Development of Cross-Language Embeddings for Extracting Chemical Structures from Texts in Russian and English

Alexey I. Molodchenkov, Dmitry A. Deviatkin, Sergey A. Loginov, Alexey Yu. Lupatov, Alisa M. Gisina, Anton V. Lukin

Abstract—**This study is dedicated to describing an algorithm for implementation cross-lingual embeddings to extract chemical structures from texts in both Russian and English. The proposed algorithm focuses on fine-tuning of pre-trained models based on transformer architecture. After analyzing existing models, mBERT and LaBSE were selected. The training datasets for these models included texts related to chemistry and adjacent fields of science. Fine-tuning was done using a collected set of scientific articles and patent texts in Russian and English. For English, the ChemProt corpus was also used. The model was trained on tasks such as masked language modeling and entity recognition. Comparisons were made with several models, including BioBERT. The results of the experiments showed that the proposed implementation of embeddings more effectively solve the task of recognition chemical structure names in texts in both Russian and English.**

Keywords—**Embeddings; transformer architecture; information extraction; chemical structures.**

I. Introduction

Methods for extracting information from texts across various domains (chemistry, biochemistry, crystallography, medicine) are fundamental in creating professional tools for automating searches through large, heterogeneous document collections (patents, scientific publications, dissertations, clinical trial results). At the same time, training such methods requires the availability of annotated text corpora, whose creation is a labor-intensive process due to the high complexity of annotation. Addressing this challenge requires the involvement of highly qualified specialists in the respective fields. As a result, such methods are underdeveloped for many languages except English, and developing general approaches to creating these methods under conditions of limited linguistic resources is of significant scientific importance.

Most models (except domain-specific ones) are trained on general datasets. This means that such corpora typically consist of texts from news, various genres of literature, as well as internet resources and encyclopedias. In our study, we address the problem of developing algorithms for cross-lingual information extraction from texts in Russian and English in the domains of chemistry, biochemistry, and related fields. Unfortunately, the share of data from such domains in standard datasets is extremely small due to the specific nature of the knowledge domain. Solving these tasks requires domain-specific fine-tuning of models to correctly interpret domain information. Without such refinements, even the largest and most advanced models will not be able to extract and correctly interpret context from specialized information. One of the stages in training cross-lingual models for entity extraction from texts involves pre-training the model to build embeddings.

To solve this task, we focused on models that could be fine-tuned on our dataset. Several approaches and models were considered for solving the problem posed in our research. The first approach was the application of Bag-Of-Words and Skipgram models [1], as they represent some of the earliest successful examples of implementing cross-lingual embeddings. Additionally, various modifications of convolutional neural networks (CNNs) [2] and recurrent neural networks (RNNs) [3] were applied. However, our preliminary experiments showed that these approaches were highly inefficient. Therefore, the main interest was in transformer-based models [4]. Two models were selected for fine-tuning, and their training datasets included texts from the domains of chemistry, biochemistry, and related fields. An additional dataset was collected for fine-tuning.

II. Data collection

The search for texts was conducted using the SciApp digital platform for aggregation and analysis of scientific and technical information [5]. The search for Russian-language texts was performed across the following resources: cyberleninka.org, FIPS. Inventions, Dissertation Abstracts, Russian Scientific Conferences, and Russian Journals. For English-language texts: PubMed, foreign journals, USPTO patents, and Eurasian patents from EAPO. The search query consisted of one term from a list of terms denoting types of interactions between a chemical substance and a protein. The list was based on an analysis of representative scientific

Alexey I. Molodchenkov is with the Federal Research Center "Computer science and Control" of the Russian Academy of Sciences, and with the RUDN University, Moscow, Russia (corresponding author to provide e-mail: aim@tesyan.ru).

Dmitry A. Deviatkin is with the Federal Research Center "Computer science and Control" of the Russian Academy of Sciences, Moscow, Russia

