УДК 004.89
DOI: https://doi.org/10.17721/AIT.2024.1.01

Maiia BOCHAROVA, PhD Student
ORCID ID: 0009-0004-3875-5019
e-mail: bocharova.maiia@gmail.com
Odesa I. Mechnikov National University

Eugene MALAKHOV, DSc (Engin.), Prof.
ORCID ID: 0000-0002-9314-6062
e-mail: eugene.malakhov@onu.edu.ua
Odesa I. Mechnikov National University

GENERAL-PURPOSE TEXT EMBEDDINGS LEARNING FOR UKRAINIAN LANGUAGE

Background. ***Learning high-quality text embeddings typically requires large corpuses of labeled data, which can be challenging to obtain for many languages and domains. This study proposes a novel adaptation of cross-lingual knowledge transfer that employs a cosine similarity-based loss calculation to enhance the alignment of learned representations.***

Methods. ***The impact of teacher model selection on the quality of learned text representations is investigated. Specifically, the correlation between cosine similarity scores among vectors of randomly selected sentences and the transferability of representations into another language is explored. Additionally, recognizing the need for effective evaluation methodologies and the limited availability of Ukrainian resources within existing benchmarks, a comprehensive general-purpose benchmark for assessing Ukrainian text representation learning is curated.***

Results. ***A cosine-similarity based loss calculation leads to 14.2% improvement in absolute Normalized Mutual Information (NMI) score compared to using mean squared error loss when distilling knowledge from the English language teacher model into Ukrainian student model. The findings demonstrate the strong correlation between the distributions of cosine similarities of the teacher model's representations of random sentences with the quality of learnt text embeddings. Pearson's correlation between "90th percentile of cosine similarity scores distribution" and "Average NMI score" is -0.96, which is a strong negative correlation.***

Conclusions. ***This research advances information theory in cross-lingual knowledge distillation, illustrating that cosine similarity-based loss functions are superior in performance. It underscores the importance of selecting the teacher model with wide distributions of cosine similarity scores. Furthermore, a pioneering broad-scale benchmark, covering five distinct domains for Ukrainian text representation learning is introduced. The source code, pretrained model, and the newly created Ukrainian text embeddings benchmark are publicly available at https://github.com/maiiabocharova/UkrTEB.***

Keywords: ***Natural Language Processing, text embeddings, Deep Learning, Data Mining, multilingual language models, knowledge transfer, domain adaptation.***

Background

There is an ever-growing interest in text representation learning, since it offers an attractive method to not only explore large text collections and find groups of semantically similar documents (Filatov, & Kovalanko, 2020), recommend relevant documents to users, but also allows to augment the input to Large Language Models (LLMs) (Binder, & Mezhuyev, 2024), such as Retrieval Augmented Generation frameworks (RAGs), enhancing their ability to generate relevant responses (Guu et al, 2020).

Models for creating text representations in vector space like SBERT (Reimers, & Gurevych, 2019), E5 (Wang et al., 2022), and GTE (Li et al., 2023) have demonstrated the strong capabilities of specialized transformer-based models to learn text embeddings of excellence in the English language, exploiting the vast availability of high-quality datasets specifically curated for semantic similarity learning, among which NLI (Bowman et al., 2015) and SNLI (Williams et al., 2018).

Recently, a lot of researchers have been focusing on unsupervised text representation learning, using contrastive learning and diverse augmentations to construct positive text pairs. However, while not requiring labeled data, those approaches tend to perform worse than models trained with supervision (Wang, Reimers, & Gurevych, 2021). This means that learning high-quality text embedding usually requires large corpuses of labeled data, which for low-resource languages are not available. Taking into account availability of a large number of strong models capable of producing high-quality text embeddings for English language, along with the translated text-pairs, one possible solution is to distill the knowledge of the teacher English model into a student model which will work for a low-resource language. Specifically, cross-lingual alignment of representations allows to transfer knowledge learnt by the model in one (source) language to another (target) language without the need for specific data annotation in the target language.

Using neural language models to produce embeddings for getting text representation is becoming more and more popular, replacing classical "bag-of-words" approaches. The vast availability of frameworks for training and inferencing neural embeddings models only contributes to the trend. Among such frameworks SentenceTransformers library offers a vast collection of pretrained transformer-based models and techniques to either train from scratch or fine-tune existing text embedding models.

Benchmarks are especially useful in comparing the benefits of training the models and assessing how beneficial it is to use one or another data source or approach for training.

