DOI: 10.14529/jsfi250304

Document-Level Approach to Extracting Argumentation Structures from the Russian Texts of Scientific Communication

Elena A. Sidorova[1], *Irina R. Akhmadeeva*[1], *Daria V. Ilina*[1], *Irina S. Kononenko*[1], *Alexey S. Sery*[1], *Yury A. Zagorulko*[1]

The study addresses the problem of automatic extraction of argumentative structures in scientific communication texts in Russian. Such texts are characterized by a branched logical structure, including distant references and interrelations. To address these complexities, recent methodological advances attempt to leverage the text itself as a contextual foundation for extracting connections. This study presents a generative approach for extracting argumentative relations, reframing the prediction task as a problem of generating marked-up text and making it an end-to-end approach, rather than the traditional pipeline. Two Russian-language corpora were used in the experiments: the translated corpus of microtexts ruMTC and the annotated corpus of scientific communication texts ArgNetSC. A comparative analysis was conducted to evaluate the performance of T5 architecture models trained with supervised fine-tuning (SFT) and Large Language Models on various Russian-language datasets. To facilitate the analysis of long texts, a text segmentation method using a sliding window was proposed. The evaluation revealed that the highest performance in argumentative relation extraction was consistently achieved on the corpus of microtexts. Notably, the smaller models fine-tuned using the SFT method and large language models that were prompted to generate marked texts demonstrated comparable performance ($F_1 \sim 0.32 - 0.37$). For larger texts, however, this trend did not persist, as the FRED-T5 model outperformed all other models with $F_1 \sim 0.23$ on texts related to the genre of scientific articles.

Keywords: argument mining, document-level argument relation prediction, long-range argumentative relation, text2text generative language model, scientific communication.

Introduction

One of the important areas of research in scientific communication, represented by scientific and popular science texts, is the study of the logical organization of reasoning that presents and substantiates the author's position from various points of view. In the process of such reasoning, in order to convince the audience, arguments formulated in the form of premises and conclusions are given in favor of or against the thesis under consideration. The author not only proves some positions through logical reasoning, but also mentally debates with an opponent, modeling possible counterarguments. In this case, a separate argument can act as an initial premise for constructing a new argument, and its conclusion is often used as a justification for another statement. In addition, different arguments can have common premises or conclusions. As a result, they all turn out to be interconnected and form a holistic system that can be represented as an argumentation graph.

Solving problems in the field of argumentation mining requires text corpora with annotated argumentative structures. In recent years, there has been an increase in the volume and diversity of annotated data, but most works are limited to using a few of the most well-known and widely used corpora, according to [11]. Less popular datasets, unfortunately, are often ignored. This is due, first of all, to the desire to compare new methods with existing ones based on uniform benchmarks. However, as the authors of the study note, such a practice is often criticized, since the benchmark data does not always reflect the features of real texts in terms of topic and genre.

[1] A.P. Ershov Institute of Informatics Systems, Novosibirsk, Russian Federation

Not only the language and topic of the text, but also its volume can have a significant impact on the process of extracting argumentation, as the length of the possible argumentative connection increases, hence the number of pairs of statements that can potentially be related.

The largest amount of argumentatively annotated data is available for the English language, while datasets for intellectual analysis of argumentation in Russian are extremely scarce. The following datasets are known for the Russian language:

- Argumentative Microtext Corpus (ruMTC) – a corpus of argumentative essays translated into Russian, with the original argumentative annotation automatically transferred from the original texts [10];
- RuArg-2022 – a corpus of comments from users of the VKontakte social network on news texts about COVID-19 [14]; the annotation model belongs to the APE (Argument Pair Extraction) class, where for a given thesis, supporting and attacking statements are found in different texts; in this corpus, a set of statements is specified, each of which is marked as «for», «against» the thesis, or has a neutral status;
- ArgNetSC – a corpus of scientific communication texts annotated on the ArgNetBankStudio resource based on D. Walton's model [30] being the traditional model of argumentative markup.

While the first two corpora contain rather short texts with contact or short-distance relations between statements, the third corpus contains longer texts as well, that are distinguished by greater structural-content complexity determined by the organization and logical connection of their parts. Scientific communication texts are characterized by a branched logical structure, with the presence of distant references and long-distance relations between content elements. At the same time, argumentative relations are implemented at the level of the entire text, and not only within a sentence, adjacent sentences or paragraphs. To take into account such long-distance relations, Document-Level approaches have recently been actively developing, which use the entire text as a context for finding connections.

In this paper, we propose to use a generative approach to solve the Document-Level Argument relation prediction problem, which, firstly, solves the relation prediction problem not as a classification problem, but as a problem of generating marked-up text, and secondly, uses an end-to-end approach to Argument Mining (E2E-AM) instead of the traditional pipeline [16], in which argument analysis is divided into separate modules trained and applied sequentially. Unlike pipeline frameworks, end-to-end frameworks jointly optimize all subtasks by studying global characteristics and dependencies, which allows us to obtain a holistic view of argument structures. In this paper, we focus on the following research questions.

RQ1. What is the quality of the solution of the End-to-End Argument Mining problem using the Document-Level approach implemented as a text generation task?

RQ2. How does the genre and volume of the text affect the quality of argumentation extraction?

We conducted comparative experiments on two Russian-language datasets: a) the Argumentative Microtext Corpus in Russian (hereinafter – ruMTC), obtained by manually translating the first part of the corpus of the same name from English [10], and b) the ArgNetSC corpus of scientific communication in Russian [30].

The article has the following structure. Section 1 is devoted to a review of the scholarly literature on the research problem. Section 2 describes the datasets used in the study and the data preparation. Section 3 presents the experiment and its result and provides an analysis of common errors. In Section 4, the results are discussed. Conclusion summarizes the study and points directions for future work.

1. Related Work

The main task of argumentation analysis is to extract argumentatively related statements from the text based on formal models. The formal argument model proposed by Toulmin in his work [31] includes 6 components. But in practice, simplified representations of the argument structure are used for data annotating, including 2 components – premises and conclusions. Thus, the argumentative structure can be represented as a binary relation linking a pair of statements, one of which (*premise*) supports or refutes the second (*conclusion*).