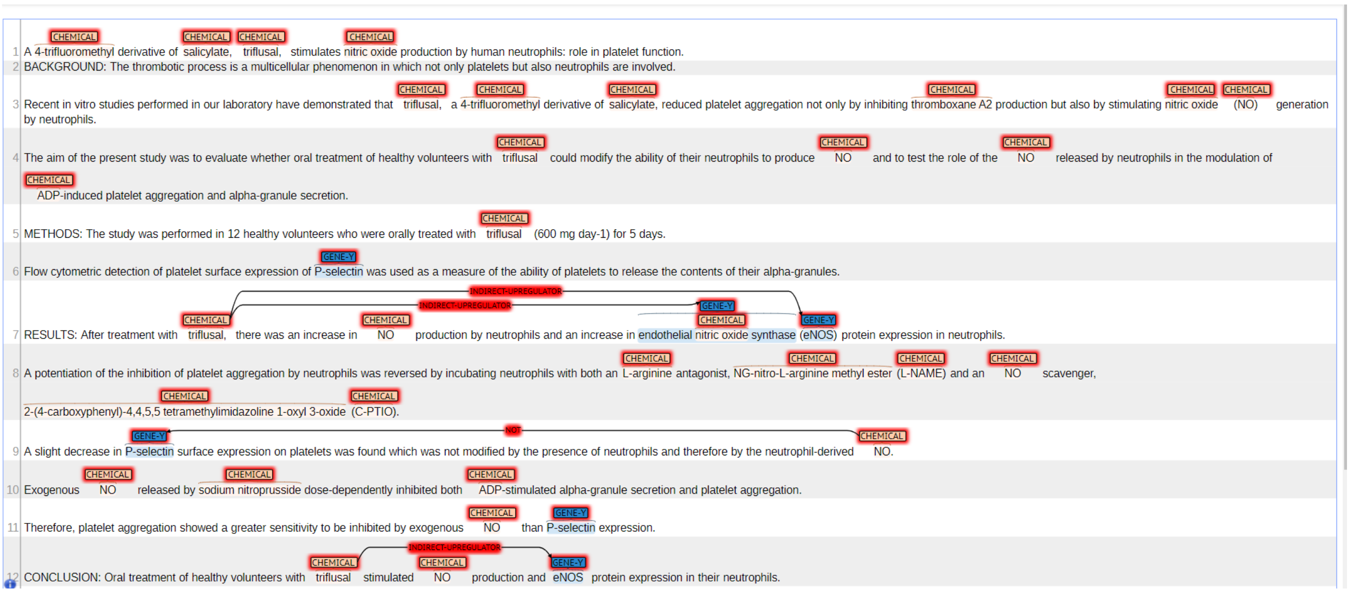
Sergey A. Loginov is with the RUDN University, Moscow, Russia

Alexey Yu. Lupatov is with the Institute of Biomedical Chemistry of the Russian Academy of Sciences, Moscow, Russia

Alisa M. Gisina is with the Institute of Biomedical Chemistry of the Russian Academy of Sciences, Moscow, Russia

Anton V. Lukin is with the Federal Research Center "Computer science and Control" of the Russian Academy of Sciences, and with the RUDN University, Moscow, Russia.

texts on the subject and included interaction types most frequently mentioned by authors, such as activator, agonist, inhibitor, antagonist, cofactor, regulator, modulator, and substrate. As an additional step, when expanding the English-language dataset, a filter was applied to select USPTO patents that had Russian patents in the same patent family. To automate this process, semi-structured patent descriptions in XML format were downloaded from the USPTO portal for the period from 2002 to 2022. Then, using the numbers of the identified patents, records with references to Russian patents distance is a generalization of Euclidean distance that in the <priority-claims> section were selected. Experts subsequently chose articles relevant to the topic from the search results. The English-language texts were further expanded using the ChemProt corpus [6]. Additionally, the dataset was enlarged by applying a function to search for similar documents. Text annotation was carried out by experts using active learning methods. Fig. 1 shows an example of the annotation process.



1 A 4-trifluoromethyl derivative of salicylate, triflusal, stimulates nitric oxide production by human neutrophils: role in platelet function.

2 BACKGROUND: The thrombotic process is a multicellular phenomenon in which not only platelets but also neutrophils are involved.

3 Recent in vitro studies performed in our laboratory have demonstrated that triflusal, a 4-trifluoromethyl derivative of salicylate, reduced platelet aggregation not only by inhibiting thromboxane A2 production but also by stimulating nitric oxide (NO) generation by neutrophils.

4 The aim of the present study was to evaluate whether oral treatment of healthy volunteers with triflusal could modify the ability of their neutrophils to produce NO and to test the role of the NO released by neutrophils in the modulation of ADP-induced platelet aggregation and alpha-granule secretion.

5 METHODS: The study was performed in 12 healthy volunteers who were orally treated with triflusal (600 mg day-1) for 5 days.

6 Flow cytometric detection of platelet surface expression of P-selectin was used as a measure of the ability of platelets to release the contents of their alpha-granules.