The Massive Text Embeddings Benchmark (MTEB) (Muennighoff et al., 2022) offers a wide range of unified tasks (e.g. summarization, reranking, clustering, pair classification, retrieval etc) and domains (news, scientific publications, reviews, comments in social media etc) for assessing embeddings in different languages. However, MTEB mainly focuses on English datasets (97 out of 185 available datasets in it contain English texts) and Chinese (43). Only 19 languages occur more than 4 times. Ukrainian, among other languages is very underrepresented, being part of only 2 datasets which are both part of the bitext mining task – namely test set of the "Tatoeba" (Tiedemann, 2020) dataset, which is a collection of crowdsourced by volunteers sentences and their corresponding translations; and Flores (Goyal, 2022), which is a benchmark dataset compiled by Meta for machine translation between English and low-resource languages.

Considering the scarcity and specialized nature of datasets available for Ukrainian language, which are focused on translation tasks, there is a necessity to establish a benchmark that could help to accurately evaluate the performance of text embeddings in Ukrainian language across various domains.

ISSN 2788-6603

Aim formulation. The improvement of the information theory for cross-lingual knowledge distillation in text representation learning for low-resource languages on the example of Ukrainian language, leveraging English – target language translated pairs and state-of-the-art English teacher model.

Introduction of a new benchmark for assessing performance of text representation learning models for Ukrainian language, containing diverse domains.

Related Works. *Benchmarks.* The practice of web crawling and data mining to collect categorized data is known to play an important role in creating the benchmarks. As such, among many others, governmental job boards were scraped to create title normalization benchmarks (Decorte et al., 2021), news documents were extracted (Lang, 1995) to make clustering and classification dataset; and categorized question titles from StackExchange were extracted (Geigle et al., 2021).

MTEB (Muennighoff et al., 2022], having unified a substantial number of different datasets into a one benchmark, has established itself as the gold-standard for evaluating English models. Furthermore, efforts for extending MTEB to some other languages have been undertaken. As such, among others, C-Pack (Xiao et al., 2023) – a package of resources for benchmarking Language embedding for Chinese language, German Text Embeddings clustering benchmark (Wehrli, , Arnrich, & Irrgang, 2023) for assessing capabilities of models for grouping German texts and Spanish Evaluation benchmark (Araujo et al., 2022) for measuring performance of Spanish language embedding models were introduced.

The overall aim of such benchmarks is to allow a fair and reproducible assessment between models.

Text embeddings models. With the introduction of BERT (Kenton, & Toutanova, 2019), the new era of neural textual representations has begun, setting new state-of-the-arts on different text processing benchmarks. However, raw BERT representations are not suitable for providing efficient text-level embeddings, due to its training objectives (Reimers, Gurevych, 2019). To overcome this limitation, a number of approaches were proposed for adapting BERT embeddings to make them representative on sentence or paragraph-level. As such Universal Sentence Encoder (Cer et al., 2018) and Sentence BERT (Reimers Reimers, Gurevych, 2019) models were introduced, trained with supervision leveraging large human-labeled datasets (Bowman et al., 2015; Williams et al., 2018).

The growing interest towards multilingual text embedding learning and bitext mining for training Neural Machine Translation (NMT) systems has led to a large number of proposed models and architectures (Artetxe, & Schwenk, 2019; Heffernan, Çelebi, & Schwenk,2022) for language-agnostic representation learning. Most of the works focus on parallel text pairs, authors of EASE (Nishikawa et al., 2022) explore the entity linking in wikipedia to construct pairs of related texts. Authors of (Feng et al., 2020; Artetxe et al., 2019) prove the beneficial effects of such learning for the languages with very limited resources, originating from the fact that during multilingual training the model might have seen similar languages.

However, as is widely known (Xu et al., 2023), training a multilingual model presents a challenge of language interference, and, when having sufficient monolingual data, specialized monolingual models tend to outperform their multi-lingual same-sized counterparts.

And, to the best of our knowledge, no specialized models for creating sentence embeddings for Ukrainian language have been released. This can be attributed to several implications, such as lack of well-prepared datasets for training and benchmarks to evaluate the capabilities of the models.

To address the challenge described above we create a specialized benchmark for Ukrainian text embeddings learning – UkrTEB, as well as study the available resources for learning Ukrainian text representations and present a novel approach for training general-purpose text embeddings model for Ukrainian language (https://github.com/maiiabocharova/UkrMTEB). Our method leverages recent advancements in transfer learning techniques to overcome the challenges posed by the limited availability of labeled data and computational resources.

