsd/medline.html) is maintained by the U.S. National Library of Medicine and available through the PubMed (pubmed.ncbi.nlm.nih.gov) search engine.

[3]In particular, from the Proceedings of the 2014 Conference on Empirical Methods in NLP (EMNLP).

Discourse Dependency TreeBank for Scientific Abstracts (SciDTB) [29].[4]

- **BIO corpus**. Includes 285 biomedical abstracts of articles from MEDLINE/PubMed. These abstracts are a sample of those used by Neves et al. [32] for the evaluation of argumentation in the biomedical domain. The sample was selected in a stratified way in order to include all annotations types considered in the referred work.

3.2. Annotation scheme

In this work we focus on the analysis of the way in which authors logically structure information in abstracts to persuade potential readers about the relevance and validity of their proposals. Our annotation scheme is aimed at capturing the underlying argumentative structure departing from its linguistic realization. It is therefore relevant to consider previous works that characterize the different constituent elements of scientific abstracts. Several works have been dedicated to the study, from a genre analysis perspective, of the rhetorical structure of scientific articles and its parts [33, 34]. Based on these works, a broad categorization of the most frequent rhetorical moves in scientific abstracts can be considered: i) **contextualization** of the research topic; ii) **limitations** in existing solutions; iii) **main purpose** of the current work; iv) description of the **methodology**; v) summary of the **main results**; vi) **conclusions**.[5] Based from this general structure of scientific abstracts we propose a fine-grained scheme that considers a sentence as the annotation unit and contains 11 ***types of units*** (Table 1) and six ***types of directed relations*** (Table 2).[6] Each of the unit types can, in turn, be mapped to a coarse-grained category. The use of fine or coarse-grained types depend on specific usages of the corpus.[7]

Table 1
Fine and coarse-grained types of units

Type of unit	Description	Coarse
proposal	high level description of the proposed approach/solution	*proposal*
proposal-implementation	processes/tools/methods that are part of the proposal	*proposal*
observation	data obtained from experiments	*outcomes*
result	direct interpretation of observed data	*outcomes*
result-means	results and the means by which they were obtained	*outcomes*
conclusion	high-level interpretation/generalization of results	*outcomes*
means	secondary methods/processes not part of the proposal	*methods*
motivation-problem	known problem/limitation addressed by the proposal	*motivation*
motivation-hypothesis	new ideas/paths for known problems/limitations	*motivation*
motivation-background	known information to support the proposed approach	*motivation*
information-additional	additional information (definitions/examples)	*other*

An annotated abstract can be seen as a directed graph with the sentences as its nodes and the relations between them as the edges. In order to gain in uniformity of the annotations,

[4] This allows us to continue exploring the interaction between argumentative, rhetoric and discourse annotation levels in scientific abstracts, as originally proposed by Peldszus and Stede [30, 31] for other textual genres.

[5] Minor variations to this general structure depend on the knowledge area.

[6] We omit the *attack* relation as there were no attacks identified in any of the abstracts analyzed.

[7] In the context of this work we use the fine-grained types.

Table 2
Relations

Relation	Description of the child node function
support	provides new supporting information/evidence for the parent
elaboration	provides additional information relevant to assess/contextualize the parent
by-means	describe methods through which supporting evidence is obtained
info-required	provides information essential to understand/contextualize the parent
sequence	describes a step that comes after the step described by the parent in a process
info-optional	provides non-essential information

reduce the level of ambiguity and simplify the annotation process, we only consider trees as valid annotations graphs. In addition to types and relations, annotators were asked to identify the unit that describe the most significant contribution of the work as the ***main unit***. Fig. 1 shows an example of the tree resulting from annotating the abstract from [35].

In previous work we considered argumentative units at sub-sentence level and explored the relation between discourse and argumentative levels [6]. Having observed that discourse-level annotations can be leveraged to identify argumentative relations within sentences we decided, in this work, to focus at the sentence level and leave the prediction of intra-sentence relations as a second step in an argument mining pipeline. This allows us to facilitate the annotation process and it also contributes to bridge the gap between annotations aimed at identifying the rhetorical role of sentences (such as AZ and CoreSC) and those aimed at finding discourse relations between -and within- them (such as SciDTB).