7 RESULTS: After treatment with triflusal, there was an increase in NO production by neutrophils and an increase in endothelial nitric oxide synthase (eNOS) protein expression in neutrophils.

8 A potentiation of the inhibition of platelet aggregation by neutrophils was reversed by incubating neutrophils with both an L-arginine antagonist, NG-nitro-L-arginine methyl ester (L-NAME) and an NO scavenger, 2-(4-carboxyphenyl)-4,4,5,5 tetramethylimidazoline 1-oxyl 3-oxide (C-PTIO).

9 A slight decrease in P-selectin surface expression on platelets was found which was not modified by the presence of neutrophils and therefore by the neutrophil-derived NO.

10 Exogenous NO released by sodium nitroprusside dose-dependently inhibited both ADP-stimulated alpha-granule secretion and platelet aggregation.

11 Therefore, platelet aggregation showed a greater sensitivity to be inhibited by exogenous NO than P-selectin expression.

12 CONCLUSION: Oral treatment of healthy volunteers with triflusal stimulated NO production and eNOS protein expression in their neutrophils.

Fig. 1: An example of annotation

The active learning process involves further annotating the most difficult examples for the model to recognize and subsequently fine-tuning the model on those examples. These textual examples are identified either by assessing the probability of their belonging to target classes or by evaluating the distance between the vector representation of the object and the vector representation of the target class centroid.

Overall, the algorithm consists of two major blocks (Fig. 2.1):

- Text annotation block.
- Model fine-tuning block.

The text annotation block comprises the BRAT annotator, which receives texts for annotation from the model (identified during the active learning process) and an oracle (expert). It forms pairs of annotated files (text + annotation) for the next step of model training.

The model fine-tuning block receives new examples for annotation from the pool of unannotated texts, retrieves annotated data from the annotation module, and ensures an iterative process of model retraining. This block consists of:

- A set of unannotated texts.
- A text preprocessing step.
- A sub-block for communication with the annotator.
- A model retraining step.
- An active learning sub-block.
- A step for identifying texts for expert annotation.
- A step for evaluating model performance.

For active learning, the Mahalanobis distance on the spectral-normalized network was used. The Mahalanobis accounts for the spread of instances from the training set across different directions in feature space. Uncertainty is measured by calculating the Mahalanobis distance between a test instance and the nearest class-conditional Gaussian distribution:

$$\mu_{MD} = \left(h_i - \mu_c\right)^T \sum{}^{-1} \left(h_i - \mu_c\right) \qquad (1)$$

where h_i is the hidden representation of the i-th instance, μ_c is the centroid of class c, and Σ is the covariance matrix for the hidden representations of the training instances.

Additionally, spectral normalization of the weight matrix in the linear layer of the "Transformer" classification block was performed. At each training step, the spectral norm υ is estimated using the power iteration method:

$$v = ||W||_2 \qquad (2)$$

This yields the normalized weight matrix:

$$W = \frac{W}{v} \qquad (3)$$

In the final step, hidden representations are computed using the normalized matrix

$$\tilde{h}(x) = \tilde{W}x + b \qquad (4)$$

and are used to calculate the Mahalanobis distance.

As a result, more than 2,000 English-language texts with a total word count of over 6 million were selected, along with over 1,500 Russian-language texts of similar size. Of these, 661 English-language texts were patents that had Russian-language equivalents.

III. Model selection

The following popular models were included in the analyzed set of models: mBERT [7], XLM-RoBERTa [8], mBART [9] (although generative, it was considered to expand the list), and LaBSE [10, 11].

In addition to these models, the following were analyzed:

1. **ERNIE-m** – this model is designed for aligning embeddings across selected languages, requiring a large corpus of texts for each language [12].
2. **LASER (Language-Agnostic Sentence Representations)** – a library with source code that calculates embeddings using multilingual text corpora, supporting over 200 languages. According to the authors, the model generalizes data to languages not used during training for specific tasks [13].
3. **VECO** – an encoder-type model primarily used for generating texts in different languages, including cross-lingual generation [14].
4. **FILTER** – an enhanced merging method that takes cross-lingual data as input for fine-tuning XLM. Specifically, FILTER first encodes the text input in the source language and its translation into the target language independently in the surface layers, then performs cross-lingual merging to extract multilingual knowledge in the intermediate layers, and finally carries out further language-specific encoding [15].
5. **UNICODER** [16] – a universal language coder insensitive to language differences. For natural language processing tasks, the model can be trained using training data in one language and directly applied to the same task in other languages. Compared to similar approaches, such as Multilingual BERT and XLM, three new cross-lingual pre-training tasks are proposed, including cross-lingual word recovery, cross-lingual paraphrase classification, and cross-lingual masked language modeling [16].

