Using Semi-automatic Annotation Platform to Create Corpus for Argumentative Zoning

Alaa El-Ebshihy[1,2,3](✉), Annisa Maulida Ningtyas[1,4], Florina Piroi[1,2], Andreas Rauber[1], Ade Romadhony[5], Said Al Faraby[5], and Mira Kania Sabariah[5]

[1] Technische Universität Wien, Vienna, Austria
{alaa.el-ebshihy,annisa.ningtyas,florina.piroi}@tuwien.ac.at,
rauber@ifs.tuwien.ac.at
[2] Research Studios Austria, Data Science Studio, Vienna, Austria
[3] Alexandria University, Alexandria, Egypt
[4] Universitas Gadjah Mada, Yogyakarta, Indonesia
[5] Telkom University, Bandung, Indonesia
{aderomadhony,saidalfaraby,mirakania}@telkomuniversity.ac.id

Abstract. Argumentative Zoning (AZ) is a tool to extract salient information from scientific texts for further Natural Language Processing (NLP) tasks, e.g. scientific articles summarisation. AZ defines the main rhetorical structure in scientific articles. The lack of large AZ annotated benchmark datasets along with the manual annotation complexity of scientific texts form a bottle neck in utilizing AZ for scientific NLP tasks. Aiming to solve this problem, in previous work, we presented an AZ-annotation platform that defines and uses four categories, or zones (*Claim*, *Method*, *Result*, *Conclusion*) that are used to label sentences in scientific articles. The platform helps to create benchmark datasets to be used with the AZ tool. In this work we look at the usability of the said platform to create/expand datasets for AZ. We present a annotation experiment, composed of two annotation rounds, selected scientific articles from the ACL anthology corpus are annotated using the platform. We compare the user annotations with a ground truth annotation and compute the inter annotation agreement. The annotations obtained in this way are used as training data for various BERT-based models to predict the zone of a given sentence from a scientific article. We compare the trained models with a model trained on a baseline AZ corpus.

Keywords: Argumentative Zoning · Annotation · Benchmark creation

1 Introduction

For any research topic, there exist various available scientific articles from conferences, journal publications etc. Usually, the abstracts of the articles do not provide enough insights about the salient information in the article. Due to this, it is usually difficult for a researcher, especially young researchers and students, to decide whether to proceed reading the full paper text or not and whether it is relevant to their own work.

Extracting salient information in scientific literature is a known challenge, and NLP techniques are becoming increasingly crucial in trying to address it. The information to be extracted is part of the main components of any research article, which are:

O. Alonso et al. (Eds.): TPDL 2023, LNCS 14241, pp. 132–145, 2023.
https://doi.org/10.1007/978-3-031-43849-3_12

the research questions, hypothesis, methodology, results and conclusions. One of the approaches that are used to identify these components is *Argumentative Zoning (AZ)* [18]. AZ refers to the examination of the argumentative status of sentences in scientific articles and their assignment to specific argumentative zones. Its main goal is to collect sentences that belong to predefined categories (i.e. zones), such as "claim" or "method". AZ is useful as a tool for downstream NLP tasks; e.g. scientific article summarisation [5,7,9,12,16] and research articles theme classification [6].

Automatic AZ identification has been approached in previous work as a supervised learning problem to train a model with annotated scientific articles [1,3,14,16–18]. The bottle neck in training these algorithms is that the training data is obtained by manual annotation of scientific articles, a work that is complex and often not feasible [20] due to the technical document structure, the length of the articles, the necessity of domain expertise. Teufel et al. [18] introduced an annotation schema of seven AZ labels , which was later updated [17]. Accusto et al. [1] proposed a fine grained annotation schema with eleven categories for AZ.

Although there are ongoing efforts to create annotated corpus as training data for AZ models, the main challenge is expanding and creating an AZ corpus on complete papers, not only abstracts, and for domains other than the Computational Linguistics (CL)[1]. With this goal in sight, we proposed in previous work, a platform for the systematic annotation and, consequently, the creation of new benchmarks to be used for training AZ identification algorithms [8]. The platform uses a simplified a schema of four labels that identify the claims, methods, conclusion and results. Sentences from a scientific article are selected and labeled by a previously trained algorithm and users are asked to verify and correct the labels. In this work, we examine the feasibility of our platform in creating an annotated AZ corpus and the use of the annotated data in automatic AZ identification. More concretely:

1. we present an annotation experiment to annotate selected scientific articles by conducting two annotation rounds (online and onsite) with bachelor and master students using the AZ annotation platform.
2. we build a new AZ corpus using collected annotations from the annotation rounds in addition to using it to expand an existing AZ corpus in a previous work [1].
3. we use the constructed AZ corpora to train Bert-based models for AZ identification and compare their performance against a baseline model.