Methods

Dataset collection. Following the MTEB's design for Ukrainian texts, aim is to ensure high diversity of collected data, allowing to simulate a broad range of real-world applications. To address this problem of diversity, different domains offering both formal and informal texts should be considered. Furthermore, the usage of the data of permissive nature should be ensured.

In the context of the average sample length diversity across various datasets, MTEB typically incorporates two distinct versions of a dataset originating from a single data source. These versions are tailored to accommodate differences in text length, with one version focusing on sentence-to-sentence comparisons for shorter texts and another version geared towards paragraph-to-paragraph comparisons intended for longer texts. Specifically, taking as an example data gathered from StackExchange, two datasets are compiled. The first dataset presents the headlines of questions, whereas the second dataset includes both the headline and its accompanying description to the models for analysis.

1. Ukrainian government legislation board

The first data source chosen for analysis is the Ukrainian government legislation board. This entails the extraction of draft law titles along with their associated categories. Nineteen of the most frequently occurring categories were selected for the benchmark, including those related to committees addressing "правоохоронна діяльність" (law enforcement activities), "фінанси, податкова та митна політика" (finance, tax, and customs policy), among others. This selection process yielded a corpus comprising 9,678 samples, with an average length of 26.1 words per sample.

2. Book titles with their descriptions

The second data source encompasses book titles, accompanied by their blurbs and corresponding categories, amounting to a total of 12,590 samples across 53 distinct categories. This dataset is distinct from the first, focusing on literary works spanning a wide array of categories. These entries encompass a broad spectrum of themes and subjects, ranging from technical domains such as "Комп'ютерна література" (computer literature) to more nuanced explorations of human psychology and relationships in "Психологія і взаємини" (psychology and relationships), as well as imaginative and fantastical narratives designed for children. Each sample within this dataset comprises an average of 126.8 words, making it possible to test the models' capabilities for handling longer paragraphs.

3. Petitions to the ukrainian government

The third source of data is a website dedicated to hosting petitions. This platform serves as a digital space where individuals can create and sign petitions on various social, political, and environmental issues. Unlike the preceding datasets focused on legislative documents and literary works, this source captures public sentiment and activism, providing insights into societal

concerns and advocacy efforts. This dataset encompasses the collection of 3,387 samples distributed across 18 distinct categories. Through the aggregation of petitions, this dataset offers a diverse array of topics, reflecting the multitude of interests and causes championed by online communities. Analyzing this data enables researchers to examine trends in public opinion, track emerging issues, and assess the efficacy of grassroots activism in shaping public discourse and policy agendas.

4. UkrQAForum

The fourth data source selected for analysis is a Ukrainian community question answering forum, chosen to capture authentic conversational language usage and cover a diverse array of topics. This dataset aims to provide insights into real-world communication patterns, reflecting the colloquial language and informal discourse commonly found in online community forums. Since users post questions not only in Ukrainian, but also in other slavic languages, the FastText language model is used to filter out non-Ukrainian language data. The dataset contains 86 distinct categories, among which "Стиль, Мода, Бренди > Бренди" (Style, fashion, brands > Brands), "Сім'я, Дім, Діти > Домашні тварини" (Family, home, children > House Pets) etc.

Like real-world texts there are grammatical mistakes, typos and surzhyk (mixture of Ukrainian and other languages, spoken in the region)

5. UkrNews

The dataset collected for the benchmark consists of the Ukrainian news articles sourced from the biggest Ukrainian news outlet. This addition to the data sources aims to provide a comprehensive view of language usage across journalism and media domains. News articles offer a distinct perspective on language usage, characterized by formal structures, journalistic conventions, and specialized terminology.

This dataset contains 23 categories, among which "війна" (war), "бокс" (boxing).

Some examples of the data samples are shown in Table 1.

Table 1

Example samples from the collected datasets

Data source	Sample	Category of the sample
Ukrainian government legislation board	Проєкт Закону про внесення змін до Закону України "Про адвокатуру та адвокатську діяльність" щодо вдосконалення окремих питань організації адвокатської діяльності	Правова політика
Book titles with their descriptions	"Легкі й швидкі рецепти неперевершених десертів для святкового дня і просто гарного настрою! Ніжний торт, ароматний лимонний чізкейк, рум'яні кексики, пиріжки, еклери… Понад 50 рецептів з оригінальними фотографіями надихнуть вас на солодкі експерименти на кухні!"	Кулінарія. Їжа та напої
Petitions to the ukrainian government	"Про порядок повірок квартирних/будинкових лічильників води"	Комунальне господарство
QA Forum	"Київ, 2-кімнатна в новобудові – по чому?"	Бізнес, Фінанси. Нерухомість
News	"День фізичної культури і спорту в Україні: листівки та побажання"	Свята

The aggregated statistics of the benchmark datasets are shown in Table 2.

