Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2025”

April 23–25, 2025

RuTermEval-2024: Cross-domain Automatic Term Extraction and Classification in Russian scientific texts

Angelina Mamontova
MSU Institute for Artificial Intelligence
mamontova@mlsa-iai.ru

Roman Ischenko
MSU Institute for Artificial Intelligence
ischenko@mlsa-iai.ru

Konstantin Vorontsov
MSU Institute for Artificial Intelligence
vorontsov@mlsa-iai.ru

Abstract

Automatic Term Extraction (ATE) is a critical NLP task for identifying domain-specific terms, which are essential for tasks like information retrieval, machine translation, and ontology construction. Cross-domain nested term extraction further complicates the task, as traditional methods often fail to handle hierarchical term structures and domain variability. This paper introduces both the CL-RuTerm3 dataset, a novel resource featuring nested term annotations across six domains (the main one is computational linguistics, also mathematics, medicine, economics, literature studies, and agrochemistry), and the RuTermEval-2024 competition, designed to evaluate term extraction systems on this data. The CL-RuTerm3 dataset, comprising 1270 abstracts and 15 full-text articles (over 165k tokens with over 37k annotated entities), is the largest of its kind for Russian scientific texts. Terms are classified into three categories based on lexical and domain specificity: specific terms, common terms, and nomens. The dataset’s unique features, such as nested term markup and cross-domain coverage, enable more realistic evaluation of ATE systems. The paper concludes with an analysis of participant approaches in the RuTermEval-2024 competition, emphasizing the effectiveness of contrastive learning. This work aims to advance ATE research by providing a robust dataset and fostering discussions on term extraction methodologies.

Keywords: ATE; automatic term extraction; nested term; dataset
DOI: 10.28995/2075-7182-2025-23-245-256

RuTermEval-2024: Кросс-доменное автоматическое извлечение терминов и их классификация в русскоязычных научных текстах

Ангелина Мамонтова
Институт ИИ МГУ
mamontova@mlsa-iai.ru

Роман Ищенко
Институт ИИ МГУ
ischenko@mlsa-iai.ru

Константин Воронцов
Институт ИИ МГУ
vorontsov@mlsa-iai.ru

Аннотация

Автоматическое извлечение терминов (ATE) – одна из важнейших задач NLP, позволяющая выявлять специфические для домена термины, которые используются в задачах поиска информации, машинного перевода и построения онтологий. Кросс-доменное извлечение вложенных терминов еще больше усложняет задачу, поскольку традиционные методы часто не справляются с иерархическими структурами терминов и новыми доменами. В данной статье представлены новый набор данных CL-RuTerm3, содержащий разметку

фрагментов с терминами и включающий тексты шести областей (главной является компьютерная лингвистика, также присутствуют области математики, медицины, экономики, литературоведения и агрохимии), и конкурс RuTermEval-2024, созданный для оценки систем извлечения терминов с использованием этих данных. CL-RuTerm3 включает 1270 аннотаций и 15 полнотекстовых статей (более 165 тысяч токенов с более чем 37 тысячами размеченных терминов) и является крупнейшим в своем роде для русских научных текстов. Термины разделены на три категории в зависимости от своей лексической и доменной специфики: специфические термины, общие термины и номены. Уникальные особенности набора данных, такие как разметка вложенных терминов и включение текстов различных научных областей, позволяют более реалистично оценивать системы ATE. В заключение статьи приводится анализ подходов участников конкурса RuTermEval-2024, подчеркивается эффективность методов контрастивного обучения. Данная работа направлена на развитие исследований в области ATE путем предоставления надежного набора данных и продолжения развития методологии извлечения терминов.

Ключевые слова: автоматическое извлечение терминов; термин; вложенный термин; набор данных

1 Introduction

Automatic Term Extraction (ATE) is an NLP task used to automatically identify and extract domain-specific terms from a collection of texts. These terms typically represent key concepts within a specialized field, such as medicine, engineering or linguistics. As units of knowledge in a specific field of expertise, extracted terms are not only beneficial for terminographical tasks, but also support and improve several complex downstream tasks, e.g., information retrieval, machine translation, topic detection, and topic modeling, etc.

Despite the significant research interest that automatic term extraction has received, it remains a very challenging task. Terms are generally defined as "textual expressions that denote concepts in a specific field of expertise" (Tran et al., 2023); however, such definitions leave room for many questions about the fundamental nature of terms. Some of the most fundamental differences in terms' basic characteristics are term POS-pattern (only nouns and noun phrases or including other POSes), minimum term frequency and length (in number of tokens). More difficult to quantify are differences such as how specialized or domain-specific a lexical unit needs to be before it is considered a term. The lack of agreement among researchers on even basic characteristics of task is a significant hurdle for all aspects of ATE, from data collection to benchmarking and comparative research in general.