2 Related Work

Argumentative zoning (AZ) is defined as "the analysis of the argumentative status of sentences in scientific articles". The theory of AZ was formalized by Simone Teufel in her PhD thesis in 1999 [18]. There, Teufel introduced a manual annotation scheme for scientific articles, focusing on argumentative zones and rules with predefined zones (i.e. labels) to annotate 48 computational linguistic (CL) papers by categorizing each sentence into one of 7 zones: *BKG, OTH, OWN, AIM, TXT, CTR* and *BAS*. In her work, Teufel provided an approach that combined traditional hand-engineered features, meta-discourse features, and classification techniques to automatically classify sentences in

[1] Previous work mainly focused on CL domain.

scientific articles into argumentative zones [18]. It included the release of 80 CL hand-annotated articles, where each sentence was labeled with one of the above mentioned zones. Though the corpus is a strong gold standard, it is relatively small in size.

Later, a more fine grained schema of 17 zones was introduced, extending to the Chemistry domain [16, 17]. This work was a step towards showing the applicability of the AZ theory to different domains. However, the small corpus issue is still present, with the annotation of only 30 papers from the Chemistry domain and 9 papers from the Computational Linguistics domain. Recently, deep learning methods have been used in automatic AZ identification [14], with the obvious conclusion that the AZ identification is sensitive to the type of embedding used. In addition to the CL and chemistry domains, the AZ theory was applied to other domains; e.g. biomedical, physics and biochemistry domains [10, 11, 13].

Argument mining is similar to AZ. Accuosto et al. [1–3] define argument mining as the automated process of identifying arguments, their components, and relationship within text. In subsequent studies, they proposed an annotation schema for identifying argumentative units and relations specifically in scientific abstract [1–3] and introduced an annotation schema that aimed to add argumentative components and relations to a small subset of 60 abstracts obtained from the SciDTB corpus [19]. Their objective was to identify the boundaries of each argumentative unit [2, 3], experimenting with sentences as argumentative units [1]. As a result of this work, the authors published a manually annotated corpus of 225 CL abstracts and 285 biomedical abstracts and used it to fine-tune a BERT based model for argument mining.

Bless et al. [4] introduced an annotation tool for LaTeX documents that differs from previous approaches in identifying argumentative zones. Their tool employs scientific knowledge graphs to annotate machine-actionable metadata (i.e. zones), specifically allowing researchers to annotate their publication while writing the manuscript.

To address the challenge of creating labeled corpora for AZ identification, we presented an AZ annotation platform [8] . Inspired by previous work [1, 18], we define a simplified AZ annotation schema. This schema is used to assign labels to selected sentences in scientific articles, which users can then correct or agree with, thus facilitating the annotation process.

3 Manual Annotation for AZ Corpus Creation

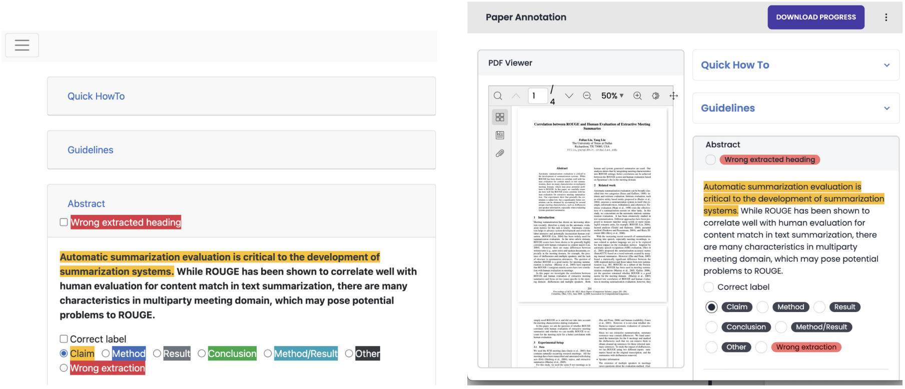
Aiming to create AZ corpus, we conducted an annotation task to collect sentences from selected scientific articles labelled with one of four predefined argumentative categories (i.e. zones): *Claim*, *Method*, *Result*, and *Conclusion*. We want to assure the good performance of the annotators by comparing their annotation against a ground truth and to assess the same level of understanding of annotators of the task. We did two annotation rounds with students from Telkom University[2] who were asked to annotate selected scientific articles using an AZ annotation platform [8]. The user interface of the platform in the second annotation round[3] differs slightly from the one used in the first round[4] (see

[2] https://telkomuniversity.ac.id/.