Most sentences in computational linguistics abstracts contain one type of argumentative unit. A relatively frequent exception to this are sentences in which mentions to methods are included in the results. For this specific case we introduce in our schema the unit type *results-means*.[8] For other cases in which more than one type of unit can be identified, our scheme allows annotators to register this information by assigning a second type to the sentence. In the annotation process annotators were asked to weight the relevance of the different types of information contained in the sentence to make a decision with respect to the main and secondary types.

It is frequent to find, in abstracts, that authors build up supporting evidence or justifications for implicit or explicit claims in more than one sentence. Consider the example in Fig. 1. Nodes (4) (*motivation-problem*) and (7) (*motivation-background*) provide partial information that, when considered together, contribute to justify the proposed work described in node (2). From a discourse analysis perspective, this would be represented by a multi-nuclear relation which could be annotated by introducing a different type of node in the argumentative tree. This, on one hand, introduces some practical difficulties in the automatic processing of the annotations, as will become evident when we describe the experiments in Section 4 and, on the other hand, does not allow to capture the hierarchical relation between nodes (4) and (7). We opt, instead, to introduce the relation *info-required* to account for these cases. In this example, we indicate that there is an *info-required* relation that goes from node (7) to node (4) and a *support* relation that goes from node (4) to node (2). When looking for supporting evidence for the sentence in node (2), therefore, we would consider not only their direct children but

[8]Examples for all types of units are included in the supplemental material.

also the chains of sentences below them linked by *info-required* relations. Depending on the specific dimensions of the argumentation quality to analyze, therefore, different subsets of units and relations can be considered. Most argument mining works focus on logic aspects of argumentation and, in particular, in the arguments' cogency [1]. This argumentative dimension is conveyed by relations of type *support* (or *attack*). It has been noted, nevertheless, that the different argumentative dimensions correlate with the perceived overall argumentative quality of the text [1, 36]. Should we consider only *support* relations, our annotations would not capture the link between a proposal and its implementation details, which improves the text's clarity and persuades the reader about the validity of the proposal and, therefore, the perceived overall argumentative strength of the text.

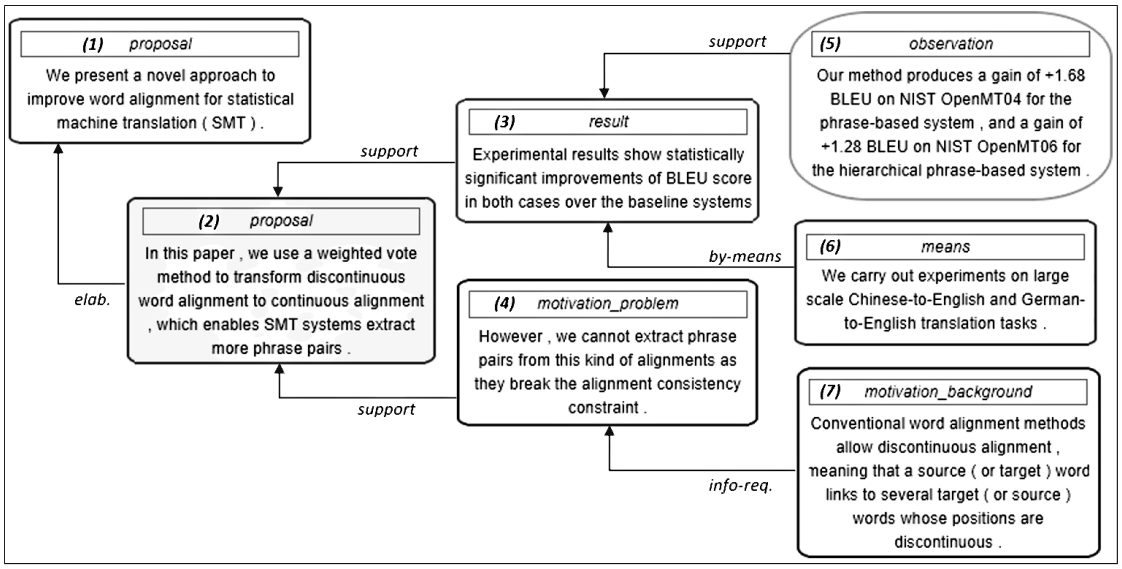

Figure 1: Example of argumentative tree. The main unit corresponds to the node with grey background.