We defined a *term* as follows: it is a word or collocation (multiple syntactically connected words) naming a concept, object, feature or action characteristic of a certain scientific area (domain). The main property of a term is its domain specificity, i.e. belonging to a certain domain. Differences in the so-called lexical specificity, i.e. usage only among a limited group of experts, are expressed by assigning of the labeled terms to different classes.

The task of identifying "flat" terms in the document $D = \{w_0, w_1, \dots, w_{n-1}\}$ is a recognizing entities $t = \{w_i, w_{i+1}, \dots, w_{i+m-1}\},\ 1 \le m < n$, moreover if a word is included in one term, it cannot be included in any other term (formula 2.1). By *nested terms* annotation we mean both classical nested entities setups – a term can be a substring of any other term, or terms can only be intersecting (have a common part). In this setup word can be part of several terms, and also an entire term can be included in another one (formula 2.2). Maximum depth (number of terms in one text fragment) is not limited, nor are the classes of nested terms – they can be either the same or different.

$$t \subseteq D \mid t = \{w_i, w_{i+1}, \dots, w_{i+m-1}\},\ D = \{w_0, w_1, \dots, w_{n-1}\}, 1 \le m < n,\ a \ne b \implies t_a \ne t_b \quad (1)$$

$$\forall\, w, a, b\colon\ w \in t_a \mid a \ne b \implies w \notin t_b \quad (2.1)$$

$$\exists\, w \mid w \in (t_a \cap t_b), \qquad a \ne b \quad (2.2)$$

Cross-domain nested term extraction presents unique difficulties. The syntactic and semantic properties of terms can vary significantly between domains (examples of terms in each class and domain can be seen in Table 2). Existing term extraction approaches often struggle to maintain robust performance when applied to diverse datasets due to their reliance on domain-specific heuristics, frequency-based statistical measures, or supervised models trained on limited annotated corpora. Furthermore, the hierarchical nature of nested terms complicates extraction, as traditional ATE methods primarily focus on flat term structures, failing to account for the compositional relationships between short and long term candidates.

domain	specific term	common term	nomen
computational linguistics	генитивная ИГ pluralia tantum «ё»-омограф разрешение омонимии претренированная ruGPT3 модель конструкция с внешним посессором	слог словарь web-браузер татарский язык ключевое слово	C++ тезаурус RussNet Национальный корпус русского языка Идеографический словарь О.С.Баранова
agrochemistry	α-амилаза 9,10-дифенилантрацен N-удобрение H+/OH−-равновесие диоксид кремния 4-польные севообороты с короткой ротацией	почва растение сорт удобрение сельское хозяйство	Реаком-Хелат Бора Канская лесостепь штамм PGPB Pseudomonas plecoglossicida 2,4-D
literature studies	имажинист 6-стопный ямб дискурс телесности образно-мотивный комплекс	перевод глава публикация советский писатель	«Евгений Онегин» А.С. Пушкин Всероссийский союз поэтов газета «Северная пчела»
medicine	HLA-ген фактор фон Виллебранда травма периферической нервной системы her2-позитивный рак молочной железы	лечение мозг болезнь лекарственное средство	SARS-CoV-2 Thymus Serpyllum L. Федеральный регистр доноров костного мозга
economics	ВВП виртуальная валюта рынок труда модель глобального доминирования публично котируемые нефинансовые корпорации	компания ресурсы организация	БРИКС Великая депрессия Новый банк развития
mathematics	базис s-сплайн (0,∞]-значные функции метод наименьших квадратов	формула точка произведение	Галуа монография Л. Фукса «Бесконечные абелевы группы»

Table 1: Examples of annotated terms in each class and domain

Creating domain-specific corpora with annotated terms is time- and effort consuming. When manual term annotation is involved, inter-annotator agreement is notoriously low and there is no consensus about an annotation protocol. This leads to a lack of available resources. Moreover, it means that the few available datasets are difficult to combine and compare, and often cover only a single language and domain. Also, very few ATE datasets include Russian texts, so the creation of such a dataset is relevant. Additional features of our dataset are nested term annotation and including several domains, which allows to considering the term extraction task in a more real-world setting. Of course, we do not claim

to be fully correct about the chosen methodology and the completeness of the dataset, given that it is based on texts of only one language. In order to compare ATE models and discuss the relevant issues using the created dataset, the RuTermEval-2024 competition was prepared within the framework of the Dialogue conference.