[3] https://riset.fanzru.dev/login.

[4] https://ir-group.ec.tuwien.ac.at/artu_az_annotation/.

Fig. 1), where the modifications include usability improvements collected in a feedback form after the first annotation round.


Quick HowTo
Guidelines
Abstract
□ Wrong extracted heading
□ Correct label
● Claim ○ Method ○ Result ○ Conclusion ○ Method/Result ○ Other
○ Wrong extraction
Paper Annotation
DOWNLOAD PROGRESS
PDF Viewer
1 / 4
50%
Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries
Quick How To
Guidelines
Abstract
○ Wrong extracted heading
○ Correct label
● Claim ○ Method ○ Result
○ Conclusion ○ Method/Result
○ Other ○ Wrong extraction

Automatic summarization evaluation is critical to the development of summarization systems. While ROUGE has been shown to correlate well with human evaluation for content match in text summarization, there are many characteristics in multiparty meeting domain, which may pose potential problems to ROUGE.

Automatic summarization evaluation is critical to the development of summarization systems. While ROUGE has been shown to correlate well with human evaluation for content match in text summarization, there are many characteristics in multiparty meeting domain, which may pose potential problems to ROUGE.

(a) V1 (b) V2

Fig. 1. The UI of the annotation platform: V1 is used in the first annotation round, and V2 in the second round.

In the following, we give more details about the AZ annotation platform, the participants per each annotation round, the selection of the scientific articles to annotate, and describe each annotation round.

3.1 AZ Annotation Platform

In this section, we give a brief overview of the annotation platform [8]. Our platform takes PDF scientific articles as input from the user, and selects/highlights sentences from each section of the article based on their similarity with the abstract sentences. We define four AZ categories that cover the main components of scientific articles: *Claim*, *Method*, *Result* and *Conclusion*. Each selected sentence is labeled with one of these categories where we use a pre-trained BERT model based on the approach proposed by Accuosto et al. [1] to predict the argumentative category of the sentences. We map each of the original argumentative category labels to one of our defined four AZ categories (Table 1). At the end of the process, the platform uses the annotations to create summaries for the annotated article. Evaluating the generated summaries, however, is not the focus of this paper.

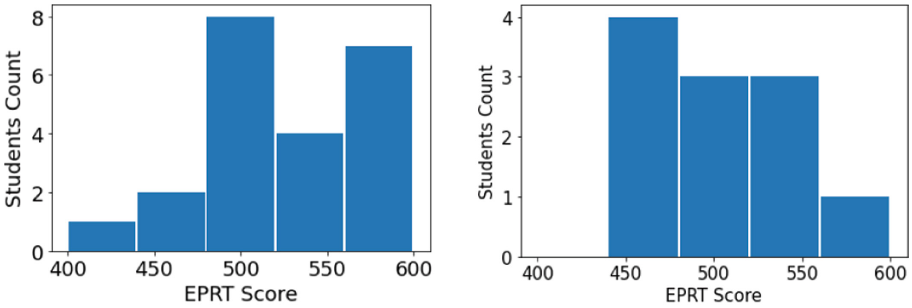
3.2 Participants

The study was conducted with bachelor and master students in their last year, who volunteered after call for participation in the annotation study. 22 and 11 students responded to the calls for the first and the second annotation rounds respectively. To understand their comprehension skills and English language proficiency, we asked the

Table 1. Mapping the Accuosto et al. [1] annotation schema to ours.

AZ categories in [1]	Our AZ categories
proposal	Claim
proposal_implementation	Method
observation	Result
result	Result
result_means	Method
conclusion	Conclusion
means	Method
motivation_problem	Claim
motivation_hypothesis	Claim
motivation_background	Claim
information_additional	Claim

students to fill a questionnaire in which they indicated their EPrT score[5] Fig. 2 shows the distribution of the EPrT scores among participants in round 1 (Fig. 2a) and round 2 (Fig. 2b). Most of the participants have score greater than 450 which shows that they can understand common phrases in academic text.



(a) Annotation round 1

(b) Annotation round 2

Fig. 2. Distribution of EPrT scores among participants.

