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

April 23–25, 2025

Comparing Transformer-Based Approaches for Term Recognition in Russian texts

Semak V. V.
Lomonosov Moscow State University
Moscow, Russia
vlad.semakk@gmail.com

Bolshakova E. I.
Lomonosov Moscow State University
HSE, Moscow, Russia
eibolshakova@gmail.com

Abstract

The paper describes an experimental comparative study of three transformer-based approaches applied for automatic term recognition (ATR) in Russian texts. The approaches include sequential labeling of word tokens in a given text, phrase classification with context sentence enclosing the phrase, and text span prediction by using vector representations obtained by contrastive learning. The BERT-based models were trained and evaluated on the data of RuTermEval-2024 competition for nested term identification and classification, which encompassed three tasks: binary term identification, term recognition with classification (into one of predefined types), and cross-domain term recognition. The experiments have shown that the span prediction models based on contrastive learning outperform the other models across all three RuTermEval tasks, but at the same time demonstrate the most significant decrease in quality in the cross-domain task.

Keywords: automatic term recognition; ATR; nested terms; transformer-based term recognition models; RuTermEval competition

DOI: 10.28995/2075-7182-2025-23-323-330

Сравнение подходов на основе трансформеров для распознавания терминов в русскоязычных текстах

Семак В. В.
МГУ имени М. В. Ломоносова
Москва, Россия
vlad.semakk@gmail.com

Большакова Е. И.
МГУ имени М. В. Ломоносова
НИУ ВШЭ, Москва, Россия
eibolshakova@gmail.com

Аннотация

В статье описывается экспериментальное сравнительное исследование трех подходов на основе трансформеров, примененных для автоматического распознавания терминов в русскоязычных текстах. Подходы включают последовательную разметку слов-токенов в заданном тексте, классификацию фраз с учетом включающего фразу контекстного предложения, и предсказание спанов (отрезков) текста с использованием векторных представлений, полученных методом контрастивного обучения. Модели на основе архитектуры BERT были обучены и протестированы на данных соревнования RuTermEval-2024 по идентификации и классификации вложенных терминов, которое охватывало три задачи: бинарную идентификацию терминов, распознавание терминов с классификацией (на один из предопределенных типов) и кросс-доменное распознавание терминов. Эксперименты показали, что модели классификации спанов, основанные на контрастивном обучении, превосходят другие модели на всех трех задачах RuTermEval, но в то же время демонстрируют наиболее значительное снижение качества в кросс-доменной задаче.

Ключевые слова: автоматическое распознавание терминов; вложенные термины; модели извлечения терминов на основе трансформеров; соревнование RuTermEval

1 Introduction

Automatic term recognition (ATR), or automatic term extraction [5, 10, 14] is a natural language processing task aiming to recognize in texts and extract terms — words and phrases denoting concepts of specialized problem domains (e.g., *refactoring*, *legacy code, transfer learning, surface tension force*). With rapid development of scientific domains and emergence of new terms, ATR became essential for such applications as automated construction of terminology dictionaries, thesauri, glossaries and subject indexes, as well as machine translation of scientific texts. Despite being studied for some decades, ATR remains a challenging task, since proposed methods still can not achieve human-like performance (which is explained by complex semantic nature of terminology and high variability of term features across different domains).

The traditional and well-studied statistical approach to ATR [5, 14] relies on extraction from text phrases matching some grammatical patterns (such as *ADJECTIVE+NOUN* for terms like *linear function* and *magnetic field*) and then ranging the obtained list of extracted term-candidates by frequency-based statistical measures (e.g., MI or C-value) in order to obtain true terms on the top of the ranked list. Such techniques is sense-agnostic and give only 30–60% average precision for extracted terms, requiring human efforts to select terms from ranked term-candidates list or additional domain-specific heuristics for ranking and selecting terms.

Machine learning (ML) for automatic term extraction began to be applied since the 2000s, leveraging traditional supervised methods (like logistic regression, random forests, CRF and so on) to automatically integrate multiple term features into a binary classifier to categorize term-candidates as terms or non-terms. Such classifiers take into account many term features: grammatical, statistical, orthographical, and contextual, thus achieving higher precision of term extraction compared to models based on the traditional approach. In particular, in the work [7] machine-learning model trained with about 130 various features showed F1-score of 55% against 28% for traditional ATR methods.