2 Related Work

Problems in creating data for an ATE task usually begin with ambiguity in understanding the search object, which strongly depends on the goals of ATE step, i.e., what the identified units are to be used for next, and actual belonging to terms from a linguistic point of view takes a back seat. It is worth noting that the problem of different understanding of the ATE search object also comes from the ambiguous linguistic nature of the term, its "multifacetedness". There are only few large annotated resources available for the task and they are usually monolingual and cover only a single domain. Since term characteristics, and therefore also ATE performance, can vary greatly between languages and domains, this is a serious drawback.

The most widely used datasets are GENIA (Kim et al., 2003) with 2,000 abstracts from MEDLINE database and ACL-RD-TEC 2.0 (Qasemizadeh & Schumann, 2016) with 300 abstracts from the ACL Anthology Reference Corpus, both in English. Although all annotated datasets exist: CRAFT (Cohen et al., 2017), TTC (Gornostay et al., 2012), KAS-term and KAS-biterm (Ljubešić et al., 2018), etc., the general lack of large, multilingual, and multi-domain resources remains a critical limitation. Given that term characteristics – and by extension, ATE performance – vary significantly across languages and domains, this constraint is a considerable drawback.

In the vast majority of datasets, the ATE task is set just in recognition setup. More interesting is the approach implemented in the ACTER dataset. It is a specialized corpora in three languages and four domains, which markup includes four classes. The main division of classes was made by the level of lexical specificity (its comprehensibility only to a domain expert or any person) and domain specificity (belonging to a certain domain or being used in any research): so there appeared special/specific terms – domain-specific and lexically specific, common terms – domain-specific but not lexically specific (comprehensible even to a non-specialist) and out-of-domain terms – specific only lexically, but used in any domain. The Named Entities class was additionally introduced for unique names of objects of any domain (Terryn et al., 2020). Nevertheless, the TermEval shared task (Terryn et al., 2020) did not use multiclass annotations, it was conducted just in an extraction setup. Despite the novelty of the ACTER partitioning approach, it has not received due attention, although, in our opinion, it is the classification of the units marked up in scientific texts depending on their lexical and domain specificity that can help to prepare a universal dataset for the ATE task for its various applications.

To our knowledge, there are no open datasets for multi-domain nested ATE task in Russian scientific articles. There are several annotated corpora for other task and texts, for example, RuDReC corpus (Tutubalina et al. 2021) was made for NER in consumer reviews about pharmaceutical products (RuDReC also includes not-labeled part of health-related user-generated texts from various Internet sources), not scientific articles or its abstracts. The known NEREL-BIO dataset (Loukachevitch et al., 2023) is the largest annotated dataset of scientific articles in Russian (766 abstracts), but it needs a strong adjustment to be used for ATE purposes, since it was created to solve the NER problem and contains 41 semantic classes of the searched units, some of which are not domain-specific. Also, NEREL-BIO is monodomain, and scaling its markup methodology to other domains is hardly possible due to differences in the composition of semantic classes.

It should be noted that due to the complexity of creating ATE annotated datasets, most of them consist only of abstracts, but not the full texts of scientific articles, and the ATE models using existing data usually work within this limitation. We believe this approach is appropriate, but to understand models' capability to solve the problem on full articles, a small number of such texts were included in the development and test sets of first and second tracks.

The RuTermEval shared task aims to provide a valuable new resource while simultaneously advancing understanding of the state-of-the-art in ATE, identifying existing strengths and weaknesses, and inspiring novel approaches in the field. The CL-RuTerm3 dataset includes six domains and enables participants to train and evaluate their systems using diverse and detailed data. Despite using the

TermEval approach, our shared task is the first competition dedicated to the task of classifying terms by their specificity, since TermEval was conducted as an extraction setup (without classification).

3 Dataset

The CL-RuTerm3 dataset of abstracts and full texts of scientific articles was prepared specially for the RuTermEval–2024 shared task. The basis of our dataset were the proceedings of the conference "Dialogue" for 2000–2023 in Russian (1055 abstracts and 15 full-texts articles of computational linguistics domain). To test the scalability of the models to other domains, about 220 abstracts from five additional scientific areas (mathematics, medicine, agrochemistry, literature studies and economics) were also included in dataset. More detailed quantitative description of our dataset is presented in Table 1.

The uniqueness of the material is that linguistics has hardly been considered before in the ATE task (apart from ACL RD-TEC datasets) and is a new material for cross-domain experiments and for analyzing the differences in NLP processing of term systems of different groups of sciences. An additional feature of CL-RuTerm3 is markup of nested terms.