There are many datasets, in which such relations were annotated: IAC [32], NoDE (Natural language arguments in online DEbates) [5], UKP-PE [27, 28], RuArg-2022 [14], etc. The Argumentative Microtext Corpus [24] and its Russian-language version [10] were annotated following more complex schemes, reducing, however, the complex arguments to sets of binary relations.

Typically, each corpus consists of texts of a specific genre. For example, the CDCP corpus [22] is used to analyze legal documents, the AbstRCT corpus [21] – for medical-related research, and the DrInventor [15] and SciDTB [1] corpora are used to analyze scientific publications and abstracts, respectively. The UKP-PE corpus of short essays [28] is widely known.

The standard solution of argumentation analysis is to build a pipeline that sequentially solves the following problems: identifying argumentative segments (ADUs), establishing the ADU type, determining the argument type and establishing relations between ADUs [16]. To solve these problems, BERT and BERT-based models are traditionally used [26, 33]. When extracting argumentative relations at short distances (if the premise and conclusion are within the same sentence or in adjacent sentences), the use of discourse markers and argumentation indicators [25], rhetorical relations [2] contributes to improving the results, but they are of little help when extracting long-distance relations.

Recent document-level approaches fall into several categories: sequence labeling methods (e.g., BIOLabel [29]); global context methods that capture document-wide information through question-answering frameworks (DocMRC [19]) or memory mechanisms (MemNet [9]); generative methods that use sequence-to-sequence models (e.g., BART-Gen [17]) for argument extraction, etc.

In general, the pipeline approach has been criticized: pipeline experiments, in general, suffer from the fact that error propagation occurs not only within each step, but also from one to another; the inflexibility of the models used is also noted [35]. In this regard, alternative approaches have recently been developed: end-to-end argument extraction methods based on a network architecture built on a biaffine parser [8, 35] and the use of Text2Text generative models [13]. In [7], an approach based on a biaffine dependency parser was applied to Russian-language texts, which also used rhetorical trees to clarify the boundaries of ADUs.

The idea of considering the AM task as a text generation task arose from related areas of NLP. Thus, the Translation between Augmented Natural Languages methodology [3, 12, 23] uses a pre-trained T5 encoder-decoder model, which has proven its effectiveness in the tasks of extracting relationships between entities, resolving coreference, constructing RST structure, etc.

In recent years, with the growth of pre-training methods, the development of a unified generative structure for solving a variety of tasks within a specific field has attracted increasing attention [6, 18, 20, 34]: solving various subtasks of named entity recognition, information extraction, tonality analysis and other areas, such as understanding dialogue and multimodal referencing. The paper [4] presents a unified generative platform (UniASA) adapted for various tasks of structured argument analysis: a) E2E-AM Task, b) Argument Pair Extraction (the task is designed to extract pairs of arguments discussing the same point from two interrelated documents), c) Argument Quadruplet Extraction (sentence-level, four-component argument structure used mainly for discussion analysis).

Our work extends a similar approach to E2E-AM by applying it to a Russian-language dataset, characterized by a more complex conceptual and argumentative structure. This adaptation necessitated several key modifications: the development of a novel text annotation scheme, an investigation into the significance of argument sequencing – a problem salient in longer texts that remains unaddressed in prior literature – and the segmentation of lengthy texts into chunks.

2. Datasets

Two annotated text corpora were used in the study: 1) **ruMTC** – the first part of the English language Argumentative Microtext Corpus translated into Russian [10] and 2) **ArgNetSC** – the annotated corpus of scientific communication texts.

The corpus of microtexts is widely known and is often mentioned in studies on automatic argumentation analysis. It includes 112 texts (576 sentences) on various topics up to 10 sentences long. Each ADU (in this dataset, each sentence is an ADU) is labeled as supporting or disputing the main thesis of the text; statements are organized into a graph with the following relations: «support», «rebuttal» (attack to an ADU), «undercut» (attack to a relation between statements), «additional» (for combining multiple premises) and «example» (support by example) [24].

A subset of 160 texts was selected from the **ArgNetSC** corpus – a Russian-language corpus with complex argumentative markup. This dataset comprised short and medium-length texts from three subgenres: 30 popular science news texts (**News**), 30 academic paper reviews (**Reviews**), and 100 full-length academic papers (**Articles**). The inclusion of texts with considerable non-argumentative content was found to introduce noise, as they were causing argument-free chunks. Consequently, such texts were systematically filtered out during the dataset compilation process. The length of the texts and the specifics of the genre and topic determine the presence of long-distance argumentative relations, i.e. links between statements that are at least one paragraph apart. The average span of these relations was 330, 502, and 793 characters for the three subcorpora, respectively, notably exceeding their average paragraph lengths (188, 240, and 301 characters). Identifying such relations presents a significant NLP challenge, as the candidate pair space grows combinatorially and grammatical mechanisms become less effective over long distances. Russian-language corpora with argumentative markup, presented in Tab. 1, were used to prepare the data.

Table 1. Data statistics

Corpus	Number of texts	Mean text length (in symbols)	Mean number of words	Mean number of sentences	Mean number of arguments
ruMTC	112	446	61	4	4
Reviews	30	2860	356	19	20
News	30	4105	521	28	34
Articles	100	9431	1165	63	54

Argumentative relations were simplified in the manner shown in Fig. 1 for training data preparation. Each relation R with more than one premise was decomposed into simple binary relations linking each of the premises $p_1, \dots, p_n$ with a conclusion C of the same type R. Each relation between a statement and another argument was replaced by a relation between the statement and the premises of that argument. A simplified argument structure was adopted to establish a baseline, reducing model complexity. While such an approach may lead to a risk of overlooking certain high-order relations, we believe it provided a necessary foundation for future work.