Although machine learning approach improves the quality of term recognition, the trained machine classifiers remain sensitive to domain differences, showing worse results for texts in another (different of target) problem domain.

Recent papers [4, 6, 9, 11] have proposed new ML approaches to term recognition, leveraging modern deep learning transformer-based techniques (mainly BERT [2]) in order to further increase ATR quality and to tackle the problem of cross-domain transfer for trained classifiers.

The works [9, 11] apply sequential labeling of word tokens in the text being processed, this approach was initially-proposed and widely-used for named entity recognition (NER) in texts, well-known NLP problem [3, 12] close to ATR. Sequential labeling implies token-level classification, when each word token is assigned a label indicating whether it is part of term (or part of named entity in NER). Another approach presented in [4, 6] relies on binary classification of phrases (term-candidates) preliminary extracted from text, which is somewhat similar to traditional ML classifiers for ATR, but the phrases are classified together with their enclosing sentence (i.e. with their context). Experimental results showed that ATR models built within these approaches achieve up to 69-75% of F1-score for term extraction, the results depend on datasets and specialized domains taken for training, It should be noted that most transformer-based models were built and evaluated with the data from ACTER corpus [8] that contains detailed manual annotations of terms and encompasses terms from texts in three natural languages (English, French, and Dutch) and for four specialized domains. For Russian, for a long time there was no freely available corpus (or even dataset) suitable for training various ATR models and for comparing approaches. Recent paper [1] reports a comparison of ATR approaches, in relation to Russian texts, but only for a dataset with terms in domains of mathematics and programming.

In this work, a comparative study of the transformer-based ATR approaches was conducted with annotated data created for RuTermEval-2024 competition devoted to nested term identification and classification. The competition provides datasets with manually annotated terms including named entities, nested terms (such as *differential equation* and *linear differential equation*) were annotated as well. RuTermEval-2024 include three tasks (each task implies identification of nested terms): (1) binary term identification, (2) term recognition with classification (into one of predefined types), (3) cross-domain term recognition and classification. Besides two above-described ATR approaches, namely sequential labeling of tokens and phrase classification with context sentence, we also considered the approach that gives state-of-the-art results for nested NER, that is text span prediction by using vector representations obtained by contrastive learning [13].

The experiments with BERT-based models trained and evaluated in the considered approaches have shown that the span prediction models based on contrastive learning outperform the other models across all RuTermEval tasks, but at the same time demonstrate the most significant decrease in quality in the cross-domain task. In the competition, our best models achieve F1-score of 77% for binary term identification (Task 1), 70% for term recognition with classification (Task 2), and only 47% for cross-domain recognition with classification (Task 3), placing second for each task.
The next section presents a short overview of the transformer-based approaches studied in our experimental study. Then, we describe implementation of our BERT-based models for term recognition and discuss results of experiments with them. The final section presents conclusions.

2 Approaches for ATR

Transformer-based phrase classification with context sentence was initially proposed for term recognition in the paper [4]. In this approach, various phrases (e.g., n-grams) are first extracted from the text being processed, together with their corresponding context sentences, and then each phrase-sentence pair is categorized by BERT-classifier, for determining whether the phrase is term or not. For training such binary classifier, besides positive phrase-sentence pairs taken from a dataset with annotated terms, negative pairs of non-terms and their enclosing sentences are to be formed.

In the work [4] n-grams were taken as term-candidates for classification, and necessary sets of positive and negative sample pairs were built with multilingual ACTER corpus [8] (negative samples were formed by random selection of n-grams). The trained binary transformer-based classifiers (specifically, RoBERTa for English and CamemBERT for French were used) were evaluated and compared with XGBoost model, which was built by traditional ML techniques with the same data, through feature engineering with certain linguistic and statistical term features. Experiments demonstrated significant advantage of transformer-based phrase-sentence classification approach: XGBoost model showed high precision but low recall, thus giving about 27% F-score, while the BERT-based classification models showed in average 48%. Contextualized embeddings of BERT-based classifiers obviously enhance detection of terms, thus increasing precision of their prediction, and this transformer-based approach also eliminates huge feature engineering. Another important circumstance is equal handling of all extracted term-candidates, which enables recognition of nested terms, but at the same time, extracting from the text all possible phrases (as term-candidates) is quite expensive computational task.