3.3. Annotation process

The annotation guidelines used in this work are made available online.[9] The first step of the annotation process is to have the texts of the abstracts splitted into sentences. In the case of the CL corpus, the source files used are already segmented into sentences and elementary discourse units in the SciDTB corpus. For the biomedical abstracts the sentence segmentation is done by means of the syntok tool.[10] The annotation was done by means of a modified version of GraPAT (Graph-based Potsdam Annotation Tool) [37] according to the specific needs of the task.

We first developed and adjusted the annotation scheme with the CL corpus and then used it to annotate the BIO corpus. One of the goals of this work is to assess the applicability of the proposed scheme to different domains. In Section 3.5 we analyze the main differences between both corpora and the resulting annotations. The annotation of the CL corpus was done by

[9] https://github.com/LaSTUS-TALN-UPF/SciARG/blob/main/Annotation_Guidelines_Arguments_SciDTB.pdf
[10] github.com/fnl/syntok

three expert annotators, a_1, a_2, a_3 (two NLP researchers and one computational linguist) in three rounds. The first two rounds were aimed at training the annotators, clarifying doubts and making the necessary adjustments to the annotation scheme and tool. As a result of the whole process 225 CL abstracts were annotated, having 30 abstracts annotated by the three annotators to compute inter-annotator agreement. For the BIO corpus only two of the CL annotators (a_1, a_2) could participate in the annotation process. In this case no training phase was needed and there were no substantial modifications to the annotation tool or scheme. As a result of this process 285 abstracts were annotated, of which 50 were annotated by both annotators. For our experiments we split both sub-corpora into training and test sets, as described in Section 5.1.

3.4. Agreement

In this section we assess the reliability of our annotations by considering inter-annotators' agreement. Table 3 shows the agreements obtained for CL and BIO sub-corpora. In the case of CL, where we have three annotators, we report the average of the pairwise agreements with their corresponding standard deviations. In order to compute the agreements we consider exact matches between pairs of labels assigned by two annotators. For the parent attachment task the label is the absolute position of the parent sentence in the document. In addition to each sub-task-specific agreement, we report the agreement observed when considering simultaneous exact matches for all the tasks. It is relevant to note that annotations within one document cannot be consider as completely independent from each other, which presents a limitation when interpreting the significance of Cohen's κ coefficient.[11] In Table 3, therefore, in addition to Cohen's κs, we directly report the accuracy obtained for the different tasks without any presupposition with respect to the independence of the annotations. When considering a pair of annotated documents as labeled trees, the accuracy indicates the number of changes (with respect to the number of nodes) that would be necessary to do in one tree to obtain the other one. It can, therefore, be interpreted as an edit-distance measure that allows to estimate the degree of agreement between two annotators with respect to the argumentative roles of the nodes when considering the document as a whole. In all cases annotation agreements fall between moderate and substantial levels: substantial agreements are obtained in general in the CL corpus (and almost perfect agreement when coarse-grained types are considered), while agreements in the identification of unit types and relations are lower in the BIO corpus. In addition to the fact that the annotation scheme was designed and adjusted specifically for the CL domain, lower agreements are expected in BIO as abstracts have a higher level of complexity than CL ones in terms of their structure, the number of units that they contain and their lengths, as shown in Section 3.5. It is also relevant to note that annotators have a high level of familiarity with CL texts while they are not experts in the BIO domain. When analyzing discrepancies in the annotation of the BIO corpus we observe that units of types *observation* and *result* give origin to systematic disagreements between annotators a_1 and a_2. In fact, annotator a_2 annotated as *observation* 64% of the units that annotator a_1 annotated as *result*, which makes us believe that a clear distinction between these two types is difficult to establish without specific domain

[11] As a decision made at one node of the argumentative structure affects decisions made in other nodes. This problem has already been observed by Marcu et al. [38] when evaluating inter-annotator agreement of discourse annotations.

knowledge. When coarse-grained types are considered, in fact, these two types of units are not distinguished and the level of agreement reaches 0.93 Cohen's κ. This also affects the attachment of these units to their parents as they are also considered differently in terms of the argumentative role that they play.