3.1 Term annotation

The markup was conducted by three annotators – specialists in linguistics, information technology and mathematics – single assessor for each document with collective discussion of challenging cases. Quality control was performed by moderator with experience in the field of terminology, ATE and dataset creation. The moderator and assessors created an assessor's guideline with a detailed description of the annotation task and a breakdown of correct and incorrect annotations. This helped assessors to make consistent decisions and make the whole process more transparent. Nevertheless, term annotation remains an ambiguous and subjective task, and we do not claim that ours is the only possible interpretation.

For the task of classifying terms, we proposed to divide them according to the degree of lexical specificity of the term (its familiarity among ordinary people who are not domain experts) and denotatum uniqueness:

1. specific term – terms that are both domain-specific and lexically-specific;
2. common term – domain-specific terms (known and used by non-specialists);
3. nomen – unique names of objects belonging to a particular domain, including nomenclature names (datasets, programming languages, corpora and dictionaries, scientific schools, writers and scientists, etc.).

Classification of terminological units by specificity (lexical and domain) was first applied in the ACTER dataset, in which classification was done in 4 classes (Specific Terms, Common Terms, Out-of-Domain Terms and Named Entities). In contrast to the ACTER, we dropped the Out-of-domain terms class. Units of this class that denote mathematical concepts (e.g., *p-value*, *confidence interval*, etc.), but belong to out-of-domain terms because of their use in any domain were marked by us as specific terms, and general scientific vocabulary that has neither lexical nor domain specificity (*method*, *research*, *experiment*, etc.) was omitted.

Markup was conducted in the format of a sequence labeling task – fragments with a term were identified and classified into three classes (a markup example is shown in Figure 1). In the output, the markup of each text is represented as a list of triples [start_index, end_index, term_class] for each labeled term.

Figure 1: Markup sample from the CL-RuTerm3 dataset

domain	type of texts	text count	token count	annotation count	unique term count	lemmatized term count
computational linguistics	abstracts	1,053	97,296	22,673	10,684	7,812
	full texts	15	32,254	5,190	2,303	1,539
agrochemistry	abstracts	55	10,321	3,110	1,591	1,267
literature studies	abstracts	60	13,211	2,478	1,578	1,218
medicine	abstracts	40	7,825	1,877	1,138	907
economics	abstracts	30	6,559	1,065	747	609
mathematics	abstracts	32	2,924	746	482	312
Σ	full texts	15	32,254	5,190	2,303	1,539
	abstracts	1,270	138,136	31,949	16,002	11,903
	all	1,285	170,390	37,139	17,652	12,938

Table 2: Quantitative characteristics of CL-RuTerm3 dataset

Consider some aspects of markup using the text in Figure 1 as an example. In phrase "National corpus of the Russian language" the following terms are identified: “National corpus of the Russian language” (as unique name of domain-specific product), “corpus”, “Russian language” and “language”.

Words “national” and “Russian” are not annotated because we do not labeled adjectives without the substantive defined by them as independent term. An exception to the rule is that if such a word is not near each other (forming a discontinuous term), the adjective is annotated as a single-word independent term, as can be seen in the example “morphological and taxonomic annotation of texts”, where “morphological” is a term, but “taxonomic” is not.

The phrases “national corpus” and “corpus of the Russian language” are not annotated because they are not separate scientific units that are often reproduced in the domain or have paradigmatic relations with other terms (in contract to “parallel corpus” or “dialect corpus”). If a term is supplemented by a feature, but the meaning of this phrase does not differ in any way from the sum of the meanings of its parts, phrase’ reproducibility in the domain and/or the existence of systemic relations with other terms become significant.

The full description of markup rules requires a separate scientific coverage due to its volume because of the complexity of the markup task and its linguistic multifacetedness. In this article we will mention only the main features of our markup rules. In general, each labeled term should remain a term even out of context, being part of the lexical system of a particular domain.

To the assessor, an expression or word is most likely to be a term if:

- It is a regular name (reproduced as a result of a single act of speech production and predominantly in an observable speech form).

- It has a definition.
- It can be found in dictionaries or ontologies of the relevant domain.
- It is related to the valid term (e.g., they are hyponyms of the same hyperonym or are in a generic-species relationship).
- It is an abbreviation used for the valid term.
- It has an abbreviation used in the particular domain.
- It is an object of research in the relevant domain.

To improve the quality of the final dataset markup, inaccuracies in the identification of term boundaries were corrected. Additional discussions were also carried out on the markup of terms that had a significant number of labels of different classes, as it is preferable for a term to belong to a particular class.