Sequence labeling approach to ATR does not require preliminary extraction of terms candidates, instead, terms occurrences are directly detected in texts by classifier performing tagging of text tokens (similar to the task of named entity recognition [12]). As a rule, BIO tag scheme is used, and each word token in the text is labeled by the classifier with one of three tags: B (beginning of term), I (inside term word) or O (word not belonging to term). The paper [9] describes such sequence labeling classifiers built with recurrent neural network (RNN) and with certain types of input word embeddings, and their comparison is reported, with an analogous classifier built with CRF (conditional random field) method and a vast set of linguistic and statistical features. All classifiers were trained on data from multilingual ACTER corpus, the best performance with 75% F-score was achieved by RNN model with multilingual BERT embeddings and additional using training samples from several languages.

In the work [11], sequential labeling approach for ATR was applied to Slovenian texts, BERT-based models were trained on RSDO5 corpus containing manually annotated data from four specialized domains (biomechanics, chemistry, veterinary science, and linguistics). Experiments in cross-domain settings were conducted, when models were trained in one specialized domain and tested in another. Evaluation of implemented cross-domain models (12 classifiers) showed F-scores ranging from 64% to 71%, which means possibility of domain transfer for term recognition models.

These two above-described approaches were studied and compared in cross-language experiments in the work [6] exploiting pretrained multilingual model XML-RoBERTa and ACTER corpus. Trained binary classifiers for phrase-sentence pairs demonstrated F1-score up to 58%, while the token labeling approach showed the highest F1-score of 69.8%. At the same time, analogous comparison for Russian texts performed in recent work [1] showed some advantage of binary phrase-sentence classification (70,5% against 52,5% of F1-score), so the question about superiority of one approach over another remains open. It should be noted that for recognizing nested terms, the sequential labeling approach has limitations compared with phrase-sentence classification, not all possible cases of nested terms may be recognized, or more complicated (than BIO-scheme) token labeling is required.

The problem of recognizing nested entities in texts is quite important for named entities (NE), and attempts to tackle the problem has led to paradigm shift in NER, from predominant sequence labeling to span prediction (see, for example, [3]). The new paradigm implies identifying NE as text spans (fixed by their boundaries) and also classifying them into predefined set of entity types (person, location, and so on). The state-of-the-art results for NER were obtained within this research approach in the work [13] that describes Binder, new architecture and framework for training NER models.

Binder (BI-encoder for NameD Entity Recognition) performs NER as classification of text spans by using their vector representations obtained by contrastive learning. Two pretrained BERT encoders are employed to separately map into the same vector space text spans to be classified and also textual descriptions of named entity types (e.g., “person” entity type may be described as “individuals or groups of individuals identified through proper nouns”). The common vector space makes it possible to calculate similarity between representations of the spans and the considered entity types. Contrastive learning follows two objectives: (1) maximizing similarity scores between representations of entity types and named entity spans, and (2) minimizing similarity scores between representations of entity types and spans for non-entities (and also between representations of entities and inappropriate entity types), thus separating non-entity spans from entity mentions. The contrastive learning with these objectives is supplemented with dynamic thresholding loss to work out specific dynamic threshold to distinguish entity spans from non-entity ones.

The described Binder's approach makes it easy to handle both nested and flat NE, and trained Binder models outperform previous NER models, across several benchmarks, achieving up to 90% F1-score for nested NER (e.g., on ACE2005 and GENIA corpora) and 95% F1-score for flat NER.