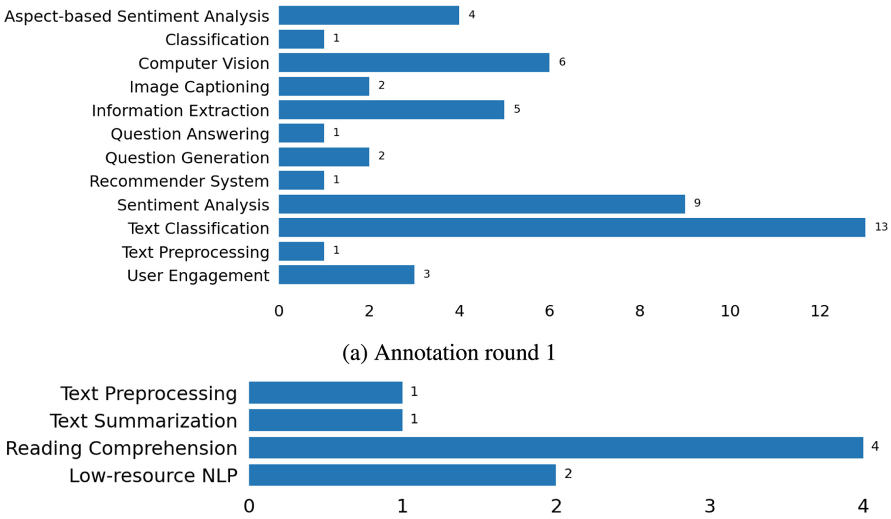
3.3 Selecting and Assigning Articles

We selected 48 and 8 scientific articles to annotate for the first and the second annotation rounds, respectively. Most of the articles were selected from the ACL anthology[6] as the annotators' background is in the Text Mining and Natural Language Processing

[5] English Proficiency Test. https://lac.telkomuniversity.ac.id/en/course/eprt-preparation/.

[6] https://aclanthology.org/.

domain. Figure 3 shows the distribution of the selected articles according to the topic of the article per each round. In each round, one of the selected papers was assigned to all of the students in the annotation round and was, thus, used to assess the quality of the student annotations against a ground truth annotation[7]. This allowed us to remove low quality annotations from the final set of annotations. The rest of the papers were assigned to three students, each, in both rounds so we can have later better judgement for the evaluation of the Inter-Annotator-Agreement (IAA) rates.



(a) Annotation round 1

(b) Annotation round 2

Fig. 3. Main topic distribution of the selected articles.

3.4 Annotation Rounds

As previously mentioned, two annotation rounds were held using two different versions of the platform's user interface while maintaining the same core model (i.e. backend):

Annotation Round 1: In this round we used the version of the platform which was published in our previous work [8]. Annotation instructions were given to the students via an online session and the guidelines[8] were provided as further material to them. The students were given a two weeks time frame to complete the annotations offline and deliver their annotated data.

[7] The first author of this paper is the ground truth annotator.

[8] Annotation guidelines available at: https://owncloud.tuwien.ac.at/index.php/s/lqyUgQmAbZg2cf3.

Annotation Round 2: based on the feedback collected from the previous round, in the second round we used a slightly modified interface. This round was held as an onsite one day workshop with the students. The annotation workshop motivation and instructions were presented to the students. In addition, assistance was given to the students during the annotation workshop by answering their questions.

4 Making Use of Annotations

In this section we describe how we make use of the collected annotations to train a classification model for the ***AZ identification*** task. In this task we predict the AZ category (i.e zone) of a given sentence from scientific article. That is, given a sentence from a scientific article we predict its argumentative category by labeling it with one of our predefined zones: *Claim*, *Method*, *Result*, or *Conclusion*.

We train a Bert model for AZ identification, the ***AZ-Bert*** model, following on the approach proposed by Accuosto et al. [1]. Using the same parameter settings as in Accuosto et al. [1], we train several AZ-Bert models on different training corpora and compare the models performance using the Computational Linguistic (CL) test corpus in Accuosto et al. [1].

We utilize the Computational Linguistic (CL) training corpus from Accuosto et al. [1] to train a baseline AZ-Bert model, we refer to this corpus as **SciArgCL** (Table 2). The SciArgCL is composed of 225 abstract sentences labeled with one of the 11 labels, as shown in Table 1. Before the training, we transform the original AZ categories of the SciArgCL corpus to our AZ categories (Table 1).

We use the collected annotations from each round to construct training corpus where we consider different combinations of data to build corpus for training each AZ-Bert model, as shown in Table 2. We use two strategies to construct corpus from the annotated data: (1) using the full annotated data without processing (we identify this data from the first and the second rounds with the ids **R1** and **R2** respectively), and (2) defining criteria to filter out the corpus from low quality annotations (identified by **FR1** and **FR2** for the first and the second round respectively). The details of constructing and filtering the training corpus are mentioned in Sect. 5.