Table 2

Statistics of the collected datasets

Dataset Name	Size per split	Classes per split	Number of categories	Number of samples	Average sample length (chars)	Average word lengths
UkrLawDrafts	9,687	19	19	9,687	186.8	26.1
UrkBookBlurs	1250 to 1693	11 to 15	48	16,678	748.4	123.7
UkrPetitions (s2s)	3,474	18	18	3,474	75.75	10.8
UkrPetitions (p2p)	3,474	18	18	3,474	1332	197
UkrQAForum (s2s)	4380 to 8475	11 to 15	86	155,788	42,3	7.9
UkrQAForum (p2p)	4380 to 8375	11 to 15	86	155,781	311	56.9
UkrNews (s2s)	8,159 to 8,457	15	23	74,746	72.3	12.3
UkrNews (p2p)	8,159 to 8,457	15	23	74,746	152.4	25.4

Where s2s means that only headline sentences are taken as samples, and in the p2p scenario respectively the concatenated headline and description paragraphs are used.

Training ukrainian sentence embeddings model. Vast amounts of parallel text chunks which offer datasets like OpenSubtitles (Lison, & Tiedemann, 2021) or wikimatrix (Schwenk et al., 2019) have been collected and made publicly available.

ISSN 2788-6603

A number of methodologies for learning multilingual sentence representation models were proposed, from which two main families of approaches for training embedding models leveraging translated text pairs can be identified.

The first one consists in training a dual encoder model using translation ranking loss with inbatch negative samples (Feng et al., 2020). This training task consists of aligning the embeddings of the source and target languages in a shared space. Models trained with this approach learn to project the embeddings of the same sentence in different languages close to each other, while putting the representation of other sentences which are not direct translation of each other further apart. However, as has been shown by multiple studies (Wang et al., 2022) that, while the strategy of using inbatch negative samples is quite efficient, it requires substantial batch sizes to reach optimal performance, and bigger batch sizes tend to be computationally intensive. The second drawback of such a training strategy lies in the fact that models trained in such a way tend to struggle with estimating the similarity of sentences which are not direct translations of each other (Reimers, & Gurevych, 2020).

The second approach involves utilizing a teacher model proficient in representing sentences in a language with abundant resources. The student model is then trained to emulate the representations generated by the teacher model for the target language text and produce embeddings closely aligned with those of the teacher model in the vector space (Reimers, Gurevych, 2020]. This approach mitigates the influence of batch size on the training process, ensuring that the quality of embeddings learned by the student model depends primarily on the abilities of the teacher's models and the size of the dataset, as well as its relevance to the texts which the student model will encounter while solving downstream tasks.

However, models trained for a specific language tend to outperform multilingual models of the same size.

In LASER3 (Heffernan et al., Tiedemann 2022), for example, researchers propose to use a language family specific model to both benefit from increased numbers of samples and transfer additional knowledge from other similar languages.

Bitext data aligned with rich-resource language (English in our case) is required to train the student Ukrainian text representation model. Training data is collected from the Opus website (Tiedemann, 2012] and consists of a combination of several corpuses. Since the inconsistent quality of the data and language contamination, language deduplication and heuristics-based cleaning is applied.

Specifically texts are cleaned using the following steps:

1. Deduplication.
2. Language identification and filtering of samples which have non-Ukrainian target texts.
3. Filtering out the samples that differ in length by more than two times.

The statistics of the combined dataset are provided in Table 3.

Table 3

Statistics of the bitext datasets used for training the model

Data Source	Number of parallel sentences	Number of unique sentences after filtering
OpenSubtitles2018	877,780	419,004
Wikimedia	757,910	617,809
SciPar_Ukraine	306,813	306,813
QED	215,530	198,100
TED2020	208,141	198,876
Tatoeba	175,502	168,480
ELRC-5179-acts_Ukrainian	129,942	126,361
ELRC-5180-Official_Parliament_Ukraine_Ukrainian_laws_EN	116,260	115,494
ELRC-5181-Official_Parliament_Ukraine_abstracts_UK_laws	61,012	60,885
ELRC-5174-French_Polish_Ukraine	36,228	35,597
Total	2,885,118	2,264,340 Globally unique: 2,202,117

After careful consideration we opt out from using the automatically mined sentence pairs like WikiMatrix dataset, because upon manual inspection it is discovered that such datasets are of very low quality for English-Ukrainian language pairs (e.g. "And they ask you what they should spend" is a pair of "Потім запитали: чи не ти пророк?", "How are you feeling?" with "Який спосіб запропонували б ви?" etc), which underscores the lack of well-performing models for ukrainian language.