Thus, the above-described transformer-based approaches and models for ATR have been tested on various datasets for several languages (but few experiments for Russian), sometimes showing contradictory results in quality, which most likely depends on language, specialized domains of processed texts, and evidently data for training. So we have chosen two approaches — phrase classification with context sentence and sequential labeling of tokens — for our experimental study within RuTermEval-2024 competition for Russian texts. As the competition was devoted to nested terms, we also considered in our comparison Binder's span prediction approach (which has not previously been used for ATR), with its adaptation to Russian and to tasks of the competition. In particular, instead of textual descriptions of named entity types, text definition of concept “term” was used, as well as textual descriptions of term types (classes).

3 Experiments and Results

3.1 RuTermEval Tasks and Dataset

RuTermEval-2024 competition on nested term identification and classification in Russian scientific texts includes three tasks (challenges). Binary term identification (Task 1) aims to detect position boundaries of all terms in a given text. Term recognition with classification (Task 2) requires not only to detect terms’ positions but also to classify each term into one of three predefined classes:

- specific terms − terms that are lexically specific and belonging to the considered scientific domain (e.g. *идиома − Eng. idiom, эпистемическая модальность − Eng. epistemic modality*);
- common terms − terms that are domain specific and at the same time may be known and used by non-specialists, as they consist of words of general lexicon (e.g. *словарь − Eng. dictionary, носитель языка − Eng. native speaker*);
- nomens − names of domain-specific objects, such as names of databases and datasets, programming languages, corpora, dictionaries, programs and software packages, etc. (e.g. *Интернет − Eng. Internet, тезаурус Роже − Eng. Roget’s thesaurus*).

Cross-domain term recognition and classification (Task 3) involves detecting all term positions in the text and their classification into one of the above-described classes, which is similar to Task 2, but for texts from scientific domains different from the domains used for training ATR models.

For evaluating participant’s term recognition models, RuTermEval-2024 provided CL-RuTerm3 dataset with annotated terms, including nested terms, such as *interrogative pronoun* — *pronoun*). The dataset contains: 1) training subset, which covers 850 abstracts and certain full-text Russian scientific papers in domains of linguistics and computational linguistics; 2) the subset for development (testing and evaluation of models), with 103 abstracts and 10 full-text papers in the same domain; 3) the subset for testing and evaluation in Task 3, which includes 100 abstract and texts from two different problem domains (literary studies and agricultural technology).

We compared some characteristics of CL-RuTerm3 with ACTER corpus that is the most often used in research papers on ATR. The training subset of CL-RuTerm3 is comparable in size with ACTER subcorpora (about 60 K tokens in each specialized domain and language), but the subset significantly differs in diversity of included terms (e.g., English subcorpus of ACTER on heart failure domain includes 2519 unique terms with nomens, while CL-RuTerm3 subset contains 7832). This is most likely explained by broader domain of CL-RuTerm3 (two wide specialized domains: linguistics and computational linguistics) and also by fragmentary nature of it (mainly abstracts of papers not full paper texts). Our analysis revealed that most terms of the training subset are encountered in texts rarely or extremely rarely − see Fig. 1, demonstrating distribution for frequencies of unique terms' occurrences, across term classes. Nomens are the most poorly represented in the texts (296 nomen were found only once). This feature may evidently influence quality of the trained models.