Table 2. Description of corpora used to train AZ-Bert models in different experiments.

Description	Training data name	Sentences number
Baseline data set	SciArgCL	1048
Expansion with whole set of annotation	SciArgCL + R1	4268
	SciArgCL + R2	1556
	SciArgCL + R1 + R2	4776
Expansion while removing low quality annotations	SciArgCL + FR1	2997
	SciArgCL + FR2	1369
	SciArgCL + FR1 + FR2	3318
Collected annotation as a standalone corpus	FR1	1949
	FR2	321
	FR1 + FR2	2270

We have two types of the experiments: (1) *Expansion* experiments where we expanded the SciArgCL dataset with combinations of R1, R2, FR1 and FR2 (Table 2) to train AZ-Bert models and measure the impact of the expansion on the model performance, and (2) *Standalone* experiments where we use FR1 and FR2 to construct a standalone training corpus, from collected annotations, for AZ-Bert to measure whether we can construct a training corpus for AZ identification using our annotation platform. We consider only the FR1 and FR2 for the *Standalone* experiments because the AZ-Bert model performed better on the expanded data using FR1 and FR2.

5 Results

In this section, we present the results for (1) the annotation task (Sect. 3) by assessing the quality of the annotation using ground truth annotation and measure the same level of understanding for the annotators of the task by means of Inter Annotator Agreement (IAA), and (2) the AZ-Bert experiments (Sect. 4) by comparing the performance of trained AZ-Bert against a baseline model.

5.1 Annotation

Annotation performance: recall, in Sect. 3.3, that we assigned one paper to be annotated by all students during each annotation round[9]. The first author of this paper annotated both articles where we consider her annotation as ground truth annotation. We compute the metrics: *Precision*, *Recall*, and *F-measure* of the students' annotation against the ground truth to measure the students annotation performance. We compute each metric per each zone. In Table 3, we report the average performance among all students annotation per each zone and the weighted average (W. Average) performance. In terms of performance results, we notice the following:

1. Generally, sentences which belong to the *Claim* zone are easy to be identified this is because these sentences usually contain clear phrases that make them easy to be labeled (e.g. *"In this work, we propose"*, *"We present"*, etc.).
2. In the first round, the performance of identifying sentences belonging to the *Method* zone is very low. This is because the sentences that were extracted and originally labelled as *Method* sentences by the platform were sentences which described previous work and not the original work of the annotated paper. For example the sentence: "*In the SemEval 2017 Task 4 (Rosenthal et al., 2017), a thorough 5x coverage annotation scheme is used (each tweet is annotated by at least five people).*" describes methodology in a previous work which is cited by the article being annotated, however it was labelled by the platform as a *Method* sentence. According to our definition of the *Method* zone, that it should contain sentences that define methodology for the annotated paper not a previous work. For the case of a sentence belong to a

[9] In the first round: [Sentiment Analysis: It's Complicated!] (Kenyon-Dean et al., NAACL 2018) In the second round: [Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging] (Schmid & Laws, COLING 2008).

previous work, we define an extra category called *Other*, this category identifies sentences describing previous work and we ignore the sentences with *Other* category from the collected annotated corpus. However, assessing the annotation results, it became obvious that the use of the *Other* category was not clear to the students in the first round which we clarified for them in the second round.

3. The annotation performance in the second annotation round is higher than that of the first. This is expected since the second round took place on-site, under direct guidance. For clearer analysis of the second annotation round performance, we divided the annotators of this round into two groups: (1) *Old* annotators - participated in both rounds, and (2) *New* annotators - participated in the second round only and we calculated the weighted average performance of each group (rows OW. Average and NW. Average in Table 4). As expected the performance of *Old* annotators is higher than that of the *New* ones since they have prior knowledge of the task.

Table 3. The average of the students annotation performance (in terms of *Precision*, *Recall* and *F-measure*) during each annotation round per each zone and the weighted average performance of all zones.