Table 3
Agreement in CL and BIO sub-corpora

	Cohen's κ		Accuracy	
Task	**CL (avg. pairwise)**	**BIO**	**CL (avg. pairwise)**	**BIO**
Fine-grained unit type	0.77 ± 0.004	0.66	0.81 ± 0.004	0.72
Coarse-grained unit type	0.94 ± 0.016	0.93	0.96 ± 0.010	0.96
Parent position	0.72 ± 0.075	0.49	0.77 ± 0.062	0.54
Relation type	0.79 ± 0.026	0.43	0.84 ± 0.019	0.58
Main unit	0.92 ± 0.042	0.94	0.97 ± 0.013	0.99
All combined	0.59 ± 0.055	0.39	0.61 ± 0.053	0.40

3.5. Corpus statistics and analysis

Substantial differences can be observed between the CL and BIO sub-corpora. Abstracts in BIO are, in general, longer and argumentatively more complex than those in CL (Table 4). It is frequent in BIO to find abstracts that describe a series of experiments, each one with their results. In some cases, results from one experiment are used to motivate and/or justify new ones. This level of detail is not present in CL abstracts. In general, the description of research outcomes and their interpretation is much more complex in BIO abstracts, which leads to a significant difference in the number of units of type *observation*, *result* and *conclusion* when compared to CL abstracts.[12]

Table 4
Statistics of CL and BIO sub-corpora

Statistics	**CL**	**BIO**	**Statistics**	**CL**	**BIO**
Number of abstracts	225	285	Avg. #tokens/unit	24.4 ± 9.9	30.1 ± 14.2
Total number of units	1199	2787	Max. #tokens/unit	101	155
Avg. #units/abstract	5.3 ± 1.7	9.8 ± 3.1	Min. #tokens/unit	5	5
Max. #units/abstract	13	25	Forward relations	32%	34%
Min. #units/abstract	2	2	Backward relations	68%	66%

The distinction between the plain report of observed data, the interpretation of results and the extraction of conclusions from them is more ambiguous in BIO than in CL and, therefore, differentiating these types of units is more difficult. The distances between units and their parents are greater in BIO. In fact, nearly 19% of the times a unit is 5 or more units away from its parent. In CL this occurs only in 2% of the cases. In 69% of the cases CL units are only one or two units away from its parent when considering the CL corpus. In BIO this occurs only 58% of

[12]While in CL 3% of the units are of type *observation*, 19% of type *result* or *result-means* and 4% of type *conclusion*, in BIO there are, respectively, 18%, 26% and 11% units of these types. More details are provided in the supplemental material.

the times. In both domains *backward* relations are more frequent than *forward* relations: the parent occurs before the child 68% and 66% of the times in CL and BIO, respectively.

4. Experiments

In this section we describe the experiments carried out in order to, given a scientific abstract, predict the nodes and relations needed to represent its argumentative structure.

4.1. Tasks

- **Unit type**: Given the text of a sentence, predict its type. The class to predict in this case is one of the 11 fine-grained types described in Table 1.
- **Relation direction**: Given two sentences, predict whether a forward or backward relation exists between them (e.g: whether the first unit is a child of the second one in the argumentative tree or vice versa). We model this task as a three-class classification task where, given two sentences, the possible classes to predict are *forw*, *back* or *none*, indicating, respectively, that there is a directed relation from the first to the second sentence, from the second to the first sentence, or that the two sentences are not related.
- **Relation type**: Given a sentence, predict the label of the relation with its parent (its argumentative and/or discourse function) or *none* for the root node. The class to predict in this case is one of the 6 relations described in Table 2.
- **Main unit**: Given the text of a sentence, predict whether it is the *main unit*.[13]

There are clear links between the four tasks. For instance, the main unit is, in most cases, the root of the argumentative tree. Associations can also be established between a unit's type and its function: in the most frequent case, units of type *result* are used to *support* units of type *proposal* or *conclusion*. It is therefore natural to explore the possibility of training the tasks jointly, in a multi-task setting, which we compare to the results obtained when training the results independently, in single-task settings.