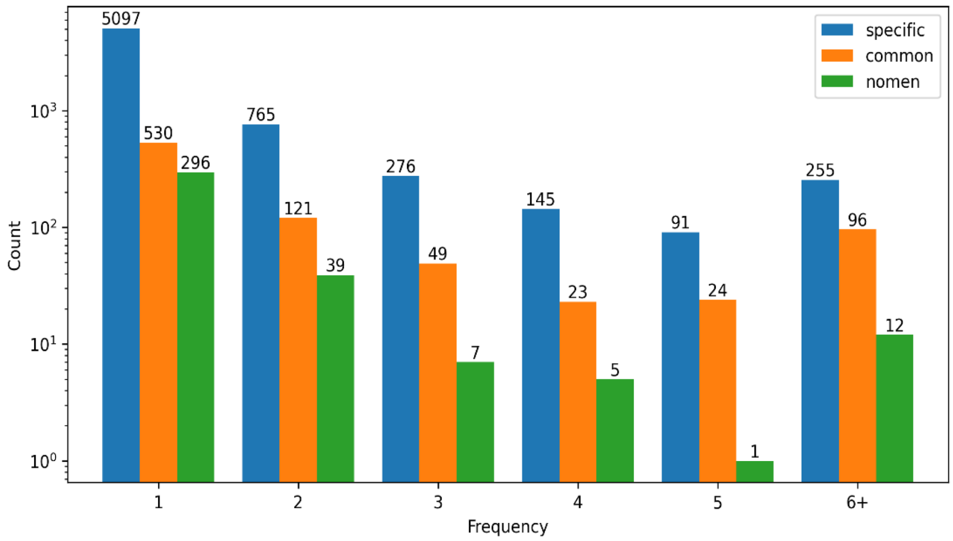

Figure 1: Statistics of term occurrences for term classes in training CL-RuTerm3 subset

3.2 Training of Models

For our experiments, we used pretrained ruBert-base [14] model from SberDevices — BERT [1] initially trained for Russian on texts from Russian Wikipedia, news, books, web, film subtitles and so on. For each of RuTermEval-2024 tasks, ruBert-base models were fine-tuned according to the considered approaches:

- classification of phrase-sentence pairs, when the fine-tuned model classifies a phrase from a given input pair as term or non-term (binary classification) or additionally produces particular term class (specific, common, nomen) for the recognized term (multiple classification);
- sequential labeling of words, when each word token in a given text is classified according to BIO tagging scheme (as part of term or not), or additionally produces particular class (specific, common, nomen) for such recognized token;
- span prediction based on contrastive learning, when for a given text the trained model processes all possible text spans with length up to predefined value and recognizes class for each span: binary (is it term or non-term) or particular term class (specific, common, nomen).

Eventually, nine ATR models were trained for experiments (three models for each of the three competition's tasks), all of them were fine-tuned for 10 epochs with AdamW optimizer, batch size of 8, and learning rate 5e-5. Among data provided for training in the competition, 10% portion was randomly selected for validation during training itself, and the model with the best F1-score on the validation data among all epochs was taken as the final model.

For classification of phrase-sentence pairs, positive examples were constructed with annotated terms from the training subset of CL-RuTerm3 dataset, while negative examples were generated by randomly selecting n-grams taken from the subset and consisting of adjectives and nouns, the number of negative examples was equal to the number of positive examples. The models for phrase-sentence classification and for sequential labeling of tokens were developed with the transformers library[1].

Span prediction models based on contrastive learning were implemented with the Binder framework[2] developed and described in [12], which requires to set for training, besides other parameters (e.g. maximum span length was 30 tokens of the BERT-model), natural language descriptions of units being detected and classified: that is text description for concept “term” (Task 1) and text descriptions of terms’ classes (Task 2 and Task 3). We have adapted for this purpose descriptions of RuTermEval tasks given by organizers of the competition. For example, text description for “specific term” was taken as *“Термин, обозначающий концепцию или явление, специфичное для определенной области и её лексики” (Eng. “Term denoting a concept or phenomenon specific to a particular domain and its lexicon”*).

3.3 Results in Competition

In the competition, quality of the trained models was evaluated by the F1-score. For Task 2 and Task 3, two modes for evaluation were exploited: classified, i.e. with term classes (F1-score is calculated based on predicted terms' positions and classes); and unclassified, i.e. without term classes (F1-score is calculated only based on predicted terms’ positions, predicted terms’ classes are ignored). The results of our models are presented in Table 1 (the unclassified mode for Task 2 is identical to Task 1).

Approach of Model	Task 1	Task 2		Task 3	
		Classified	Unclassified	Classified	Unclassified
Phrase classification	0.44	0.41	0.44	0.33	0.37
Sequential labeling	0.57	0.55	0.57	0.40	0.44
Span prediction	**0.77**	**0.70**	**0.77**	**0.47**	**0.51**

Table 1: Quality of term recognition in Leader board

The experiments demonstrate that the models of span prediction approach consistently and noticeably outperform the models of alternative approaches (sequential labeling and phrase-sentence classification), across all RuTermEval tasks. Specifically, the span prediction models achieved 77% F1-score for binary term identification (Task 1), 70% for term recognition with classification (Task 2), and 47% for cross-domain term recognition and classification (Task 3). Models of all the considered approaches exploit BERT contextual embeddings, so the advantage of the span prediction approach is most likely explained by contrastive learning for spans’ representations in the common vector space.