After final deduplication the obtained dataset contains 2,202k bitexts.

In (Reimers, & Gurevych, 2020) authors propose to use Mean Squared Error (MSE) as a loss function. MSE is a commonly used loss function in regression tasks and measures the average squared difference between the predicted and actual values. In the context of text representation learning, MSE compares the embeddings produced by the student model with those generated by the teacher model, aiming to minimize the discrepancy between the two sets of embeddings. However, training with MSE loss can lead to unstable training and unexpected results, especially when the embeddings of the teacher model are normalized and MSE values are small. Moreover, cosine similarity is used for the final measurement of similarities when handling downstream tasks.

Cosine similarity measures the cosine of the angle between two vectors and is particularly well-suited for comparing the similarity between vectors regardless of their magnitudes. In the context of text embeddings, cosine similarity quantifies the degree of similarity between two text representations based on the direction of their vectors in the embedding space, rather than their absolute distances and as such is more robust in capturing semantic similarity between text representations.

Taking the above into account, we adapt the M-SBERT (Reimers et al., 2020) training approach to use cosine similarity loss for training.

More formally, loss is defined as follows:

$$\text{Loss} = (1 - \cos(emb_1, emb_2))^2 , \quad (1)$$

where emb_1 – is embedding of the English sentence, produced by teacher model; emb_2 – is embedding of the Ukrainian sentence, produced by student model; cos – is a cosine similarity between two vectors.

To save computational resources we initialize the weights of the ukrainian language model from the RoBERTa-base model trained on ukrainian text data (ukr-roberta-base, 2024).

AdamW optimizer with a linear warmup for 10% of steps is used. Learning rate is set to 2e-5.

Results

For evaluation three publicly available models supporting Ukrainian language are taken. As evaluation metric, the Normalized Mutual Information (NMI) score, which is a harmonic mean of homogeneity and completeness, is used. This score can be in range 0 to 1, where 1 means a perfectly complete labeling of samples. This metric is not dependent on the absolute values of the labels and permuting the class or cluster label values does not affect the score.

The evaluation results are presented in Table 4.

Table 4

Evaluation results of NMI score

Model	UkrLawDrafts	UrkBookBlurs	UkrPetitions		UkrQAForum		UkrNews		Average
LaBSE	0,175	0,513	0,138	0,249	0,188	0,396	0,559	0,585	0,350
distiluse-base-multilingual-cased	0,138	0,410	0,162	0,222	0,144	0,326	0,416	0,485	0,289
E5	0,201	0,481	0,205	0,274	0,300	0,459	0,627	0,644	0,40
Ukr-distil_mse	0,145	0,329	0,122	0,121	0,133	0,207	0,517	0,602	0,272
Ukr-distil_cos	0,234	0,542	0,246	0,326	0,326	0,453	0,568	0,621	0,414

As can be seen from the results, there is a clear advantage of using cosine similarity loss over 0,272 average score, which accounts mean squared error loss (0,414 over to 14.2% absolute NMI score score improvement).

Dependence of quality of learn text embeddings on teacher model. Experiments using different teacher models were conducted. Below in fig. 1 the distributions of the cosine similarities calculated between all possible pairs of vectors from 5,000 random English sentences from the training set are visualized. As can be seen, E5 model shows a very narrow distribution.

The plot (Fig. 1) shows cumulative distribution of cosine similarity scores, where the x-axis represents the cosine similarity threshold, and the y-axis shows the percentage of records that have a similarity score less than that threshold

Correlation between cosine similarity scores distribution of the teacher model and performance on the downstream tasks of the student model distilled from it is shown in table 5.