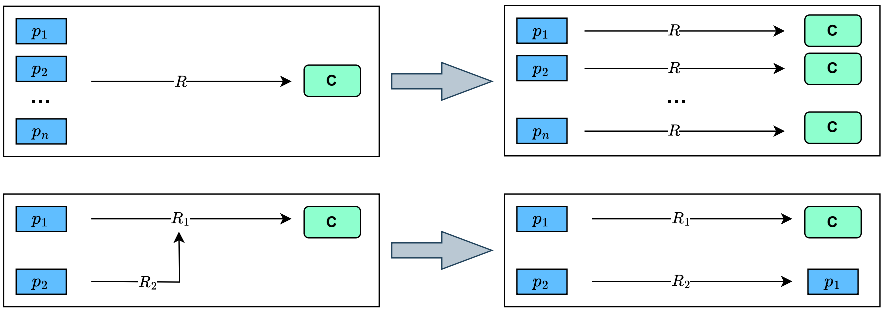
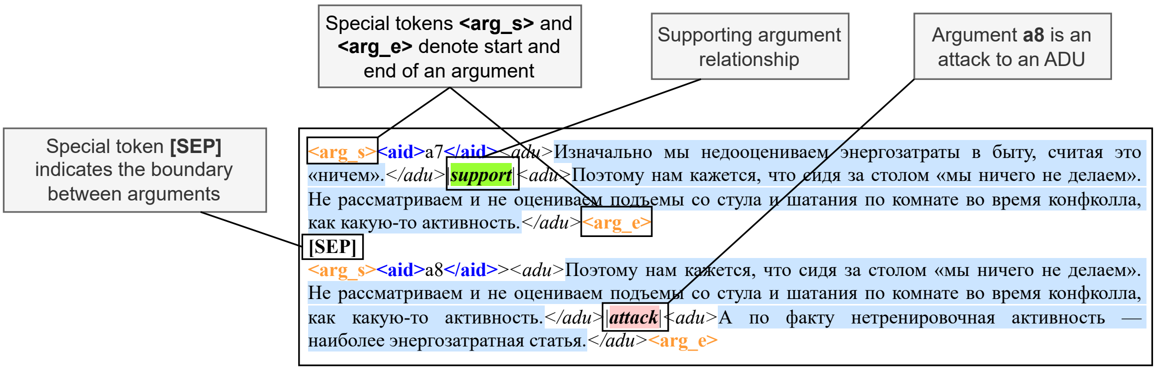


Figure 1. Transformation of a complex argument structure into binary relations; *P – premise*, *R – relation*, *C – conclusion*

When making datasets, we developed a specialized annotation scheme, enclosing the structural elements of arguments with special tokens (Fig. 2).


Special tokens **<arg_s>** and **<arg_e>** denote start and end of an argument
Supporting argument relationship
Argument **a8** is an attack to an ADU
Special token **[SEP]** indicates the boundary between arguments

<arg_s><aid>a7</aid><*adu*>Изначально мы недооцениваем энергозатраты в быту, считая это «ничем».</*adu*>|***support***|<*adu*>Поэтому нам кажется, что сидя за столом «мы ничего не делаем». Не рассматриваем и не оцениваем подъемы со стула и шатания по комнате во время конфколла, как какую-то активность.</*adu*><arg_e>
[SEP]
<arg_s><aid>a8</aid>><*adu*>Поэтому нам кажется, что сидя за столом «мы ничего не делаем». Не рассматриваем и не оцениваем подъемы со стула и шатания по комнате во время конфколла, как какую-то активность.</*adu*>|***attack***|<*adu*>А по факту нетренировочная активность — наиболее энергозатратная статья.</*adu*><arg_e>

Figure 2. An example of a text annotated using special tokens

ADU boundaries were marked with the tags `<adu>` and `</adu>`. The type of argumentative relation was marked with the symbols | (pipe). Two types of argumentative relations were considered: *support* and *attack*. The order of arguments in reference markups is of significant importance. It was observed that if the order of arguments in the markup differed from their order in the text, the model did not learn well. In the current study, arguments in the marked-up text are arranged in the order in which the premises of these arguments appear in the text.

3. Experiments

The experiments were conducted using the generative approach, allowing various argument mining tasks to be viewed as the task of generating a set of arguments. The study included a comparative evaluation of small T5-based models trained using SFT, as well as Large Language Models (LLMs) from the GPT family, across different datasets.

3.1. Implementation

Pre-trained language models supporting Russian language were used in the experimental study.

1. **mT5-base** (Multilingual T5, **580M**) is a multilingual model based on the T5 architecture;
2. **ByT5-base** (**582M**) is a modification of T5 that does not use a tokenizer and works directly with UTF-8 bytes; this model can handle any language, is more robust to noise (e.g., typos), and is easier to use because it does not require additional preprocessing;
3. **FRED-T5-large** (**820M**) is a model for Russian language based on the T5 architecture;
4. **Gemma 3 (27B)** is a multilingual and multimodal LLM that supports long context and vision inputs;
5. **gpt-oss: 20b/gpt-oss:120b** are LLMs, which are considered good for performance-efficiency trade-off; particularly the gpt-oss-120b model, which performs comparably to OpenAI's proprietary o4-mini on many benchmarks, while the smaller gpt-oss-20b competes with o3-mini.

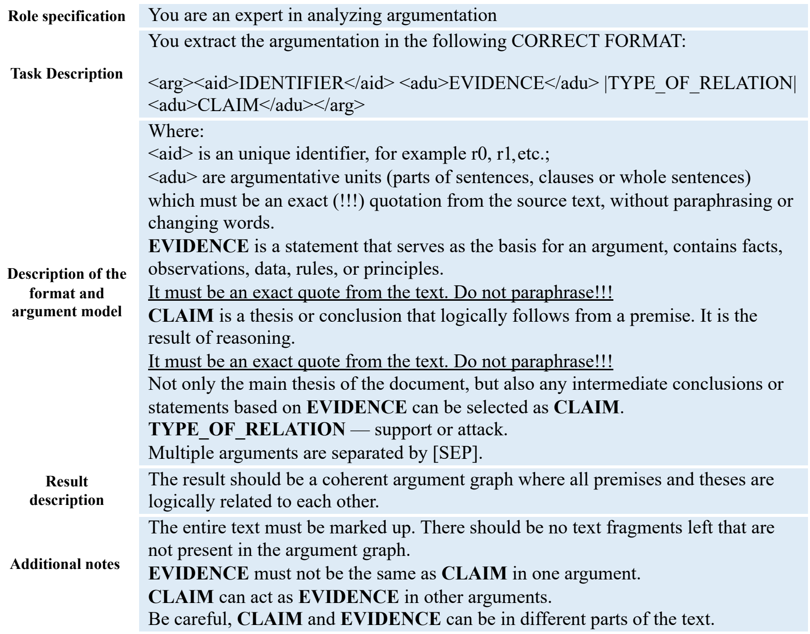
We employed SFT to train models of the **T5** architecture. The training was conducted over 20 to 50 epochs, utilizing a starting learning rate of 5×10^{-4}. To process lengthy documents from the **News**, **Reviews** and **Articles** datasets, a sliding window algorithm was employed for segmentation. This method generated overlapping chunks without regard for inherent textual units (e.g., sentences or paragraphs). The annotated text's length, averaging twice that of the source, dictated a maximum chunk size of half the model's context window. Therefore, the mT5-base model that is pretrained on 512-token sequences, was fed 256-token chunks, and the FRED-T5 model (4096-token context) received 512-token chunks. A larger chunk size for FRED-T5 was not possible due to limited computational resources. Annotations for a chunk were derived only from arguments fully contained within it, which led to the expected loss of long-range or oversized relations. This loss was measured at 21% for the 256-token chunking strategy and 7% for the 512-token strategy.