Though the span prediction approach demonstrates better quality across all tasks, it also shows the highest sensitivity to task complexity: additional term class prediction required in Task 2 led to 4-7% decrease of F1-scores (for other approaches decrease is 2-3%), while the cross-domain Task 3 resulted in a dramatic 23-26% drop in F1-scores (8-15% for other models).

One can notice evident correlation between task complexity and quality of the models: simple term identification (Task 1) yields the best results, additional term class prediction (specific terms, common terms, or nomen in Task 2) slightly degrades the quality (2-7% across the models), but change of specialized domain (cross-domain term recognition and classification, Task 3) proved particularly challenge, with F1-scores dropping by 8-26% compared to Task 2.

The results obtained in our work correspond to the second place in RuTermEval competition for all its tasks, and the gap from the winner is quite insignificant in Task 2 (F1-scores 0.6996 versus 0.6997) and relatively small in Task 1 (0.769 versus 0.794).

---

[1] https://huggingface.co/docs/transformers/index

[2] https://github.com/microsoft/binder

3.4 Comparing Precision and Recall for Models and Term Classes

Besides F1-score that measures averaged quality of the trained models, precision and recall estimate additional important characteristics, so we also have evaluated them for our models on the development subset of CL-RuTerm3, the only available for this purpose. The results are given in Table 2 for all term classes in total, and in Table 3 for the classes separately. For sequential labeling model, recall is consistently worse than precision, this may be explained by difficulties of the approach to recognize nested terms. For phrase classification, situation varies for different Tasks and term classes, while for span classification, recall and precision often are approximately the same, except Task3, when recall drops sharply (probably due to good adaptation to the training subset).

Approach of Model	Task 1			Task 2			Task 3		
	P	R	F1	P	R	F1	P	R	F1
Phrase classification	0.67	0.68	0.66	0.36	0.52	0.40	0.27	0.28	0.26
Sequential labeling	0.77	0.57	0.65	0.68	0.52	0.58	0.51	0.24	0.32
Span classification	**0.83**	**0.78**	**0.79**	**0.79**	**0.74**	**0.76**	**0.63**	**0.25**	**0.33**

Table 2: Quality of term recognition on development subset

The data in Table 3 for three considered term classes shows that in both Tasks specific terms are recognized by sequential labeling model somewhat better than common terms, but for phrase classification the situation is opposite. As for span classification model, in Task 2 it gives about equal results, but it has worse quality for common terms. Nomens are recognized very poorly by all models, the worse result (0.03 of F1-measure, in cross-domain Task 3) belongs to span classification model, in this term class it unexpectedly loses to other approaches. Such worse quality for nomens is likely caused by their diversity (personal names, names of text corpora, databases, etc.) and by simultaneously rare frequency in the dataset used for training.

Approach of Model	Specific			Common			Nomen		
	P	R	F1	P	R	F1	P	R	F1
Task 2									
Phrase classification	0.35	0.50	0.37	0.41	0.53	0.42	0.11	0.17	0.12
Sequential labeling	0.68	0.56	0.60	0.60	0.39	0.44	**0.30**	**0.21**	**0.23**
Span classification	**0.77**	**0.75**	**0.75**	**0.76**	**0.77**	**0.75**	0.26	0.15	0.18
Task 3									
Phrase classification	0.23	0.23	0.21	0.29	**0.33**	**0.26**	0.17	**0.17**	**0.15**
Sequential labeling	0.55	0.26	0.34	0.25	0.14	0.15	**0.22**	0.10	0.13
Span classification	**0.64**	**0.31**	**0.40**	**0.36**	0.12	0.16	0.19	0.02	<u>0.03</u>

Table 3: Quality of recognition for term classes

Our analysis of errors in term classes recognized by the models revealed that all they more often tend to get confused with classification between specific and common terms, e.g., specific term *часть речи* (*parts of* speech) was incorrectly categorized as common, while *деловая коммуникация* (*business communication*) was classified as specific instead of common. However, distinguishing between these classes is difficult even for a terminologist, due to the lack of clear boundaries between term domains.