4.2. Experimental setup

Transformer-based encoders [39] such as BERT [40] currently provide state-of-the-art performance for semantic text classification tasks. For our experiments we make use of the BERT implementations available as part of HuggingFace's Transformers library [41]. We use the cased version of SciBERT [42] as base model, as it is trained on texts in the same domains as the ones covered by our corpus. We apply the standard method of considering the representation of the `[CLS]` token and feed it into linear classifiers. A softmax function is then applied to the classifier's output in order to obtain the distribution of probabilities for the predicted labels. In the multi-task setting the BERT layers are shared among all the tasks only training independently the task-specific heads. We follow the common practice of modeling the identification of relations between pairs of sentences as a classification task using as input the sequence obtained

[13] The main unit is considered to be the unit where the main proposed approach/solution is described.

by concatenating the tokens occurring in each sequence separated by the [SEP] special token. In order to predict the relations present in a given abstract we consider all the pairs formed by a sentence and the sentences that occur after it in the text. As a result of the prediction we should obtain the label *forw* if the first sentence is a child of the second one in the argumentative tree, *back* if the relation is established in the opposite direction, and *none* if the two sentences are not in a direct relation. The most frequent case, given any two sentences, is that they are not related. In order to train the model with more positive examples, when a relation exists between two sentences we sample it twice in the training set: once for each direction, with the corresponding *forw* / *back* labels.[14] For evaluation we consider each pair only once, in the order in which they appear in the text.

When fine-tuning our models in each domain, we consider the median number of tokens in the input sequences and set the maximum sequence length to its double. We consider cross-entropy as the loss function to optimize. We use the Adam optimizer with a learning rate of 2e-5 and a warm-up period of 10% of the learning steps. We set a dropout probability of 0.1 for multi-task settings and 0.2 for single-task ones. The batch size used is of 16 instances with gradient accumulation of 2 batches. These hyperparameters were set based on five-fold cross-validation evaluations in the training set. While the general recommendation is to fine-tune BERT for 2 to 4 epochs [40], we observed that more epochs were required to train our tasks, considering the large number of classes and the relatively small number of training instances.[15] In Section 5 we report the results obtained for each task for 5, 10 and 15 epochs, so it is possible to observe how each combination of task and domain impact on the training time required both in single and multi-task settings, which would be more difficult to observe if we considered either the best cross-validation epoch or a fixed number of epochs, as frequently done when using BERT. We also observed that the models' overall performances improved when including, as additional tokens, information about the sentences positions in the abstracts as well as their relative distance and order. We add special tokens to the standard BERT tokenizer to represent this information.[16]

As mentioned in Sections 1.1 and 3, our annotation scheme was specifically developed to account for argumentative types and relations in computational linguistic abstracts. One of the goals of this work is to explore i) the applicability of this scheme to other scientific disciplines and ii) whether models trained in the CL domain can be easily adapted to predict the argumentative structure of abstracts in other scientific areas. In particular, in the BIO domain. We use the newly annotated set of biomedical abstracts in order to respond to both research questions.

We are also interested in exploring to what extent models trained with annotations in CL contain task-specific information that can be exploited to predict argumentative types and relations in scientific abstracts with a more complex structure and in another discipline. We therefore analyze the results obtained by keeping the weights of a model fine-tuned with the CL abstracts fixed and only training a linear classifier on top of it with the BIO abstracts.

[14] I.e.: if sentence s_2 is a child of sentence s_1 in the argumentative tree, we include the instances (s_2, s_1, *forw*) and (s_1, s_2, *back*) in the training set.

[15] While there are 3 classes and 13,874 training instances for the BIO/*relation type* task, we only have 1,049 training instances and 11 classes for CL/*unit type*.

[16] I.e.: "[CLS] [AFTER] [DISTANCE-1] [POS-1] This paper presents ... [SEP] We observe ..."

5. Results

We present in this section the performance of the models trained in each domain (BIO, CL), as well as the adaptation of CL models to BIO by training a small number of additional parameters.

5.1. Evaluation

In the CL domain 30 abstracts were annotated in common by three annotators (a_1, a_2, a_3). This set is used for evaluation, while the rest of the 195 abstracts is used to train the models. We generate a set of consensus annotations by assigning, to each instance, the majority label considering the annotations by a_1, a_2 and a_3. In the few cases in which there is total discrepancy among the three annotators we keep the label assigned by the annotator with the highest average of pairwise agreement with the other two annotators. For BIO it is not possible to do this since we only have two annotators (a_1, a_2) that annotated 50 abstracts, while the rest of the 235 abstracts were annotated only by one annotator (a_1). Therefore, in these experiments we only use the annotations produced by a_1. As test set, in this case, we consider the subset of 35 abstracts[17] annotated by a_1 with the highest levels of agreement with the annotations made by a_2, keeping the remaining 250 abstracts annotated by a_1 as training set.