Final dataset consists of 1,285 text in Russian (over 165k tokens) with ≈37k sequence annotations, provided by domain and terminology human experts, more detailed quantitative description of our dataset is presented in tables 2 and 3.

domain	type of texts	annotation count	class	annotation count	lemmatized term count	mean lemmatized term frequency
computational linguistics	abst.	22,673	specific	15,946	6,748	2.385
			common	5,906	593	9.708
			nomen	821	471	1.741
	full texts	5,190	specific	3,826	1,236	3.224
			common	1,005	81	10.444
			nomen	359	222	1.617
agrochemistry	abst	3,110	specific	2,343	1,121	2.090
			common	655	68	9.632
			nomen	112	78	1.436
literature studies	abst	2,478	specific	927	485	1.903
			common	490	59	8.373
			nomen	1,061	674	1.574
medicine	abst	1,877	specific	1,603	851	1.895
			common	234	27	8.296
			nomen	40	29	1.379
economics	abst	1,065	specific	945	560	1.686
			common	54	16	3.438
			nomen	66	33	2.000
mathematics	abst	746	specific	617	270	2.226
			common	101	26	4.500
			nomen	28	16	1.750

Table 3: Distribution of terms in each class and domain

4 Setup

4.1 Tasks

The RuTermEval-2024 Shared task features three sub-tasks:

1. Nested term extraction (in sequence labeling format);
2. Nested term extraction (in sequence labeling format) and classification (labels are specific, common, nomen);
3. Nested term extraction (in sequence labeling format) and classification (labels are specific, common, nomen) in cross-domain task.

All tracks assume a transfer learning task, so the test set in each of them includes such texts, the likes of which were absent in the train set. Thus, in tracks one and two, the test set includes texts of a different genre – in addition to abstracts as in the train set, it includes full articles. In the third track, the test set consists of texts of only domains that were not present in the training set.

4.2 Evaluation

For all tracks, only full term matches were considered.

The metric for the first task – term extraction without class consideration – is the averaged F1-score across all documents, with abstracts and full texts treated separately:

$$F1_{T1} = \frac{1}{2}\left(\frac{1}{n_{abst}}\sum_{i=1}^{n_{abst}} F1_{abst_i} + \frac{1}{n_{ft}}\sum_{i=1}^{n_{ft}} F1_{ft_i}\right)$$

where n_{abst}, n_{ft} represent the number of abstracts and full-text documents, respectively, and $F1_{ann_i}$, $F1_{ft_i}$ denote the F1-score for each individual text.

To evaluate the quality of solutions for the second task, a weighted F1-score across various classes was employed:

$$F1_{T2} = \frac{1}{6}\left(3 * F1_{spec} + 2 * F1_{comm} + F1_{nom}\right)$$

where $F1_{spec}$, $F1_{comm}$, $F1_{nom}$ represent the F1-scores for specific, common terms, and nomens, respectively. The weights for each class were set according to their importance to the overall task.

Each individual $F1_t$ is calculated similarly to the metric from the first task:

$$F1_t = \frac{1}{2}\left(\frac{1}{n_{abst}}\sum_{i=1}^{n_{abst}} F1_{abst_i} + \frac{1}{n_{ft}}\sum_{i=1}^{n_{ft}} F1_{ft_i}\right)$$

where $t \in \{spec, comm, nom\}$.

For the third task, a weighted F1-score analogous to that used in the second task was applied:

$$F1_{T3} = \frac{1}{6}\left(3 * F1_{spec} + 2 * F1_{comm} + F1_{nom}\right)$$

However, since full texts were not available for the third task, each individual $F1_t$ was calculated as the averaged F1-score across all abstracts:

$$F1_t = \frac{1}{n_{abst}}\sum_{i=1}^{n_{abst}} F1_{abst_i}$$

where $t \in \{spec, comm, nom\}$.

For the second and third tracks, an additional F1-score without class consideration (analogous to the first track) was computed. This allowed for a separate determination of the quality of term extraction without classification.

4.3 Baseline

To compare participants' models with the simplest solution baseline was prepared. All the entities labeled as terms were selected with their annotations count, as well as their total occurrence in the corpus (in any grammatical form). Further, the optimal ratio of the number of markings as a term to its frequency in the corpus was empirically selected – 0.35. With the obtained bag of terms (with retained classes), a test set was marked up to obtain the required kind of markup.

4.4 Dataset Splits

The training data is the same for all subtasks, but in first track the term classes were removed. The training data consists of 850 texts (77k tokens, 18k annotations) in computational linguistics domain.