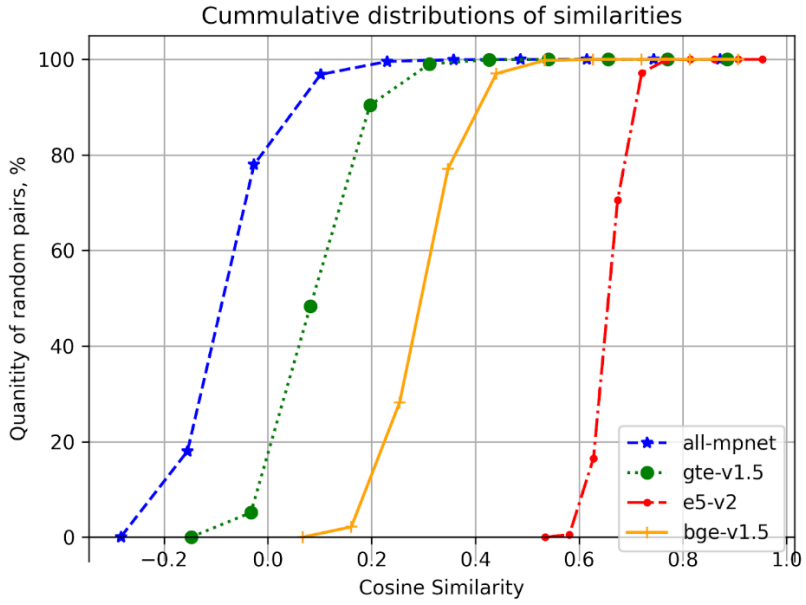

Fig. 1. Cumulative distributions of similarity scores between random sentence pairs

ISSN 2788-6603

Table 5

Correlation between cosine similarity distribution scores and knowledge transferability

Teacher model	UkrLawDrafts	UrkBookBlurs	UkrPetitions		UkrQAForum		UkrNews		Average score	90th percentile of cosine similarity
e5 base	0,062	0,245	0,054	0,045	0,065	0,127	0,308	0,414	0,165	0,75
bge base	0,113	0,342	0,124	0,112	0,133	0,215	0,516	0,601	0,269	0,48
gte base	0,128	0,364	0,105	0,199	0,132	0,263	0,448	0,548	0,273	0,31
all-mpnet	0,234	0,542	0,246	0,326	0,326	0,453	0,568	0,621	0,414	0,16

As can be seen from Table 5, there exists a strong correlation between the distribution of cosine similarities between vectors of unrelated sentences and quality of the sentence embeddings learnt by student model. Pearson's correlation between "90th percentile of cosine similarity" and "Average score" is – 0.96, which is interpreted as a strong negative correlation.

The lower similarity between unrelated vectors leads to more stable training and learning by the student model of more distinct sentence embeddings, which in turn enhances the model's ability to accurately capture the semantic nuances of the text.

This results in improved performance on downstream tasks, as the embeddings are able to better represent the underlying information and differentiate between various content types. Therefore, selecting a teacher model with a lower average cosine similarity score is crucial for effective knowledge transfer and the overall success of the student model.

Discussion and conclusion

In this paper information theory for cross-lingual knowledge distillation obtained further development. It has been shown that using cosine similarity-based loss function leads to significant improvements (14.2% absolute NMI score improvement) compared to using the mean squared loss function when distilling knowledge from the well-trained model.

The influence of the teacher model selection and correlation between the cosine similarity distribution which the teacher model produces was proved to influence the quality of learn embeddings of the student model. It has been established that there is a strong negative correlation between "90th percentile of cosine similarity scores distribution" and "Average NMI score" obtained on clustering benchmark.

A new and first of its kind benchmark for Ukrainian text representation learning has been introduced, covering 5 distinct domains.

The model and Ukrainian text embeddings benchmark are freely available at https://github.com/maiiabocharova/UkrMTEB.

Authors' contribution. Maiia Bocharova – literature overview, development of methods and methodologies of the research, empirical data collection, analysis of results and conclusions. Eugene Malakhov – consultation, ideas and guidance.

References

Abdelali, A., Guzman, F., Sajjad, H., & Vogel, S. (2014). The AMARA Corpus: Building parallel language resources for the educational domain. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.).*The Proceedings of the 9th International Conference on Language Resources and Evaluation* (pp. 1856–1862)/ In Lrec.

Araujo, V., Carvallo, A., Kundu, S., Cañete, J., Mendoza, M., Mercer, R. E., & Soto, A. (2022). Evaluation Benchmarks for Spanish Sentence Representations. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.). *In Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6024–6034), Marseille, France. European Language Resources Association.