For the experiments with LLMs, texts were also divided into chunks, which, unlike the previous experiment, were aligned along sentence (paragraph) boundaries. This allowed to exclude incomplete contexts that may introduce noise from the source data for prompt. Note that both LLMs were employed in a zero-shot setting. The experiments were conducted locally using Ollama, the temperature was set to 1.0. Figure **??** shows the prompt template.

The prompt included the role specification that limited the subject area, problem statement and detailed description of the structure of the expected response. When analyzing the results of LLMs, a number of features were identified that allowed us to adjust the prompt.

– The model tended to paraphrase the original text. To counter such behavior we added the following requirement to the description of each component of the markup structure: *Must be an exact quote from the text. Do not paraphrase!!!.*

– The model tended to pay attention only to the main thesis and directly related arguments, so we also added the following requirement: *You must mark up all the text. There should be no fragments of the text that are not present in the argumentation graph.*



Role specification	You are an expert in analyzing argumentation
Task Description	You extract the argumentation in the following CORRECT FORMAT: <arg><aid>IDENTIFIER</aid> <adu>EVIDENCE</adu> |TYPE_OF_RELATION| <adu>CLAIM</adu></arg>
Description of the format and argument model	Where: <aid> is an unique identifier, for example r0, r1,etc.; <adu> are argumentative units (parts of sentences, clauses or whole sentences) which must be an exact (!!!) quotation from the source text, without paraphrasing or changing words. **EVIDENCE** is a statement that serves as the basis for an argument, contains facts, observations, data, rules, or principles. It must be an exact quote from the text. Do not paraphrase!!! **CLAIM** is a thesis or conclusion that logically follows from a premise. It is the result of reasoning. It must be an exact quote from the text. Do not paraphrase!!! Not only the main thesis of the document, but also any intermediate conclusions or statements based on **EVIDENCE** can be selected as **CLAIM**. **TYPE_OF_RELATION** — support or attack. Multiple arguments are separated by [SEP].
Result description	The result should be a coherent argument graph where all premises and theses are logically related to each other.
Additional notes	The entire text must be marked up. There should be no text fragments left that are not present in the argument graph. **EVIDENCE** must not be the same as **CLAIM** in one argument. **CLAIM** can act as **EVIDENCE** in other arguments. Be careful, **CLAIM** and **EVIDENCE** can be in different parts of the text.

Figure 3. The prompt template for argument extraction

3.2. Results

Table 2 summarizes the results of the conducted experiments. We utilized Precision (P_{adu}), Recall (R_{adu}), and $F1$ score ($F1_{adu}$) as evaluation metrics for ADU Extraction part and $F1$-score for extraction of unlabeled ($F1_{urel}$) and labeled (support, attack) relations ($F1_{rel}$). ADUs were compared at the character level using the Dice coefficient (equivalent to the F_1 score), and partial matches above a predefined threshold (equal to 0.8) were considered correct.

For the SFT experiments, the data were randomly partitioned into training, validation, and test sets, comprising 72%, 8%, and 20% of the total data, respectively. To ensure statistical robustness, this procedure was repeated across 10 distinct stratified splits (folds) for each dataset, and a full training-validation-testing cycle was conducted on each fold. LLM-based experiments were conducted on the test sets of each fold. The results, reported in Tab. 2, are presented as the mean performance across all folds with a 95% confidence interval.

The **ByT5** model was applied exclusively to the **ruMTC**-based dataset. Operating directly on UTF-8 bytes without a tokenizer caused the attention tensor to expand rapidly, which – given available computational resources – precluded its application to longer texts from ArgNetSC.

As it can be seen from the results, the highest result of the SFT approach was obtained on microtexts of the **ruMTC** corpus. Small sizes of both source and marked-up texts allowed to fit them completely into the context window of the model without information loss. This can also explain the lack of quality improvement for this dataset in argument extraction when moving to a larger model (**mT5** vs. **FRED-T5**).

LLMs applied to the **ruMTC** data demonstrated performance comparable to that of fine-tuned models. The highest performance for extracting argument relations (without type classification) was achieved by the **gpt-oss-120b** model ($F1_{urel} = 0.42$). When relation types were considered, the **gpt-oss-120b** model performed best on the same data ($F1_{rel} = 0.37$).

On texts of other genres, the fine-tuned models performed better than LLMs, and in total the results were expectedly lower. This may be due to both the presence of non-argumentative