Development and test set for first and second subtasks is also the same, they belong to the train domain. Development data includes 103 abstracts and 10 full-text articles (25k tokens, 5k annotations), test set consists of 100 abstracts and 5 full-text articles (27k tokens, 5k annotations).

The development and test set for the third track consists of abstracts from domains, which were absent in train data. Development data contains 115 texts (24k tokens, 6k annotations). The test set consists of 102 texts (17k tokens, 4k annotations), but more diverse in domain component.

5 Results and Discussion

Seven teams participated in the RuTermEval–2024. Six teams took part in the final testing phase (distribution across tracks: 5, 5, 3). We provide descriptions of the solutions submitted by the top teams. The names are listed as the participants registered them in CodaLab. The final results are presented in separate tables for each track (in the case of multiple submissions, only the best one is considered).

We summarize participating teams' methods. A detailed description of their work can be found in their own articles.

fulstock [LAIR RCC MSU] participant who won all tracks treated nested term extraction as a NER task using the Binder model (Zhang et al., 2023), which employs contrastive learning to extract nested terms. Text sequences and class descriptions are encoded using RuRoberta-large, with embeddings mapped to a shared vector space. The model aligns entities of the same type closer to a class-specific center (defined by entity descriptions) while distancing unrelated subsequences, optimizing for term extraction accuracy.

Rank	Participant and team	Scores					
		with classification			without classes		
		F1-class_weighted (avg. per docs)	Precision (avg. per docs)	Recall (avg. per docs)	F1 (avg. per docs)	Precision (avg. per docs)	Recall (avg. per docs)
Track 1							
1	**fulstock [LAIR RCC MSU]**	–	–	–	**79.40%**	**79.69%**	**80.12%**
2	**VladSemak [VSemak]**	–	–	–	76.85%	79.40%	75.72%
–	***baseline***	–	–	–	61.69%	78.75%	52.68%
3	**ivan_da_marya**	–	–	–	56.19%	68.03%	50.36%
4	ragunna [KiPL SPBU]	–	–	–	53.49%	58.64%	51.83%
5	angyling	–	–	–	53.33%	42.29%	76.23%
Track 2							
1	**fulstock [LAIR RCC MSU]**	**69.97%**	70.45%	**70.53%**	**77.79%**	78.77%	**78.00%**
2	**VladSemak [VSemak]**	69.96%	**72.18%**	69.20%	77.26%	**79.85%**	76.38%
–	***baseline***	58.26%	67.09%	57.03%	61.69%	78.75%	52.68%
3	**VatolinAlexey [ai]**	57.97%	63.18%	58.79%	63.47%	68.97%	60.75%
4	ragunna [KiPL SPBU]	50.43%	55.98%	51.31%	52.04%	57.07%	50.44%
5	angyling	31.37%	27.83%	42.32%	53.33%	42.29%	76.23%
Track 3							
1	**fulstock [LAIR RCC MSU]**	**48.23%**	53.23%	48.85%	**60.38%**	67.90%	56.77%
2	**VladSemak [VSemak]**	46.54%	**57.85%**	46.22%	50.88%	**71.26%**	42.75%
3	**angyling**	43.70%	45.16%	**51.18%**	52.81%	49.36%	**60.74%**
–	*baseline*	22.76%	40.14%	30.93%	11.81%	34.87%	7.59%

Table 4: All participants' scores

VladSemak used the *span classification* approach (Binder) and contrastive learning with two BERT encoders to map text spans and term descriptions into a shared vector space. It evaluates all spans up to a set length, identifying terms via vector similarity. The method maximizes similarity for term spans with their type descriptions and minimizes it for non-terms.

ivan_da_marya team used a span classification approach, generating lemmatized n-grams (length 1 to 4) and encoding them with SBERT. Span and sentence vectors, along with POS tags, were concatenated and fed into a KNN classifier to identify terms.

angyling tested one-shot prompting on the Qwen2.5 (An el at., 2024). The prompt included an explanation of the task with one example from the training set, and an explanation of the correct form of answer inference. Next, all terms extracted by the model were used to final annotation in nested setup. All entities were labeled as specific class.

The majority of submissions of first track demonstrated higher precision than recall. However, in the winning solution, recall played a decisive role in the final evaluation. Notably, the baseline is competitive with top-performing solutions, and approaches provided by other participants failed to this simple model. A similar pattern was observed in the second track. Solutions outperforming the baseline achieved significantly higher recall. Additionally, both leading solutions performed best on specific terms, while their performance on nomens was noticeably weaker. Interestingly, the baseline handled common terms exceptionally well, but underperformed on specific terms because of low recall. As expected, baseline showed the highest precision in identifying specific and nomen terms, and score on common terms was also quite high. All participants' scores per class are summarized in the Table 5.