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. In L. Lee, M. Johnson, B. Roark, A. Nenkova (Eds.). *Transactions of the Association for Computational Linguistics*, 7, 597–610. https://doi.org/10.1162/tacl_a_00288

Binder, M., & Mezhuyev, V. (2024). *A framework for creating an IoT system specification with ChatGPT*. Internet of Things, 27, 101218. Institute of Industrial Management, University of Applied Sciences FH JOANNEUM, Austria. https://doi.org/10.1016/j.iot.2024.101218

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, J. Su (Eds.). *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 632–642). https://doi.org/10.18653/v1/D15-1075

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., & Kurzweil, R. (2018). Universal sentence encoder. In E. Blanco, W. Lu (Eds.). *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 169–174). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2029

Decorte, J. J., Van Hautte, J., Demeester, T., & Develder, C. (2021). Jobbert: Understanding job titles through skills. *In International workshop on Fair, Effective And Sustainable Talent management using data science (FEAST)* (pp. 1–9). As part of ECML-PKDD.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. In S. Muresan, P. Nakov, A. Villavicencio (Eds.). *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 1 (Long Papers)* (pp. 878–891). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62

Filatov, V., & Kovalenko, A. (2020). Fuzzy systems in data mining tasks. *In Advances in Spatio – Temporal Segmentation of Visual Data* (pp. 243–274). Springer. https://doi.org/10.1007/978-3-030-35480-0_6

Geigle, G., Reimers, N., Rücklé, A., & Gurevych, I. (2021). *TWEAC: transformer with extendable QA agent classifiers*. https://doi.org/10.48550/arXiv.2104.07081

Goyal, N., Gao, C., Chaudhary, V., Chen, P. J., Wenzek, G., Ju, D., & Fan, A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. In B. Roark, A. Nenkova (Eds.). *Transactions of the Association for Computational Linguistics,* 10, 522–538. https://doi.org/10.1162/tacl_a_00474

Grabar, N., & Hamon, T. (2017). Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. *In Computational linguistics and intelligent systems (COLINS 2017)*. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. *In International conference on machine learning* (pp. 3929–3938). JMLR.org.

Heffernan, K., Çelebi, O., & Schwenk, H. (2022). Bitext mining using distilled sentence representations for low-resource languages. In Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.). *Findings of the Association for Computational Linguistics: EMNLP* (pp. 2101–2112), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.154

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, T. Solorio (Eds.). *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1 (Long and Short Papers)* (pp. 4171–4186). https://doi.org/ 10.18653/v1/N19-1423

Lang, K. (1995). Newsweeder: Learning to filter netnews. *Proceedings of the 12th International Conference on Machine Learning* (pp. 331–339).

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning. https://doi.org/10.48550/arXiv.2308.03281

Lison, P., & Tiedemann, J. (2016). *Opensubtitles2016:* Extracting large parallel corpora from movie and tv subtitles. N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.). *Proceedings of the 10th International Conference on Language Resources and Evaluation* (pp. 923–929). In Lrec.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive text embedding benchmark. In A. Vlachos, I. Augenstein (Eds.). *In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2014–2037). Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.148

Nishikawa, S., Ri, R., Yamada, I., Tsuruoka, Y., & Echizen, I. (2022). EASE: Entity-aware contrastive learning of sentence embedding. In M. Carpuat, M. Marneffe, I. Ruiz (Eds.). *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3870–3885). Association for Computational Linguistics.

Reimers, N. & Gurevych, I. (2019). SentenceBERT: Sentence embeddings using siamese BERT networks. In K. Inui, J. Jiang, V. Ng, X. Wan (Eds.). *Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Hong Kong: China (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410.

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In B. Webber, T. Cohn, Y. He, Y. Liu (Eds.). *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4512–4525), online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1351–1361). Association for Computational Linguistics.

Schwenk, H., Wenzek, G., Edunov, S., Grave, E., & Joulin, A. (2021). CCMatrix: Mining billions of high-quality parallel sentences on the web. In P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.).*In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 1 (Long Papers)* (pp. 6490–6500), online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.115

Tiedemann, J. (2012). *Parallel data, tools and interfaces in OPUS* (pp. 2214–2218). In Lrec.

Tiedemann, J. (2020*).* The Tatoeba Translation Challenge--Realistic Data Sets for Low Resource and Multilingual MT. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, Ba. Haddow, M. Huck, A. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri (Eds.). *In Proceedings of the Fifth Conference on Machine Translation* (pp. 1174–1182). Association for Computational Linguistics.