Label	Round 1			Round 2		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Claim	**0.85 ± 0.13**	**0.87 ± 0.16**	**0.85 ± 0.12**	**0.95 ± 0.05**	**0.85 ± 0.17**	**0.89 ± 0.11**
Method	0.13 ± 0.28	0.27 ± 0.46	0.15 ± 0.30	0.86 ± 0.12	0.63 ± 0.11	0.72 ± 0.10
Result	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.61 ± 0.19	0.72 ± 0.25	0.61 ± 0.15
Conclusion	0.49 ± 0.50	0.53 ± 0.51	0.5 ± 0.5	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
W. Average	0.68 ± 0.20	0.61 ± 0.17	0.61 ± 0.17	**0.82 ± 0.06**	**0.70 ± 0.10**	**0.74 ± 0.08**
OW. Average				0.83 ± 0.05	0.71 ± 0.03	0.75 ± 0.03
NW. Average				0.80 ± 0.06	0.69 ± 0.13	0.72 ± 0.10

Inter Anotator Agreement (IAA): Each paper was annotated by three annotators in addition to the paper that was annotated by all of the students during each annotation round. We use these annotations to compute the pair wise inter-annotator's agreement using Kappa κ Cohen score to assess the same level of understanding of different annotators for the task definition. We report in Table 4 the average with the standard deviation, minimum and maximum of the pairwise agreements. Noticing the results in the first row, the agreements are *moderate* for both annotation rounds. We were expecting based on the annotation performance results (Table 3), that the agreement in the second annotation round should be higher than that of the first. For further analysis of the results, we filtered the annotations by removing low quality annotations considering the average F-measure (Table 3) as a threshold for the annotation quality and removed all instances of annotations for annotators with performance less than the threshold. Then, we recomputed the agreement using the filtered corpus (last three rows in Table 4). We notice that the interpretation of the agreement raised to *substantial* for both rounds. After filtering the annotators, we notice that the minimum agreement (the last row in

Table 4) of the second round is relatively high compared to the first round which shows that the annotators tends to have the same understanding of the task in the second annotation round more than the first round.

Table 4. The pairwise agreement between annotators using Cohen's κ coefficient.

Description	Cohen's κ	Round 1	Round 2
Considering whole set of annotation	Avg. pairwise	0.519±0.279	0.450 ± 0.143
	Maximum	1.000	0.754
	Minimum	0.013	0.058
Removing low quality annotations	Avg. pairwise	0.768±0.258	0.605 ± 0.095
	Maximum	1.000	0.754
	Minimum	0.090	0.419

5.2 AZ Identification

We looked at the usefulness of the collected annotated corpus in addressing the AZ identification task (recall Sect. 4). To construct the training corpus from collected annotated sentences, we label the sentences by considering the majority voting of papers assigned by multiple annotators, where we broke ties randomly, and the ground truth annotations for the single papers in both rounds. We trained several AZ-Bert models using different combination of training corpora, as shown in Sect. 4 and Table 2. For the expansion and creation of standalone corpus, we experiment per annotation round and merging corpora from both rounds. To filter out the low quality annotations, we consider the average F-measure (Table 3) as a threshold for the annotation performance where we removed all instances of annotations for annotators with performance less than the threshold to build the final filtered corpus.

Table 5 shows the performance of the AZ-Bert models trained using different corpora on the CL test set used by Accuosto et al. [1] in terms of *Precision*, *Recall* and *F-measure*. To assure that the results are not random, we repeat each experiment four times and we report for the average performance and the standard deviation. In the following, we discuss the results of each of the *Expansion* and *Standalone* experiments.

Expansion Experiments: We notice that the models trained by expanding SciArgCL with R2 and FR2 (**bolded** values in Table 5) achieves higher performance than the baseline and the best performance is achieved after filtering for low quality annotations (i.e. the row **SciArgCL + FR2**). We measure the significance *F-measure* improvement of the model trained using SciArgCL + FR2 over the one using SciArgCL using a t-test, we got $p\_value = 0.049$ which we interpret as statistical significant change. On the other hand, expanding SciArgCL with data collected from the first annotation round (i.e. R1 and FR1) does not help in improving the AZ-Bert performance, where the baseline is statistically significantly higher in performance. This result matches with the annotation performance results (see Table 3) where the annotation performance of the second

Table 5. Performance of repeated experiments (mean±std) of the AZ-BERT models trained on different corpus on the CL test set from [1].