Due to technical constraints and other limitations, mBERT and LaBSE were selected to solve the entity extraction task.

IV. Continual pre-training and fine-tuning

Both models have a BERT architecture, and their training process is similar, consisting of two stages: pretraining and finetuning. Both models are pretrained on large and diverse corpora, which include texts from the field of chemistry (about 7%). The architecture of these models is a multi-layer bidirectional encoder based on the original transformer.

During the pretraining stage, the model is trained (in our case, fine-tuned) on an unlabeled text corpus through various tasks to learn sentence representations from both the left and right context in all its layers. The semi-supervised tasks used in the pretraining phase are masked language modeling (MLM) and next sentence prediction (NSP) [17]. In the MLM task, 15% of input tokens are randomly masked before the sequences are processed. Once selected, a token is masked with a [MASK] token in 80% of cases, replaced with a random token in 10% of cases, and left unchanged in the remaining 10%. The data generator processes the masked sentences and selects 15% of the masked tokens to predict based on their context.

The goal of the NSP task is to predict the next sentence for jointly pretraining representations of text pairs. The model focuses on two masked sentences, A and B, to learn the relationship between them. B follows A in the original text 50% of the time, while in the other 50%, it is replaced with a sentence randomly selected from the corpus. In the NER task, sentence B corresponds to the entity labels found in sentence A. The NSP task is important for building question-answer systems and generating inferences in natural language.

However, for the entity extraction task, analysis of the articles showed that excluding NSP does not significantly degrade or improve the results, but simplifies the model training process in terms of training time. Additionally, understanding the context between words in a sentence plays a major role in entity extraction. Therefore, more attention was given to the MLM task in this study.

It is worth that the LaBSE model is trained on MLM and TLM (translation language modeling) tasks. TLM was introduced to improve cross-lingual pretraining and is an extension of MLM. Fig. 2 shows how TLM works and its difference from MLM.

TLM extends MLM to sentence pairs written in different languages. To predict a masked English word, the model can consider both the English sentence and its French translation, aligning the English and French representations. The positional embeddings of the target sentence are reset to facilitate alignment.

The mBERT model implementation was taken from [7], and the LaBSE model was adapted for English and Russian [11].

The MLM task includes the following concepts:

- $W=[w_1,w_2,\dots,w_n,]$ – the input sentence of length n
- $\hat{W}$ – the sentence with some tokens masked [MASK]
- θ – model hyperparameters
- $P(w_i| \hat{W}, \theta)$ – the probability of correctly predicting word w_i

The task is to maximize the log-likelihood function:

$$\sum_{i=1}^{n} \log P\left(w_i \middle| \widehat{W},\ \theta \right) \qquad (5)$$

The fine-tuning loop for the models consists of the following steps:

1. A certain percentage of randomly selected words are replaced with [MASK].
2. The sentences are passed to the model.
3. For each token, the model generates a prediction as a list of words from the model's vocabulary, along with scores for each word.
4. Loss calculation.
5. Weight update.

Perplexity [18] is used as the quality metric.