Table 2. Experimental results

Dataset		Model	ADU Extraction			$F1_{urel}$	$F1_{rel}$
			P_{adu}	R_{adu}	$F1_{adu}$		
ruMTC		mT5-Base	0.86 ± 0.02	0.78 ± 0.03	0.81 ± 0.02	0.39 ± 0.03	0.32 ± 0.03
		ByT5-Base	0.85 ± 0.01	0.64 ± 0.03	0.72 ± 0.02	0.24 ± 0.02	0.17 ± 0.02
		FRED-T5-Large	**0.92 ± 0.03**	**0.91 ± 0.02**	**0.91 ± 0.03**	0.38 ± 0.06	0.31 ± 0.05
		gpt-oss:20b	0.77 ± 0.05	0.75 ± 0.04	0.74 ± 0.04	0.33 ± 0.05	0.29 ± 0.04
		gemma3:27b	0.86 ± 0.03	0.82 ± 0.03	0.84 ± 0.03	0.34 ± 0.04	0.30 ± 0.03
		gpt-oss:120b	0.80 ± 0.03	0.82 ± 0.02	0.81 ± 0.03	**0.42 ± 0.04**	**0.37 ± 0.03**
ArgNetSC	Reviews	mT5-Base	**0.71 ± 0.04**	0.52 ± 0.05	0.58 ± 0.04	0.10 ± 0.03	0.09 ± 0.03
		FRED-T5-Large	0.70 ± 0.04	0.64 ± 0.05	0.65 ± 0.04	**0.19 ± 0.03**	**0.18 ± 0.03**
		gpt-oss:20b	0.60 ± 0.03	0.56 ± 0.05	0.56 ± 0.04	0.06 ± 0.02	0.04 ± 0.02
		gemma3:27b	0.62 ± 0.02	**0.78 ± 0.03**	**0.68 ± 0.02**	0.13 ± 0.01	0.12 ± 0.01
		gpt-oss:120b	0.61 ± 0.03	0.68 ± 0.06	0.63 ± 0.04	0.15 ± 0.02	0.13 ± 0.02
	News	mT5-Base	0.58 ± 0.05	0.55 ± 0.03	0.55 ± 0.03	0.10 ± 0.02	0.10 ± 0.02
		FRED-T5-Large	0.64 ± 0.05	0.57 ± 0.03	**0.59 ± 0.03**	**0.14 ± 0.01**	**0.13 ± 0.01**
		gpt-oss:20b	0.49 ± 0.03	0.45 ± 0.03	0.46 ± 0.02	0.07 ± 0.02	0.06 ± 0.01
		gemma3:27b	**0.66 ± 0.03**	0.49 ± 0.01	0.55 ± 0.02	0.11 ± 0.03	0.11 ± 0.03
		gpt-oss:120b	0.59 ± 0.03	**0.59 ± 0.02**	0.58 ± 0.02	0.13 ± 0.03	**0.13 ± 0.03**
	Articles	mT5-Base	**0.70 ± 0.04**	0.43 ± 0.06	0.51 ± 0.05	0.11 ± 0.007	0.10 ± 0.006
		FRED-T5-Large	0.70 ± 0.04	**0.83 ± 0.01**	**0.74 ± 0.02**	**0.25 ± 0.01**	**0.23 ± 0.01**
		gpt-oss:20b	0.47 ± 0.01	0.62 ± 0.02	0.52 ± 0.01	0.08 ± 0.01	0.08 ± 0.005
		gemma3:27b	**0.73 ± 0.04**	0.61 ± 0.07	0.66 ± 0.05	0.22 ± 0.03	0.17 ± 0.03
		gpt-oss:120b	0.54 ± 0.02	**0.83 ± 0.02**	0.65 ± 0.02	0.12 ± 0.01	0.11 ± 0.01

zones and the loss of long-range relations and the fact that chunks often did not correspond to the logical structure of the text: paragraphs, sections or sentences.

The recall of ADU detection on texts in the ArgNetSC corpus improved significantly when utilizing the larger **FRED-T5 model**, especially on the **Articles** (0.42 vs. 0.82). The average length of an ADUs in **Articles** is greater than in **News** and **Reviews**, so expanding the context window was particularly critical for processing articles. Due to the smaller context size seen during training, some ADUs were «lost» by the **mT5** model.

3.3. Analysis of Argumentation Extraction Errors

Error analysis was conducted by experts who participated in the annotation of scientific communication texts. Tables of comparison of arguments found by models with reference ones (type I errors) and tables of comparison of reference arguments with arguments found by the model (type II errors) were considered separately.

3.3.1. Segmentation errors

Argumentative analysis involves identifying text fragments that make sense from the point of view of argumentation – ADUs. In general, they are statements based on propositions. Whole sentences are most often identified automatically. However, the analysis of manual segmentation by human annotators shows that these can be both smaller and larger fragments of text, depending on its length and genre.

Scientific and popular science texts (i.e., in this case, the corpus of scientific communication as a whole) are characterized by a high density of argumentation, which contributes to the persuasiveness and effectiveness of the impact on the reader. However, in the experiment for various subgenres and models, low values of recall for segmentation can be noted, i.e. identification of a lower density of argumentation compared to human annotation. This corresponds, first of all, to a smaller volume of predicted ADUs as shown in the list below and in Tab. 3.

1. Complex ADUs. Boundaries are set incorrectly in complex ADUs consisting of several sentences.
2. ADUs representing the source of information. Segmentation errors are observed in multicomponent structures denoting someone's speech or opinion: direct speech with inserted segment corresponding to the act of speaking *X explains*; reported speech or opinion *According to X*, etc. Separation of the source indicator is an absolute rule for annotators, due to the peculiarities of the analysis of argumentation from the source (expert, witness, etc.), while models either combine speech indicator with the main proposition, or do not isolate such a segment at all.
3. Subordinate clauses. Subordinate clauses connected to the main clause by means of a subordinate conjunction or a conjunction word *where*, *what*, *because*, *as a result of what*, *since*, etc. are not distinguished into independent ADUs – in this case, the correctness of the segmentation is determined by the semantics of the conjunction/conjunction word (usually a cause-and-effect relationship).
4. Nested phrases. Independent ADUs do not include embedded phrases that represent a comparison or exemplification of the situation in the main sentence.
5. Incomplete Propositions. Collapsed propositions represented by prepositional constructions with a substantive predicate (verbal noun) are not distinguished as independent ADUs.
6. Discontinuous structures. There is a lack of identification of discontinuous structures, the necessity of which is demonstrated by the example in Tab. 3 (it also presents the above-mentioned errors related to exemplification and collapsed proposition).