Experiment Type	Training data	Results		
		Precision	Recall	F-measure
Baseline	SciArgCL	0.686±0.016	0.684±0.008	0.683±0.013
Expansion	SciArgCL + R1	0.622±0.043	0.609±0.040	0.614±0.041
	SciArgCL + R2	**0.692±0.028**	0.677±0.037	0.682±0.030
	SciArgCL + R1 + R2	0.613±0.042	0.602±0.038	0.607±0.040
	SciArgCL + FR1	0.632±0.021	0.628±0.013	0.629±0.017
	SciArgCL + FR2	**0.715±0.028**	**0.720±0.037**	**0.716±0.030**
	SciArgCL + FR1 + FR2	0.645±0.042	0.630±0.038	0.636±0.040
Standalone	FR1	0.628±0.008	0.564±0.015	0.589±0.010
	FR2	0.581±0.056	0.570±0.021	0.568±0.024
	FR1 + FR2	0.639±0.005	0.585±0.007	0.610±0.006

annotation round is better than that of the first. When the annotation instructions were carefully clarified for the students, the annotators performance increased and it helps in an overall improvement of the AZ identification task. This verifies also the usefulness of collecting annotated corpus using our AZ annotation platform, when the annotation quality is high, to extent corpus for AZ identification.

Standalone Experiments: with these experiments, we aim to study the usefulness of the annotation platform to construct a standalone AZ corpus using collected annotations. We chose to train AZ-Bert models with the filtered corpus (i.e. FR1 and FR2) only and ignore the whole set based on the *Expansion* experiments results. As shown in Table 5, the performance of the AZ-Bert model built using the standalone corpus is significantly less than the baseline which is verified by a t-test which gives a $p\_value < 0.05$ when we measure the significance of the high value of the baseline F-measure with respect to the models trained using the standalone corpus. The performance of the AZ-Bert model of trained with FR1+FR2 achieves higher performance than that from each round alone (i.e. FR1 alone and FR2 alone) with significant F-measure improvement ($p\_value = 0.005$). This result shows that increasing collected corpus with more annotated data helps in the AZ-Bert model improvement. It is expected that results from the collected annotation would not improve over the baseline, but we assume that this result accepted given that: (1) the annotations are done in a semi-automatic way which reduces the effort and the time for the annotation process, (2) they are done by students compared to expert annotators who build the SciArgCL corpus [1], (3) the number of collected annotated articles are fewer compared to the base line (56 articles vs. 225 articles), and (4) the tool helps to collect annotation on paper level with less effort, by suggesting automatic annotations, if it is compared to do annotation on abstract level [1].

6 Conclusion and Future Work

With the aim to solve the problem of creating benchmark data for Argumentative Zoning (AZ) identification, we proposed in a previous work an AZ annotation platform that helps user to annotate given PDF scientific articles with a simplified AZ schema of four zones: *Claim*, *Method*, *Result*, and *Conclusion*. In this paper, we present our work on the design and execution of an annotation experiment to collect sentences from scientific articles labeled with AZ categories, using the platform. The experiment consisted of two annotation rounds, online and onsite, with bachelor and master students from Telkom University. The aim of the annotation experiment was to collect AZ annotated corpus where we evaluated the students annotation performance using ground truth annotation and using the agreement between annotators to analyse the students understanding of the task. We utilize the collected annotations to train AZ-Bert models using different training corpora and compare the performance of the trained models with an AZ-Bert model trained on a baseline corpus (SciArgCL). We experiment with two settings: expanding the SciArgCL corpus with collected annotations and using the annotations as a stand alone training corpora. Though only one model achieved better performance over the baseline, we consider that the performance is accepted given that the platform helped to reduce the cost of the annotation process and creating AZ corpus in terms of time and effort and without need of domain experts.

As future work, we plan to use the platform to create benchmark data set which helps for scientific articles summarisation. By its original design, the platform generates two types of article summaries at the end of the annotation process; one is based on improving a previous work [7] and the other using the users annotations. We collected feedback for the generated summaries using pre and post-questionnaires during the annotation rounds described in this paper. We are planning to use the collected feedback to refine the summarisation pipeline as a step to build informative summaries [15] for scientific articles using the argumentative zones. In addition, we plan to analyse the potential of the tool to create AZ corpora for domains other than the Computational Linguistics domain.