In Table 5 we report the results obtained by the set of experiments described in Section 4 when evaluating against the *consensus* annotations. For the types of units and relations we use weighted-averaged F1-scores, as we want to consider the contribution of each label to the results in proportion to their frequency. For the evaluation of the direction of the relations, instead, we use macro-averaged scores, which are more sensitive to the minority classes. If we were to use micro-averaged or weighted-averaged scores in these cases we would obtain misleading high numbers for F1, given the large proportion of *none* labels which are correctly classified. This is also the case for the prediction of the main unit. As expected, considering the greater argumentative complexity of the BIO abstracts, which is also reflected in the lower levels of inter-annotator agreements, the performance of the models trained and evaluated with the BIO annotations is lower than the one obtained with the CL annotations (Table 5). In the BIO domain the models trained jointly in a multi-task settings tend to perform better than those in which these tasks are trained independently. In the case of CL the difference between both settings is less evident: while there is a clear advantage of the multi-task setting in the prediction of the types of units, better results are obtained for the prediction of the parent relations in a single-task setting. We also observe that the BERT models fine-tuned with the CL annotations (CL-BERT) without in-domain fine-tuning perform competitively when compared to the models in which BERT is fine-tuned with the BIO annotations.

It is relevant to note that the CL-BERT model with frozen weights performs significantly better in the prediction of the BIO annotations than the frozen SciBERT encoder that we consider as baseline. This confirms that the model fine-tuned with CL annotations is able to capture information about the argumentative structure of scientific abstracts independent of the specific discipline in which it was trained.

[17] The number of 35 is chosen in order to keep the training-test sets percentages similar in both domains.

Table 5
Evaluation of BERT-based models in CL and BIO consensus test sets (F1-scores)

	CL-BERT Multi-task - CL Test Ann.				CL-BERT Single-task - CL Test Ann.			
Ep.	**R. Dir.**	**R. Type**	**U. Type**	**Main**	**R. Dir.**	**R. Type**	**U. Type**	**Main**
5	0.8221	0.7629	0.7959	0.9076	**0.8478**	0.7894	0.7659	0.9263
10	**0.8380**	0.7945	0.7897	0.9146	0.8456	0.7874	**0.7991**	**0.9376**
15	0.8262	**0.8216**	**0.8259**	**0.9243**	0.8263	**0.8159**	0.7801	0.9281

	BIO-BERT Multi-task - BIO Test Ann.				BIO-BERT Single-task - BIO Test Ann.			
Ep.	**R. Dir.**	**R. Type**	**U. Type**	**Main**	**R. Dir.**	**R. Type**	**U. Type**	**Main**
5	**0.7046**	**0.7425**	0.6375	0.9120	0.6953	0.7170	**0.6797**	0.8543
10	0.6973	0.7416	0.6717	0.9164	0.6612	**0.7221**	0.6593	**0.8676**
15	0.6929	0.7394	**0.6951**	**0.9249**	**0.6993**	0.7208	0.6738	**0.8676**

	CL-BERT Frozen weights - BIO Test Ann.				SciBERT Frozen weights - BIO Test Ann.			
Ep.	**R. Dir.**	**R. Type**	**U. Type**	**Main**	**R. Dir.**	**R. Type**	**U. Type**	**Main**
5	0.6928	0.7239	0.6207	**0.8769**	0.5162	0.5722	0.4934	0.8005
10	0.7075	**0.7269**	**0.6604**	0.8738	0.5408	**0.6331**	0.5364	0.8398
15	**0.7080**	0.7220	0.6588	0.8738	**0.5575**	0.6305	**0.5411**	**0.8441**

6. Conclusions

In this work we propose a new sentence-level annotation scheme for the identification of argumentative units and relations in scientific abstracts, which we apply to the annotation of 510 documents in two highly specialized domains: computational linguistics and biomedicine. The resulting corpus, as well as the code used to train and evaluate models trained with it, is made publicly available. The results obtained in our experiments encourage us to think that, in spite of the fact that the annotation scheme was originally developed and refined for the CL domain, it can be successfully applied to other scientific disciplines. This work also opens up new research paths, including further exploration of domain adaptation techniques for argument mining models in challenging domains as is the case of scientific articles.