Table 3. Examples of segmentation errors.
Labels (a), (b), etc., denote the ADUs comprising the fragment

Error type	**Expert segmentation**	**Model segmentation**
Complex ADUs	*There are many methods for detecting a mask on the face, and most of them are a combination of other methods. But they can all be divided into two categories*	*But they can all be divided into two categories.*
ADUs representing the source of information	(a) *This is a compelling, evidence-based case for freshwater fishing at the end of the last ice age,* (b) *– Potter noted*	*This is a compelling, evidence-based case for freshwater fishing at the end of the last ice age*
Subordinate clauses	(a) *AR devices are the future of surgery,* (b) *because they can significantly reduce the number of medical errors,* (c) *and they can also be used to teach surgery*	*AR devices are the future of surgery, because they can significantly reduce the number of medical errors, and they can also be used to teach surgery*
Nested phrases	(a) *and they can also be used to teach surgery,* (b) *as the sports medicine specialist from Aglaya, Christopher Heading, did.*	*and they can also be used to teach surgery, as the sports medicine specialist from Aglaya, Christopher Heading, did.*
Incomplete Propositions	(a) *Mennonite dialects are generally recognized as German,* (b) *however, due to their constant migration to regions with other languages or German dialects,* (c) *a simple mention of this is not enough.*	(a) *Mennonite dialects are generally recognized as German,* (b) *however, due to their constant migration to regions with other languages or German dialects, a simple mention of this is not enough.*
Discontinuous structures	(a) *the ancestors of the indigenous people of this region, many of whom still depend on freshwater fish* (b) *(salmon, for example)* (a) *may have started subsistence fishing* (c) *in response to declining food resources during long-term climate change.*	*the ancestors of the indigenous people of this region, many of whom still depend on freshwater fish (salmon, for example), may have started subsistence fishing in response to declining food resources during long-term climate change.*

3.3.2. Errors of argument relations extraction

The analysis of the false positive and false negative responses of the models showed the following reasons for the incorrectly extracted relations.

Common causes. Expectedly, the facts explained by the causes of false results common to AM tasks were found: proximity in the text of clauses and sentences combined into one pair (one or neighboring sentences); lexical and semantic similarity (presence of the same words and synonyms in the statements of a pair); presence in a pair of relations similar to argumentative ones but not being them.

Technical reasons for false positive responses. The use of a generative approach led to the appearance of false positive results due to incorrect segmentation; this type of results also included pairs of statements that, in the expert annotation, are connected by an argumentative relation, but indirectly, through other statements.

Peculiarities of recognition of typical reasoning models (schemes of argumentation). The analysis of true- and false-negative responses revealed that the SFT approach has a greater coverage of schemes that are recognized with a quality greater than 10% than the LLM approach – 12 vs. 9. In general, the pairs realizing the Example and Cause to Effect relations are well recognized by the largest number of models on the largest number of subcases.

The relations Expert Opinion, Part to Whole and Sign are recognized in some cases most well (up to 28%), in other cases very poorly (up to complete absence of true positives).

Expert Opinion. This scheme is well recognized on the **News** subcorpus by SFT models; both LLM models for the same subcorpus and **mT5** models for the **Articles** subcorpus performed poorly. The low results are unexpected, since our previous experiments, which solved the problem of binary classification of pairs of statements, showed high results for this scheme (up to $F1_{urel}$=0.89). The analysis showed that the errors are related either to insufficient explication of the argument components or to insufficiently detailed segmentation performed by the models.

Part to Whole. This relation is well recognized by the **FRED-T5** model, but the other models perform poorly. The scheme contains two premises and a conclusion: *m is a species (part) of n; m has property G → n has property G.* The errors can be partially explained by the fact that experts use the ***Part to Whole*** scheme as a transitive rather than argumentative scheme when marking up the text, in order to preserve the integrity of the graph. Therefore, an isolated pair of sentences out of context may not contain argumentation, and the models naturally fail to detect it.

Sign. The ***Sign*** relation is well recognized by LLM models and poorly by SFT models, including **FRED-T5**. This reasoning model also has three components: *B is generally indicated as true when its sign, A, is true; A (a finding) is true in this situation → B is true in this situation.* False negative examples of this scheme, as a rule, had no explicit relation indicators, or the indicators are polysemous (may indicate not only argumentative relation: e.g., parenthetical explanations, markers *associated with*, *since*), or the segmentation of annotators is too fractional, and the selected segments are not informative enough for models. Another possible cause of errors is the simplification of arguments during data preparation, which made an isolated premise statement insufficient to support the conclusion.

Negative Consequences. In examples implementing this reasoning model, the models most frequently make errors in determining the type of relation (support or attack). This is because in the AIF ontology this scheme is supporting (*If A is implemented, bad consequences* *are likely to occur → A should not be implemented*), but due to the implicitness of some reasoning components, this scheme is more often implemented as attacking.

Table 4 demonstrates examples of the above-described errors made by the models in extracting argument relations.

Table 4. Examples of Argument Relation Extraction errors.
FP denotes False Positives, FN denotes False Negatives

Error type	**Errors in Argument Relation Prediction**	**Comments**
Incorrect segmentation (**FP**)	(a) *There are inaccuracies in the article,* |**attack**| (b) *5) There are inaccuracies in the article,*	Pairs of identical statements have been merged, either in full or truncated form.
Incorrect segmentation (**FP**)	(a) *5) The article* |**support**| (b) *The material of the article raises a number of questions:*	The statements correspond to pairs from the reference dataset, but one of the statements is truncated.
Mediated argument relation (**FP**)	<arg><aid>r6</aid><adu> *Since the main goal of our research was to compare the obtained results with the data presented in the «Russian Associative Dictionary»*</adu>|**support**|<adu>*In this study, the free association experiment was used as the main method, in which the respondent is required to give an unrestricted response to a stimulus word in the form of a response word or phrase.*</adu></arg>	In the expert annotation, the pairs of statements are mediated by others; in this example, the statement reports an intermediate research objective: *The focus of the research interest was on identifying, within the mini-group, matches with the most frequent responses recorded in the lexicographic source.*
Relation similar to argumentative ones but not being them (**FP**)	(a) *The author has collected interesting material (the corpus of examples found in each translation is given in the article), but limits himself exclusively to its quantitative analysis, presented in two tables.* |**support**| (b) *The reviewed article is devoted to the study of archaisms and historicisms in five Russian translations of The Song of Roland.*	The candidate premise is a detail of the candidate conclusion
Peculiarities of ***Part to Whole*** reasoning model (transitive scheme, **FN**)	(a) *The ATT&CK Matrix for Enterprise information security threat assessment methodology has the following merits* <...> |**support**| (b) *It helps to understand what tools attackers use, to familiarize with their techniques and tactics.* |**support**| (c) *This knowledge allows predicting the likely point of entry into organizations.*	Statement (b) is not a premise to statement (a) because it clarifies rather than proves it, but since statement (c) supports (b), the link between (a) and (b) is necessary to demonstrate the support for statement (a) by statement (c) in the graph.
Peculiarities of **Sign** reasoning model (**FN**)	(a) *No clarity in the interpretation of the term «language».* (b) *In the first sentence of the abstract we read: «one of the Germanic languages», which is known «in the world» as «Lower Germanic language (???) (dialect???) of the Mennonites».* (c) *Below it is labeled as «vernacular» for both Siberian and Canadian Mennonites.*	Thesis (a) is supported by two premises (b) and (c). Together (b) and (c) are able to prove the thesis, but separately they are insufficient to support (a).
Peculiarities of ***Negative Consequences*** reasoning model (Incorrect Relation type)	(a) *However, when dealing with large amounts of data, training diffusion models can be time-consuming and require large computational resources* |**attack**| (b) *Generalizing, diffusion models allow generating an image from a textual description by sequentially varying the noise in pixel space.*	If we explicate all the statements, the first fragment should be divided into two: (1) However, when dealing with large amounts of data, training diffusion models may not be appropriate (conflicts logically with the second statement) and (2) It can be time-consuming and computationally intensive (supporting premise to statement (1)). Thus, the model has generated a markup with a correct label, but it contradicts the «reference»markup.