Acknowledgement. This work was a part from the *Artificial Researcher in Science: Efficient Publication Mining* project (ID:2241716) and supported by the ASEAN European Academic University Network (ASEA-UNINET), Project Number: ASEA 2021 22/TU/Proj. 1

References

1. Accuosto, P., Neves, M., Saggion, H.: Argumentation mining in scientific literature: from computational linguistics to biomedicine. In: Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval, 1 April 2021, Lucca, Italy. Aachen: CEUR; 2021, pp. 20–36. CEUR Workshop Proceedings (2021)
2. Accuosto, P., Saggion, H.: Transferring knowledge from discourse to arguments: a case study with scientific abstracts. In: Stein, B., Wachsmuth, H., (eds.) Proceedings of the 6th Workshop on Argument Mining, 1 August 2019, Florence, Italy. Stroudsburg: Association for Computational Linguistics, pp. 41–51. ACL (Association for Computational Linguistics) (2019)

3. Accuosto, P., Saggion, H.: Mining arguments in scientific abstracts with discourse-level embeddings. Data Knowl. Eng. **129**, 101840 (2020). https://doi.org/10.1016/j.datak.2020.101840, https://www.sciencedirect.com/science/article/pii/S0169023X20300446
4. Bless, C., Baimuratov, I., Karras, O.: Scikgtex - a latex package to semantically annotate contributions in scientific publications (2023)
5. Contractor, D., Guo, Y., Korhonen, A.: Using argumentative zones for extractive summarization of scientific articles. In: Proceedings of COLING 2012, pp. 663–678. The COLING 2012 Organizing Committee, Mumbai, India, December 2012. https://aclanthology.org/C12-1041
6. E. Mendoza, Ó., et al.: Benchmark for research theme classification of scholarly documents. In: Proceedings of the Third Workshop on Scholarly Document Processing, pp. 253–262. Association for Computational Linguistics, Gyeongju, Republic of Korea, October 2022. https://aclanthology.org/2022.sdp-1.31
7. El-Ebshihy, A., Ningtyas, A.M., Andersson, L., Piroi, F., Rauber, A.: ARTU/TU Wien and artificial researcher@ LongSumm 20. In: Proceedings of the First Workshop on Scholarly Document Processing, pp. 310–317. Association for Computational Linguistics, November 2020. https://doi.org/10.18653/v1/2020.sdp-1.36, https://www.aclweb.org/anthology/2020.sdp-1.36
8. El-Ebshihy, A., Ningtyas, A.M., Andersson, L., Piroi, F., Rauber, A.: A platform for argumentative zoning annotation and scientific summarization. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM 2022, pp. 4843–4847. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3511808.3557193
9. Goldsack, T., Zhang, Z., Lin, C., Scarton, C.: Domain-driven and discourse-guided scientific summarisation. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval, pp. 361–376. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-28244-7_23
10. Guo, Y., Korhonen, A., Liakata, M., Silins, I., Sun, L., Stenius, U.: Identifying the information structure of scientific abstracts: an investigation of three different schemes. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, pp. 99–107. Association for Computational Linguistics, Uppsala, Sweden, July 2010. https://aclanthology.org/W10-1913
11. Guo, Y., Silins, I., Stenius, U., Korhonen, A.: Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. Bioinformatics **29**(11), 1440–1447 (2013). https://doi.org/10.1093/bioinformatics/btt163
12. Liakata, M., Dobnik, S., Saha, S., Batchelor, C., Rebholz-Schuhmann, D.: A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 747–757. Association for Computational Linguistics, Seattle, Washington, USA, October 2013. https://aclanthology.org/D13-1070
13. 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 2010). European Language Resources Association (ELRA), Valletta, Malta, May 2010. http://www.lrec-conf.org/proceedings/lrec2010/pdf/644_Paper.pdf
14. Liu, H.: Automatic Argumentative-Zoning Using Word2vec. CoRR abs/1703.10152 (2017). http://arxiv.org/abs/1703.10152
15. Saggion, H., Lapalme, G.: Generating indicative-informative summaries with SumUM. Comput. Linguist. **28**(4), 497–526 (2002). https://doi.org/10.1162/089120102762671963, https://www.aclweb.org/anthology/J02-4005

16. Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Comput. Linguist. **28**(4), 409–445 (2002)
17. Teufel, S., Siddharthan, A., Batchelor, C.: Towards domain-independent argumentative zoning: evidence from chemistry and computational linguistics. In: Proceedings of the 2009 Conference on Empirical Methods in natural language processing, pp. 1493–1502 (2009)
18. Teufel, S., et al.: Argumentative zoning: Information extraction from scientific text. Ph.D. thesis, Citeseer (1999)
19. Yang, A., Li, S.: SciDTB: discourse dependency TreeBank for scientific abstracts. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 444–449. Association for Computational Linguistics, Melbourne, Australia, July 2018. https://doi.org/10.18653/v1/P18-2071, https://aclanthology.org/P18-2071
20. Yasunaga, M., et al.: ScisummNet: a large annotated corpus and content-impact models for scientific paper summarization with citation networks. In: Proceedings of AAAI 2019 (2019)