"ukr-roberta-base" (11, August, 2024). https://huggingface.co/youscan/ukr-roberta-base

Wang, K., Reimers, N., & Gurevych, I. (2021). TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. M. Moens, X. Huang, L. Specia, S. Yih (Eds.). *In Findings of the Association for Computational Linguistics: EMNLP*. Punta Cana, Dominican Republic (pp. 671–688). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.59

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., & Wei, F. (2022). *Text embeddings by weakly-supervised contrastive pre-training*. https://doi.org/10.48550/arXiv.2308.03281

Wehrli, S., Arnrich, B., & Irrgang, C. (2023). German Text Embedding Clustering Benchmark. In M. Georges, A. Herygers, A. Friedrich, B. Roth (Eds.). *In Proceedings of the 19th Conference on Natural Language Processing.* Ingolstadt, Germany (pp. 187–201). Association for Computational Linguistics.

Williams, A., Nangia, N., & Bowman, S. R. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In M. Walker, H. Ji, A. Stent (Eds.). *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1 (Long Papers)* (pp. 1112–1122). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Xiao, S., Liu, Z., Zhang, P., & Muennighof, N. (2023). *C-pack: Packaged resources to advance general chinese embedding*. https://doi.org/10.48550/arXiv.2309.07597

Xu, H., Tan, W., Li, S. S., Chen, Y., Van Durme, B., Koehn, P., & Murray, K. (2023). Condensing Multilingual Knowledge with Lightweight Language-Specific Modules. In H. Bouamor, J. Pino, K. Bali (Eds.). *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 1575–1587). https://doi.org/10.18653/v1/2023.emnlp-main.97

Отримано редакцією журналу / Received: 08.09.24
Прорецензовано / Revised: 14.10.24
Схвалено до друку / Accepted: 23.10.24

Майя БОЧАРОВА, асп.
ORCID ID: 0009-0004-3875-5019
bocharova.maiia@gmail.com
Одеський національний університет імені І. І. Мечникова

Євгеній МАЛАХОВ, д-р техн. наук, проф.
ORCID ID: 0000-0002-9314-6062
eugene.malakhov@onu.edu.ua
Одеський національний університет імені І. І. Мечникова

ТРЕНУВАННЯ ТЕКСТОВИХ ВКЛАДЕНЬ ЗАГАЛЬНОГО ПРИЗНАЧЕННЯ ДЛЯ УКРАЇНСЬКОЇ МОВИ

В с т у п . ***Тренування високоякісних текстових вкладень зазвичай вимагає великих корпусів з анотованими даними, які може бути складно отримати для більшості мов і доменів. У цьому дослідженні запропоновано нову адаптацію крос-лінгвістичного перенесення знань, яка використовує обчислення втрат на основі косинусної подібності між перекладами текстів для кращого зіставлення отриманих векторних представлень текстів.***

М е т о д и . ***Досліджено вплив функцій втрат, а також вибору моделі вчителя на якість вивчених текстових репрезентацій. Крім того, досліджено кореляцію між розподілом косинусної подібності між векторами випадково вибраних речень моделі-вчителя та можливістю перенесення репрезентацій на іншу мову. З огляду на потребу в ефективних методологіях оцінювання й обмежену доступність ресурсів для української мови в межах існуючих бенчмарків, розроблено комплексний універсальний бенчмарк для оцінювання представлень тексту для української мови.***

Р е з у л ь т а т и . ***Обчислення втрат на основі косинусної подібності приводить до покращення абсолютного показника нормалізованої взаємної інформації (NMI) на 14,2 % порівняно з використанням середньоквадратичної функції втрат під час перенесення знань із моделі-вчителя англійської мови на українську модель-учня. Отримані результати демонструють сильну кореляцію між розподілом косинусної подібності векторів не пов'язаних між собою речень, які векторизуються моделлю-вчителем, та якістю засвоєних текстових вкладень. Кореляція Пірсона між "90-м процентилем розподілу оцінок косинусної подібності" та "середнім показником NMI" становить –0,96, що є сильним негативним зв'язком.***

В и с н о в к и . ***Це дослідження розвиває теорію інформації в галузі крос-лінгвістичної дистиляції знань, показуючи, що функції втрат на основі косинусної подібності є кращими за своїми характеристиками. Підкреслено важливість вибору моделі-вчителя із широким розподілом коефіцієнта косинусної подібності. Представлено новий широкомасштабний бенчмарк, що охоплює п'ять різних доменів для навчання представлення українського тексту. Код, попередньо навчена модель і новостворений бенчмарк для української мови опубліковано за посиланням https://github.com/maiiabocharova/UkrTEB.***

К л ю ч о в і с л о в а : ***оброблення природної мови, текстові вкладення, глибоке навчання, видобування даних, багатомовні мовні моделі, перенесення знань, адаптація до домену.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

ISSN 2788-6603