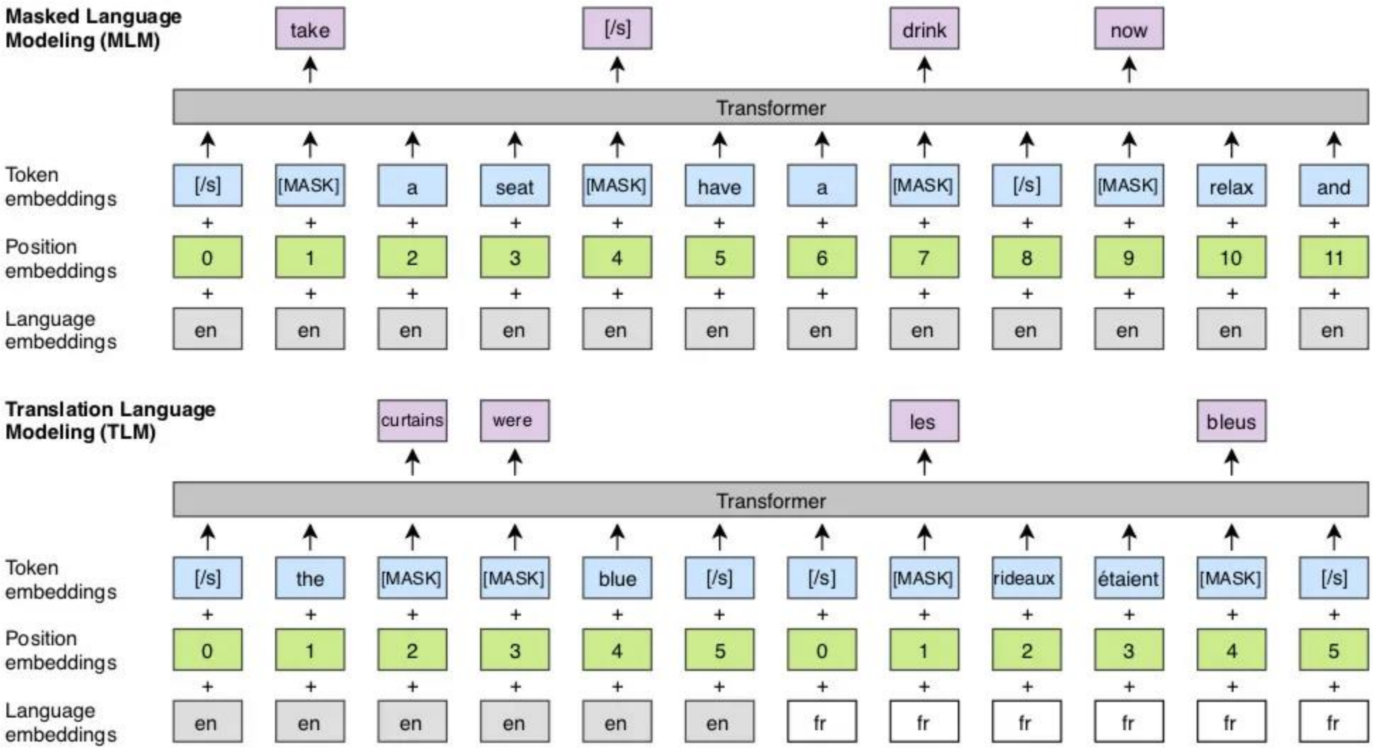


Fig. 2: The difference between MLM and TLM

The result of training is a model with weights adjusted based on the collected data.

During model training, the number of epochs was varied, and the learning curve and loss penalty functions were optimized. As a result, in the initial training of mBERT and LaBSE, the loss function values were 0.73 and 1.06, respectively, and perplexity values were 2.8 and 3.9.

Additionally, for solving the MLM task with LaBSE, a sentence alignment procedure was added based on patent texts written in both Russian and English. The alignment was based on context rather than direct translation. Consequently, the MLM task was solved using the aligned dataset, and LaBSE was fine-tuned with MLM parameters for the TLM task. Checkpoints from models trained on MLM were used for this purpose. These parameters were also applied to the mBERT model. Furthermore, the model's hyperparameters were fine-tuned during the entity extraction task. After these steps, the perplexity values for mBERT and LaBSE models decreased to 2.6 and 3.6, respectively.

V. Extraction of Named Entities

A small experiment was conducted to extract named entities—names of chemical structures. The model architectures were modified without altering the hyperparameters of the pretrained components. An additional top layer was added, serving as a classifier to solve the entity extraction task.

Next, two tokens were added to the tokenizer: B-chem and I-chem, and tokenization of words and subwords was carried out. All tags were recorded in the label2id and id2label dictionaries. After data preparation, the training process for entity extraction in the text began, followed by hyperparameter optimization. The number of epochs was set to 10 and 20.

To implement the proposed approach, additional preprocessing of the constructed dataset was performed. It included the following steps:

1. Splitting the data into two sets: training and test samples.
2. The selected models based on the BERT architecture accept sentences up to 512 characters in length as input. Therefore, all sentences exceeding this length were truncated.
3. All sentences were tokenized and grouped into sets of 128 elements (sentences).
4. A class label array was created.

The models trained using this approach were then compared with other models that were used by different groups for chemical entity extraction from texts. The comparison was based on the performance of the following models:

- mBERT,
- LaBSE,
- BioBERT – a separate multilingual model trained on PubMed articles, with a specific focus on biomedical texts, used for extracting biomedical data from texts [19],
- ChemBERTa – a model fine-tuned specifically for extracting chemical entities from texts [20],
- BERT [21].

All models were trained in the same way. For each model, a hyperparameter tuning procedure was carried out on the task of named entity recognition from texts, and the same hyperparameters were used for pretraining.

Table 1 presents the experimental results, with the f1 score used as the quality metric.