Acknowledgments

This work was (partly) supported by the Spanish Government under the María de Maeztu Units of Excellence Programme (MDM-2015-0502) and by the Research and Innovation Agency of Uruguay (ANII). We also acknowledge support from the project Context-aware Multilingual Text Simplification (ConMuTeS) PID2019-109066GB-I00/AEI/10.13039/501100011033 awarded by Ministerio de Ciencia, Innovación y Universidades (MCIU) and by Agencia Estatal de Investigación (AEI) of Spain.

References

[1] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, B. Stein, Computational argumentation quality assessment in natural language, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (ACL 2017) (Volume 1: Long Papers), 2017, pp. 176–187.

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

[3] M. Lippi, P. Torroni, Argumentation mining: State of the art and emerging trends, ACM Trans. Internet Technol. 16 (2016) 10:1–10:25.

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

[5] P. Accuosto, H. Saggion, Mining arguments in scientific abstracts with discourse-level embeddings, Data & Knowledge Engineering (2020) 101840.

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

[7] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: https://www.aclweb.org/anthology/N18-1202. doi:10.18653/v1/N18-1202.

[8] C. Stab, C. Kirschner, J. Eckle-Kohler, I. Gurevych, 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, 2014, pp. 21–25.

[9] C. Kirschner, J. Eckle-Kohler, I. Gurevych, Linking the thoughts: Analysis of argumentation structures in scientific publications, in: Proceedings of the 2nd Workshop on Argumentation Mining, 2015, pp. 1–11.

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

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

[12] S. Teufel, A. Siddharthan, C. Batchelor, 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), Association for Computational Linguistics, 2009, pp. 1493–1502.

[13] M. Liakata, S. Saha, S. Dobnik, C. Batchelor, D. Rebholz-Schuhmann, Automatic recognition of conceptualization zones in scientific articles and two life science applications, Bioinformatics 28 (2012) 991–1000.

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

[15] M. Liakata, S. Teufel, A. Siddharthan, C. Batchelor, 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, 2010.
[16] F. Dernoncourt, J. Y. Lee, Pubmed 200k rct: A dataset for sequential sentence classification in medical abstracts, arXiv preprint arXiv:1710.06071 (2017).
[17] R. Mochales-Palau, M.-F. Moens, 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), ACM, 2009, pp. 98–107.
[18] T. Goudas, C. Louizos, G. Petasis, V. Karkaletsis, Argument extraction from news, blogs, and social media, in: Hellenic Conference on Artificial Intelligence, Springer, 2014, pp. 287–299.
[19] E. Aharoni, L. Dankin, D. Gutfreund, T. Lavee, R. Levy, R. Rinott, N. Slonim, Context-dependent evidence detection, 2018. US Patent App. 14/720,847.
[20] E. Florou, S. Konstantopoulos, A. Koukourikos, P. Karampiperis, Argument extraction for supporting public policy formulation, in: Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 49–54.
[21] C. Stab, I. Gurevych, Annotating argument components and relations in persuasive essays, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 1501–1510.
[22] J. Visser, B. Konat, R. Duthie, M. Koszowy, K. Budzynska, C. Reed, Argumentation in the 2016 us presidential elections: Annotated corpora of television debates and social media reaction, Language Resources and Evaluation (2019) 1–32.
[23] I. Habernal, I. Gurevych, Argumentation mining in user-generated web discourse, Computational Linguistics 43 (2017) 125–179. doi:10.1162/COLI_a_00276.
[24] C. Schulz, S. Eger, J. Daxenberger, T. Kahse, I. Gurevych, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 35–41. doi:10.18653/v1/N18-2006.
[25] A. Lauscher, G. Glavaš, S. P. Ponzetto, An argument-annotated corpus of scientific publications, in: Proceedings of the 5th Workshop on Argument Mining (ArgMining 2018), 2018, pp. 40–46.
[26] A. Lauscher, G. Glavaš, K. Eckert, ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing, in: Proceedings of the 5th Workshop on Argument Mining (ArgMining 2018), 2018, pp. 22–28.
[27] B. Fisas, F. Ronzano, H. Saggion, A multi-layered annotated corpus of scientific papers., in: Proceedings of the 2016 The International Conference on Language Resources and Evaluation, 2016.
[28] D. R. Radev, P. Muthukrishnan, V. Qazvinian, A. Abu-Jbara, The ACL Anthology network corpus, Language Resources and Evaluation 47 (2013) 919–944.
[29] A. Yang, S. Li, SciDTB: Discourse dependency TreeBank for scientific abstracts, in: Proceed-

ings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018) (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 444–449.