4. Discussion

RQ1 discussion. In most cases, the generative approach to argument extraction shows rather low results on Russian-language texts, which is generally consistent with the work of other researchers. For example, in [4] on the English-language **MTC** corpus the $F1_{rel}$ score is 0.35, while on the Russian-language **ruMTC** corpus in our experiment $F1_{rel} = 0.37$.

On the positive side, a comprehensive analysis of the whole text through the chunking mechanism allowed us to include about 98% of all argumentative relations. Some problems arise at the boundaries of the chunks due to the arbitrariness of text partitioning, and aligning the chunk boundaries to the boundaries of sentences/paragraphs leads to an unstable context window size in the training sample, which has a bad effect on model training.

RQ2 discussion. The analysis of the dependence of the quality of argument extraction on the text length shows that such correlation takes place only for ultra-small texts, when it is possible to fully place the text in the context window (**ruMTC** vs. **ArgNetSC**). On texts larger than one chunk this ceases to play a significant role. For larger texts, genre features seem to play an important role, in particular, the complexity and type of argumentation used, the coverage of the text with argumentation, the size of the argumentative statement, etc. Thus, the experimental results show better performance when analyzing scientific articles than on other subgenres of the **ArgNetSC** corpus, despite their larger size (see Tab. 1). This can be explained by the fact that scientific articles have a stricter organization of the presentation of the material. In addition, there is less complex argumentation in this sub-corpus (in the **Articles**, 21% of binary relations are obtained as a result of simplification, in **Reviews** – 25%, and in **News** – 26%), so errors arising from the simplification of such argumentation are less frequent in this corpus.

Conclusion

The main goal of the study was to test new document-level generative approaches for solving the problem of argumentation analysis and long-range argumentative relation extraction.

Experiments conducted on Russian-language data showed that the highest results were achieved on microtexts, with small models fine-tuned with SFT showing approximately the same quality as large language models running on a specialized prompt. However, for large texts of scientific genres, this trend does not hold and the best results are obtained with the trained **FRED-T5** model.

To analyze long texts, a technique was proposed to segment them into chunks by a sliding window. The size of the chunk depended on the context window of the model used. This approach guarantees consideration of relations «fitting» into such a window.

The main types of errors that reduce argument extraction performance were identified: errors related to segmentation, to establishing an argument relation between two ADUs, and to determining the type of relation. The models often fail to recognize ADUs consisting of more than one sentence, presenting subordinate clauses, comparative turns, explanatory turns with examples, discontinuous structures, and indications of the source of information. The errors in establishing an argumentative relation between statements occurred both due to segmentation errors and general quality-reducing factors (lexico-semantic similarity, contact of statements between which a relation is falsely established, etc.), peculiarities of individual argumentation schemes. In addition, some of the erroneously identified relations in the expert markup are argumentatively related through other statements, but such relations could not be reflected in the dataset due to the peculiarities of the experiment. The most frequent errors in identifying the type of relation are explained by the implicitness of part of the reasoning elements, due to which supporting schemes are sometimes used as attacking schemes.

In further development of the approach, a stage for identifying the main thesis will be added, which is likely to adapt the document-level argument mining task for cases of long-range links over two paragraphs.

We have published our code at the GitHub[2]. All datasets used in this article are publicly available from each distributor.

Acknowledgements

This study was supported by the Russian Science Foundation (grant No. 23-11-00261), https://rscf.ru/project/23-11-00261/.

References

1. Accuosto, P., Saggion, H.: Mining arguments in scientific abstracts with discourse-level embeddings. Data & Knowledge Engineering 129, 101840 (2020). https://doi.org/10.1016/j.datak.2020.101840

2. Akhmadeeva, I.R., Kononenko, I., Sidorova, E., Shestakov, V.: Using rhetorical structures to analyze argumentation in scientific communication texts. Computational Linguistics and Intellectual Technologies (2025), https://api.semanticscholar.org/CorpusID:280935169

3. Bao, J., He, Y., Sun, Y., *et al.*: A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 10437–10449. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.713

4. Bao, J., Jing, M., Dong, K., *et al.*: UniASA: A Unified Generative Framework for Argument Structure Analysis. Computational Linguistics 51(3), 739–784 (09 2025). https://doi.org/10.1162/coli_a_00553

5. Cabrio, E., Villata, S.: Node: A benchmark of natural language arguments. In: Computational Models of Argument, pp. 449–450. IOS Press (2014). https://doi.org/10.3233/978-1-61499-436-7-449

6. Chen, Z., Chen, L., Chen, B., *et al.*: UniDU: Towards a unified generative dialogue understanding framework. In: Lemon, O., Hakkani-Tur, D., Li, J.J., *et al.* (eds.) Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 442–455. Association for Computational Linguistics, Edinburgh, UK (sep 2022). https://doi.org/10.18653/v1/2022.sigdial-1.43

7. Chistova, E.: End-to-end argument mining over varying rhetorical structures. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 3376–3391. Association for Computational Linguistics, Toronto, Canada (jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.209

[2] https://github.com/Inscriptor/doc-level-approach-arg-extraction.git

8. Dozat, T., Manning, C.D.: Deep biaffine attention for neural dependency parsing. In: International Conference on Learning Representations. Toulon, France (apr 2017), https://openreview.net/forum?id=Hk95PK9le