Table 1. Model comparison

Model	F1-score
mBERT without continual pre-training	0.65
mBERT with continual pre-training	0.80
LaBSE without continual pre-training	0.68
LaBSE without continual pre-training	0.79
BioBERT	0.65
ChemBERTa	0.39
BERT	0.14

The experiments showed that the proposed cross-lingual embedding approach significantly improves the quality of chemical entity recognition in texts.

VI. Conclusion

The task of developing information extraction methods from texts in highly specialized fields is highly relevant. These methods allow us to solve a wide range of tasks, such as developing specialized search engines, data collection services, and more. This study presents the results of implementation cross-lingual embeddings for texts in chemistry and related fields in both Russian and English. The proposed approach was tested on the task of extracting names of chemical structures from texts in Russian and English. The experimental results demonstrated the effectiveness of this approach for named entity extraction. Future work will focus on modifying the algorithm and approach to extracting chemical structure names to further improve the results.

References

[1] T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781;

[2] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel: Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation, 1(4):541-551, Winter 1989;

[3] H. Sak, A. Senior, F. Beaufays. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition https://doi.org/10.48550/arXiv.1402.1128;

[4] Attention Is All You Need A. Vaswani, L. Jones, N. Shazeer, N. Parmar, J. Uszkoreit, A. N. Gomez, Ł. Kaiser, I. Polosukhin https://doi.org/10.48550/arXiv.1706.03762;

[5] Sciapp [Electronic resource]. – URL: https://sciapp.ru/ (accessed: 19.09.2024)

[6] Taboureau O. et al. ChemProt: a disease chemical biology database //Nucleic Acids Research. – 2010. – Vol. 39. – No. suppl_1. – pp. D367-D372.

[7] mBERT base model [Electronic resource]. – URL: https://huggingface.co/google-bert/bert-base-multilingual-cased (accessed: 19.09.2024)

[8] Li, B., He, Y., & Xu, W. (2021). Cross-lingual Named Entity Recognition Using Parallel Corpus: A New Approach Using XLM-Roberta Alignment. arXiv preprint arXiv:2101.11112.

[9] Chipman H. A. et al. mBART: Multidimensional Monotone BART //Bayesian Analysis. – 2022. – Vol. 17. – No. 2. – pp. 515-544.

[10] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang. Language-agnostic BERT Sentence Embedding //arXiv preprint arXiv:2007.01852. – 2020, doi: https://doi.org/10.48550/arXiv.2007.01852.

[11] LaBSE base model [Internet resource]. – URL: https://huggingface.co/cointegrated/LaBSE-en-ru (accessed: 19.09.2024)

[12] X. Ouyang, S. Wang, C. Pang, Y. Sun, H. Tian, H. Wu. ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora, 2020, https://doi.org/10.48550/arXiv.2012.15674.

[13] Mikel Artetxe, Holger Schwenk; Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Transactions of the Association for Computational Linguistics 2019; 7 597–610. doi: https://doi.org/10.1162/tacl_a_00288.

[14] F. Luo, W. Wang, J. Liu, Y. Liu, B. Bi. VECO: Variable and Flexible Cross-lingual Pre-training for Language Understanding and Generation. 2020, https://doi.org/10.48550/arXiv.2010.16046.

[15] Y. Fang, S. Wang, Z. Gan, S. Sun, J. Liu. FILTER: An Enhanced Fusion Method for Cross-lingual Language Understanding. 2020, https://doi.org/10.48550/arXiv.2009.05166.

[16] H. Huang, Y. Liang, N. Duan, M. Gong, L. Shou. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. 2019, https://doi.org/10.48550/arXiv.1909.00964.

[17] Aroca-Ouellette, S., and Rudzicz, F. (2020). "On Losses for Modern Language Models," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), 4970–4981. Available online at: https://www.aclweb.org/anthology/2020.emnlp-main.403

[18] Lukashkina Yu. N., Vorontsov K. V. Assessing Stability and Completeness of Topic Models of Multidisciplinary Text Collections. [Electronic resource]. – URL: http://www.machinelearning.ru/wiki/images/4/4b/Lukashkina2017MSc.pdf (accessed: 19.10.2024)

[19] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. Bioinformatics, 36(4), 1234-1240.

[20] S. Chithrananda, B. Ramsundar, G. Grand. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction https://doi.org/10.48550/arXiv.2010.09885;

[21] BERT base model https://huggingface.co/google-bert/bert-base-uncased. (accessed: 19.10.2024)