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

[31] A. Peldszus, M. Stede, An annotated corpus of argumentative microtexts, in: Proceedings of the First Conference on Argumentation, Lisbon, Portugal, 2015.

[32] M. Neves, D. Butzke, B. Grune, Evaluation of scientific elements for text similarity in biomedical publications, in: Proceedings of the 6th Workshop on Argument Mining, Association for Computational Linguistics, Florence, Italy, 2019, pp. 124–135. URL: https://www.aclweb.org/anthology/W19-4515. doi:10.18653/v1/W19-4515.

[33] J. Swales, Genre analysis: English in academic and research settings, Cambridge University Press, 1990.

[34] M. B. Dos Santos, The textual organization of research paper abstracts in applied linguistics, Text-Interdisciplinary Journal for the Study of Discourse 16 (1996) 481–500.

[35] Z. He, H. Wu, H. Wang, T. Liu, Transformation from discontinuous to continuous word alignment improves translation quality, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 147–152.

[36] L. Ng, A. Lauscher, J. Tetreault, C. Napoles, Creating a domain-diverse corpus for theory-based argument quality assessment, arXiv preprint arXiv:2011.01589 (2020).

[37] J. Sonntag, M. Stede, GraPAT: A tool for graph annotations, in: Proceedings of the 2014 The International Conference on Language Resources and Evaluation, 2014, pp. 4147–4151.

[38] D. Marcu, E. Amorrortu, M. Romera, Experiments in constructing a corpus of discourse trees, in: Towards Standards and Tools for Discourse Tagging, 1999. URL: https://www.aclweb.org/anthology/W99-0307.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[41] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., HuggingFace's Transformers: State-of-the-art natural language processing, ArXiv (2019) arXiv–1910.

[42] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3606–3611.

A. Supplemental material

A.1. Types of units

Table 6
Examples for each type of unit

proposal
We present a novel approach to improve word alignment for statistical machine translation (SMT) .
proposal-implementation
We observe , identify , and detect naturally occurring signals of interestingness in click transitions on the Web between source and target documents , which we collect from commercial Web browser logs .
observation
Our method produces a gain of +1.68 BLEU on NIST OpenMT04 for the phrase-based system , and a gain of +1.28 BLEU on NIST OpenMT06 for the hierarchical phrase-based system .
result
Experimental results show statistically significant improvements of BLEU score in both cases over the baseline systems .
means
We conducted experiments on two standard benchmarks : Chinese PropBank and English PropBank .
result-means
Results on the Switchboard disfluency tagged corpus show utterance-final accuracy on a par with state-of-the-art incremental repair detection methods , but with better incremental accuracy , faster time-to-detection and less computational overhead .
conclusion
This transfer learning approach brings a clear performance gain over features based on the traditional bag-of-visual-word approach .
motivation-problem
However , fundamental problems on effectively incorporating the word embedding features within the framework of linear models remain .
motivation-hypothesis
Combining the two tasks can potentially improve the efficiency of the overall pipeline system and reduce error propagation .
motivation-background
Recent work has shown success in using continuous word embeddings learned from unlabeled data as features to improve supervised NLP systems , which is regarded as a simple semi-supervised learning mechanism .
information-additional
The structure of argumentation consists of several components (i.e. claims and premises) that are connected with argumentative relations .

A.2. Distribution of types and relations in CL and BIO sub-corpora

Table 7
Distribution of unit types in CL and BIO

Type	CL	BIO
proposal	285	289
proposal-implementation	264	274
observation	39	505
result	157	703
conclusion	54	301
means	27	58
result-means	69	31
motivation-problem	103	97
motivation-background	157	487
motivation-hypothesis	20	16
information-additional	24	26

Table 8
Distribution of relations in CL and BIO

Relation	CL	BIO
support	417	1581
elaboration	358	535
by-means	28	57
info-required	120	303
sequence	29	2
info-optional	22	24