Participant	**Specific**			**Common**			**Nomen**		
	Micro-average F1	Precision (class-wide)	Recall (class-wide)	Micro-average F1	Precision (class-wide)	Recall (class-wide)	Micro-average F1	Precision (class-wide)	Recall (class-wide)
Track 2									
fulstock	**74.81%**	77.66%	**72.16%**	69.70%	58.80%	**85.57%**	35.15%	49.33%	**27.31%**
VladSemak	74.36%	78.04%	71.01%	**73.19%**	**65.63%**	82.70%	**35.79%**	62.39%	25.09%
baseline	56.55%	**78.10%**	44.32%	70.48%	61.47%	82.57%	32.39%	**70.37%**	21.03%
VatolinAlexey	39.45%	69.52%	27.54%	43.79%	56.47%	35.76%	25.22%	61.43%	15.87%
ragunna	36.44%	52.92%	27.78%	48.13%	62.14%	39.27%	15.85%	45.61%	9.59%
angyling	43.73%	33.26%	63.80%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Track 3									
fulstock	**61.92%**	67.63%	57.09%	17.20%	18.67%	15.94%	**18.08%**	37.21%	**11.94%**
VladSemak	54.60%	**71.23%**	44.27%	18.57%	**37.50%**	12.34%	13.07%	**52.63%**	7.46%
angyling	48.39%	40.88%	**59.30%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
baseline	9.08%	42.04%	5.09%	**26.44%**	32.34%	**22.37%**	0.00%	0.00%	0.00%

Table 5: Scores of participants' models in each term class

Note that only models based on contrastive learning approach outperformed the baseline in the first and second tracks, which shows that within the domain particular work, applying to dictionary sources (a specific list of terms) can allow to quickly achieve acceptable results and even be more effective than other methods. Improving the quality requires the use of already more advanced and sophisticated methods, e.g., contrastive learning, and demands further research and experimentation.

It is worth to mention that the approach using generative LLM with just one sample (by angyling) showed very high recall comparable to the leaders, who used all labeled data of the same domain as the test set. In the more equal conditions of the third track, when no one had data from the test domain (but did have data from another domain), generative LLM approach achieved a comparable result to the leaders primarily because of its the best recall, in contrast to the top two participants, whose precision, substantially exceeding their recall values, contributed most to the final scores. Also, this approach performed best scores on the economics domain, which proved to be the most challenging for the other models (see Table 6 for a look at the quality for each domain), although best results on the math and medicine are achieved by the fulstock with using Binder model (contrastive learning).

Rank	Participant and team	Domain	F1-class_weighted (avg. per docs)	Macro-average F1	Precision (avg. per docs)	Recall (avg. per docs)
1	**fulstock [LAIR RCC MSU]**	Mathematics	**51.88%**	**53.02%**	**52.30%**	**53.75%**
		Medicine	**52.86%**	**53.94%**	57.91%	**50.48%**
		Economics	38.18%	40.24%	46.68%	35.36%
2	**VladSemak [VSemak]**	Mathematics	48.93%	49.75%	50.33%	49.18%
		Medicine	49.51%	52.63%	**66.48%**	43.55%
		Economics	40.01%	43.43%	**56.01%**	35.47%
3	**angyling**	Mathematics	47.35%	48.01%	45.65%	50.64%
		Medicine	38.76%	39.93%	35.88%	45.01%
		Economics	**46.38%**	**47.83%**	42.11%	**55.34%**
4	*baseline*	Mathematics	26.20%	30.06%	43.36%	23.00%
		Medicine	26.19%	29.44%	38.03%	24.02%
		Economics	14.52%	16.04%	18.53%	14.14%

Table 6: Scores of participants' models in each domain (Track 3)

6 Conclusion

We presented the RuTermEval-2024 shared task, the first shared task on cross-domain nested term extraction and classification for the Russian language. The CL-RuTerm3 dataset is based on the proceedings of the "Dialogue" conference from 2000 to 2023 (1055 abstracts and 15 full-text articles in the computational linguistics domain). To evaluate the models' generalizability to other domains, approximately 220 abstracts from five additional scientific areas (mathematics, medicine, agrochemistry, literature studies, and economics) were also included. To our knowledge, this dataset exceeds the volume of all existing open datasets for solving a similar problem for Russian scientific texts.