4 Conclusion

The described comparative study of modern transformer-based approaches to automatic term recognition and classification, which has been applied to Russian texts, has revealed the advantage of span prediction by contrastive learning, but at the same time this approach may be particularly vulnerable to domain shift and task complexity. Nevertheless, further experiments for the considered approaches are needed, relying on representative annotated datasets of various sizes and with texts of various specialized domains. Term annotations is a complex task even for humans (especially identification of term class), and datasets for training often vary significantly across domains and languages, and in some cases different approaches may be more effective, not only span prediction. The more challenging research task is cross-domain term recognition, which is more necessary in applications because of the lack of terminologically annotated data for new specialized domains.

References

[1] Bolshakova E.I., Semak V.V. Methods and means of term extraction from texts for terminological tasks [Metody i sredstva izvlecheniya terminov iz tekstov dlya terminologicheskikh zadach] // Programming products and systems [Programmnye produkty i sistemy]. — 2025. — V. 38 — №. 1. — pp. 5–16.

[2] Devlin, J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 NAACL Conference: Human Language Technologies, Minneapolis — 2019 — pp. 4171–4186.

[3] Fu, J., Huang X., Liu P. SpanNER: NamedEntity Re-/Recognition as Span Prediction // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conference on Natural Language Processing — 2021 — pp. 7183–7195.

[4] Hazem, A., Bouhandi, M., Boudin, F., Daille, B. TermEval 2020: TALN-LS2N System for Automatic Term Extraction // Proceedings of the 6th International Workshop on Computational Terminology — 2020 — pp. 95–100.

[5] Korkontzelos, I., Ananiadou, S. Term Extraction // Oxford Handbook of Computational Linguistics — 2014 — 2nd Ed. — Oxford University Press, Oxford — pp. 991–1012.

[6] Lang, C. et al.: Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 — 2021 — pp. 3607–3620.

[7] Terryn, A.R. et al. Analysing the impact of supervised machine learning on automatic term extraction: HAMLET vs TermoStat // Proceedings of the Int. Conference on Recent Advances in Natural Language Processing (RANLP 2019) — 2019 — pp. 1012-1021.

[8] Terryn, A.R. et al. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset // 6th International Workshop on Computational Terminology (COMPUTERM 2020) — 2020 — pp. 85-94.

[9] Terryn, A.R., Hoste V., Lefever E. Tagging terms in text: A supervised sequential labelling approach to automatic term extraction // Terminology. Int. Journal of Theoretical and Applied Issues in Specialized Communication — 2022 — Issue 28, Vol. 1 — pp. 157–189.

[10] Tran, H.T.H. et al. The recent advances in automatic term extraction: A survey // arXiv preprint — 2023 — Access mode: https://arxiv.org/abs/2301.06767

[11] Tran, H.T.H., Martinc M., Doucet A., Pollak, S. A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction // Slovenian Conference on Language Technologies and Digital Humanities — 2022 — Ljubljana — pp. 196-204.

[12] Yadav, V., Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models // Proceedings of the 27th International Conference on Computational Linguistics — 2018 — ACL, Santa Fe, USA — pp. 2145–2158.

[13] Zhang, S. et al. Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning // The Eleventh International Conference on Learning Representations — 2022 — Access mode: https://arxiv.org/abs/2208.14565

[14] Zhang, Z., Iria J., Brewster, C., Ciravegna, F. A Comparative Evaluation of Term Recognition Algorithms // Proc. of the Sixth Int. Conf. on Language Resources and Evaluation (LREC'08) — 2008 — pp. 2108–2111.

[15] Zmitrovich, D., et al. A Family of Pretrained Transformer Language Models for Russian // arXiv preprint — 2023 — Access mode: https://arxiv.org/abs/2309.10931