9. Du, X., Li, S., Ji, H.: Dynamic global memory for document-level argument extraction. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5264–5275. Association for Computational Linguistics, Dublin, Ireland (may 2022). https://doi.org/10.18653/v1/2022.acl-long.361

10. Fishcheva, I., Kotelnikov, E.: Cross-Lingual Argumentation Mining for Russian Texts. In: van der Aalst, W.M.P., Batagelj, V., Ignatov, D.I., *et al.* (eds.) Analysis of Images, Social Networks and Texts. pp. 134–144. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-37334-4_12

11. Galassi, A., Lippi, M., Torroni, P.: Multi-task attentive residual networks for argument mining. IEEE/ACM Transactions on Audio, Speech, and Language Processing 31, 1877–1892 (2023). https://doi.org/10.1109/TASLP.2023.3275040

12. Hu, X., Wan, X.: RST Discourse Parsing as Text-to-Text Generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 31, 3278–3289 (2023). https://doi.org/10.1109/TASLP.2023.3306710

13. Kawarada, M., Hirao, T., Uchida, W., Nagata, M.: Argument mining as a text-to-text generation task. In: Graham, Y., Purver, M. (eds.) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2002–2014. Association for Computational Linguistics, St. Julian's, Malta (mar 2024). https://doi.org/10.18653/v1/2024.eacl-long.121

14. Kotelnikov, E., Loukachevitch, N., Nikishina, I., Panchenko, A.: RuArg-2022: Argument Mining Evaluation. pp. 333–348 (06 2022). https://doi.org/10.28995/2075-7182-2022-21-333-348

15. Lauscher, A., Glavaš, G., Ponzetto, S.P.: An argument-annotated corpus of scientific publications. In: Slonim, N., Aharonov, R. (eds.) Proceedings of the 5th Workshop on Argument Mining. pp. 40–46. Association for Computational Linguistics, Brussels, Belgium (nov 2018). https://doi.org/10.18653/v1/W18-5206

16. Lawrence, J., Reed, C.: Argument mining: A survey. Computational Linguistics 45(4), 765–818 (01 2020). https://doi.org/10.1162/coli_a_00364

17. Li, S., Ji, H., Han, J.: Document-level event argument extraction by conditional generation. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., *et al.* (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 894–908. Association for Computational Linguistics, Online (jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.69

18. Li, Z., Lin, T.E., Wu, Y., *et al.*: UniSA: Unified Generative Framework for Sentiment Analysis. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 6132–6142. MM '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3581783.3612336

19. Liu, J., Chen, Y., Xu, J.: Machine reading comprehension as data augmentation: A case study on implicit event argument extraction. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 2716–2725. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.214

20. Lu, Y., Liu, Q., Dai, D., *et al.*: Unified structure generation for universal information extraction. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5755–5772. Association for Computational Linguistics, Dublin, Ireland (may 2022). https://doi.org/10.18653/v1/2022.acl-long.395

21. Mayer, T., Marro, S., Cabrio, E., Villata, S.: Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. Artificial Intelligence in Medicine 118, 102098 (2021). https://doi.org/10.1016/j.artmed.2021.102098

22. Niculae, V., Park, J., Cardie, C.: Argument mining with structured SVMs and RNNs. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 985–995. Association for Computational Linguistics, Vancouver, Canada (jul 2017). https://doi.org/10.18653/v1/P17-1091

23. Paolini, G., Athiwaratkun, B., Krone, J., *et al.*: Structured prediction as translation between augmented natural languages. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=US-TP-xnXI

24. Peldszus, A., Stede, M.: An annotated corpus of argumentative microtexts. In: Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015. vol. 2, pp. 801–815. College Publications, London (2016)

25. Sidorova, E., Akhmadeeva, I., Zagorulko, Y., *et al.*: An integrated approach to the analysis of argumentative relationships in scientific communication texts. Ontology of Designing 13(4), 562–579 (12 2023). https://doi.org/10.18287/2223-9537-2023-13-4-562-579, (in Russian)

26. Srivastava, P., Bhatnagar, P., Goel, A.: Argument Mining using BERT and Self-Attention based Embeddings. In: 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). pp. 1536–1540 (2022). https://doi.org/10.1109/ICAC3N56670.2022.10074559

27. Stab, C., Gurevych, I.: Annotating argument components and relations in persuasive essays. In: Tsujii, J., Hajic, J. (eds.) 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), https://aclanthology.org/C14-1142/

28. Stab, C., Gurevych, I.: Parsing argumentation structures in persuasive essays. Computational Linguistics 43(3), 619–659 (09 2017). https://doi.org/10.1162/COLI_a_00295

29. Strubell, E., Verga, P., Andor, D., *et al.*: Linguistically-informed self-attention for semantic role labeling. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 5027–5038. Association for Computational Linguistics, Brussels, Belgium (oct – nov 2018). https://doi.org/10.18653/v1/D18-1548

30. Timofeeva, M., Ilina, D., Kononenko, I.: Argumentative annotation of the scientific Internet-communication corpus: Genre analysis and study of typical reasoning models based on the ArgNetBank Studio platform. NSU Vestnik 22(1), 27–49 (2024). https://doi.org/10.25205/1818-7935-2024-22-1-27-49, (in Russian)

31. Toulmin, S.E.: The Uses of Argument. Cambridge University Press, 2 edn. (2003). https://doi.org/10.1017/CBO9780511840005

32. Walker, M., Tree, J.F., Anand, P., *et al.*: A corpus for research on deliberation and debate. In: Calzolari, N., Choukri, K., Declerck, T., *et al.* (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 812–817. European Language Resources Association (ELRA), Istanbul, Turkey (may 2012), https://aclanthology.org/L12-1643/

33. Xu, H., Ashley, K.: Multi-granularity argument mining in legal texts. In: Legal Knowledge and Information Systems, pp. 261–266. Frontiers in Artificial Intelligence and Applications, IOS Press (2022). https://doi.org/10.3233/FAIA220477

34. Yan, H., Gui, T., Dai, J., *et al.*: A unified generative framework for various NER subtasks. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5808–5822. Association for Computational Linguistics, Online (aug 2021). https://doi.org/10.18653/v1/2021.acl-long.451

35. Ye, Y., Teufel, S.: End-to-end argument mining as biaffine dependency parsing. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 669–678. Association for Computational Linguistics, Online (apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.55