The competition attracted 13 submissions from seven teams, addressing the problem through three subtasks. The best results were achieved by the solution that utilized the Binder model with contrastive learning. All solutions performed well in extracting terms, but they were slightly less effective in classifying the extracted terms. Within a single domain, only models based on contrastive learning outperformed the baseline approach, which relied solely on labeling terms found in the train and dev sets. This suggests that leveraging dictionary sources (or other resources of domain-specific terms) can enable rapid achievement of acceptable results. Further improvements will require the application of more advanced methods, such as contrastive learning and other emerging techniques.

As expected, performance declined when identifying terms from domains absent in the train and dev sets (Track 3). However, a generative LLM approach (with Qwen2.5) achieved strong results, primarily due to high recall despite minimal annotation samples (just one example from the training set). This result highlights the potential of generative LLMs for term extraction in low-resource scenarios and underscores the need for further research and experiments.

The shared task dataset, codebase and other materials are available online on GitHub[*]. We welcome the communities of NLP developers, linguists, and engineers to contribute to further research in the area.

Acknowledgements

The authors are grateful to Julian Serdyuk from Lomonosov Moscow State University for support in researching the discussed theme and Alsu Vagapova from Kazan Federal University for help in annotating the data.

[*] https://github.com/mlsa-iai-msu-lab.

References

[1] An Yang, Baosong Yang et al. Qwen2.5 Technical Report // Computing Research Repository. – 2024. – Vol. arXiv:2412.15115. Access mode: https://arxiv.org/abs/2412.15115.

[2] Cohen K. Bretonnel, Verspoor Karin et al. The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain // Handbook of Linguistic Annotation. – Dordrecht: Springer, 2017. – P. 1379–1394. – DOI: http://dx.doi.org/10.1007/978-94-024-0881-2_53.

[3] Gornostay Tatiana, Gojun Anita et al. Terminology Extraction, Translation Tools and Comparable Corpora: TTC concept, midterm progress and achieved results // LREC 2012 Workshop on Creating Cross-language Resources for Disconnected Languages and Styles (CREDISLAS), Istanbul, Turkey. – 2012. – P. 1–4.

[4] Kim Jin-Dong, Ohta Tomoko, Tateisi Yuka, Tsujii Junichi. GENIA corpus – a semantically annotated corpus for bio-textmining // Bioinformatics. – 2003. – Vol. 19. – Suppl. 1 – P. i180–i182. – DOI: https://doi.org/10.1093/bioinformatics/btg1023.

[5] Ljubešić Nikola, Erjavec Tomaž, Fišer Darja. KAS-term and KAS-biterm: Datasets and baselines for monolingual and bilingual terminology extraction from academic writing // Proceedings of the Conference on Language Technologies and Digital Humanities. – 2018. – P. 168–174.

[6] Loukachevitch Natalia, Manandhar Suresh et al. NEREL-BIO: A Dataset of Biomedical Abstracts Annotated with Nested Named Entities // Bioinformatics. – 2023. – Vol. 39. – Issue 4. DOI: https://doi.org/10.1093/bioinformatics/btad161.

[7] QasemiZadeh Behrang, Schumann Anne-Kathrin. The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). – Portorož, Slovenia: European Language Resources Association (ELRA), 2016. – P. 1862–1868. Access mode: https://aclanthology.org/L16-1294.pdf.

[8] Terryn Ayla Rigouts, Hoste Véronique, Drouin Patrick, Lefever Els. TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset // Proceedings of the 6th International Workshop on Computational Terminology. – Marseille, France: European Language Resources Association, 2020. – P. 85–94.

[9] Terryn Ayla Rigouts, Hoste Véronique, Lefever Els. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora // Language Resources and Evaluation. Springer, 2020. – Vol. 54(6). – P. 385–418. DOI: https://doi.org/10.1007/s10579-019-09453-9.

[10] Tran Thi Hong Hanh, Martinc Matej, Caporusso Jaya, Doucet Antoine, Pollak Senja. The Recent Advances in Automatic Term Extraction: A survey // Computing Research Repository. – 2023. – Vol. arXiv:2301.06767. – DOI: https://doi.org/10.48550/arXiv.2301.06767.

[11] Tutubalina Elena, Alimova Ilseyar et al. The Russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews // Bioinformatics. – 2021. – Vol. 37. – Issue 2. – P. 243–249. DOI: http://dx.doi.org/10.1093/bioinformatics/btaa675.

[12] Zhang Sheng, Cheng Hao, Gao Jianfeng, Poon Hoifung. Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning // Computing Research Repository. – 2023. – Vol. arXiv:2208.14565. Access mode: https://arxiv.org/abs/2208.14565.