UDC 004.94 **doi: 10.20998/2522-9052.2020.3.13**

Serhii Olizarenko[1], Vladimir Argunov[2]

[1] Kharkiv National University of Radio Electronics University, Kharkiv, Ukraine
[2] HIPSTO, Kharkiv, Ukraine

RESEARCH ON THE SPECIFIC FEATURES OF DETERMINING THE SEMANTIC SIMILARITY OF ARBITRARY-LENGTH TEXT CONTENT USING MULTILINGUAL TRANSFORMER-BASED MODELS

Abstract. The possibilities of determining the semantic similarity of multilingual arbitrary-length text content have been investigated using their vector representations obtained within different multilingual models based on Transformer architecture. A comparative analysis of the Transformers has been performed to select the most advantageous model for this class of problems. Also, two new unique approaches to determining the semantic similarity of a multilingual text content have been developed to be used in the HIPSTO Open AI Information Discovery Platform, the challenge being to allow arbitrary text length. Experimental and research evidence is offered to support the new approaches as a solution to the semantic similarity problem.

Keywords: Natural Language Processing; BERT; semantic similarities; news content; Deep Learning; multilingual text content; vector representation; Transformers; fine-tuning.

Introduction

The determination of the semantic similarity of texts (sentences) is one of the fundamentally important problems in the field of Natural Language Processing (NLP). In particular, the issue of semantic similarity is of paramount importance in systems of automatic search, retrieval and analysis of multilingual contents available on various news sites. In this case, the solution to the problem lies in a numerical assessment of the similarity of sequences corresponding to the multilingual text content.

The problem of determining the semantic similarity in the process of searching and analyzing a multilingual text content can be solved efficiently by using pre-trained multilingual Transformer-based models adapted through fine-tuning within the Deep Learning methodology [1].

Below the following models supporting multiple languages and based on Transformer architecture are analyzed:

- The multilingual BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding) supporting 104 languages [2];
- The multilingual DistilBERT (104 languages), which is a simplified BERT version operating 60% faster and retaining over 95% of the BERT characteristics measured in the GLUE (General Language Understanding Evaluation) test [3];
- The multilingual XLM model for 100 languages [4];

The comparative analysis of the above Transformers (see below), ranks these models as highly efficient tools for solving the problem of determining the semantic similarity. All of them, however, suffer a severe limitation on the length of an incoming tokenized text sequence – they can recognize no more than 512 tokens. The main techniques used to solve the problem are based on:

1) the cutting-off methods based on choosing the first or the final fragments of the sequence or combining the first and the final fragments of the sequence, in both cases the length of the resulting sequence being no more than 512 tokens;

2) the hierarchical methods (e.g. combining the hidden states of all fragments of the sequence) [5].

However, these methods can lead to a loss of the contextual dependence of the most important words (hence, phrases and sentences) in the text sequences, which can in turn significantly affect the quality of determination of the semantic similarity of the texts being analyzed. So, there is a problem with applying the above Transformers to texts longer than 512 tokens and saving the utmost contextual dependence of the most significant words in text sequences, which is necessary for efficient determination of the semantic similarity of an arbitrary-length text content.

To meet this challenge, two new approaches to determining the semantic similarity of an arbitrary-length text content have been proposed to be used within HIPSTO Open AI Information Discovery Platform setup:

1) preliminary machine learning-based generalization (automatic summarization) of arbitrary-length texts under comparison and the subsequent, direct determination of the semantic text similarity by solving the problem of classification of pairs of text sequences within the pre-trained and fine-tuned multilingual Transformer-based model;

2) using, on step 1, the contextualized vector representations of arbitrary-length texts obtained by sliding window method according to the “medium core” rule as an input for the pre-trained multilingual Transformer (without fine-tuning, as feature extraction problem), and on step 2 followed by determining the degree of the similarity of the content through measuring the distance between their vector representations using the selected similarity metric (for example, cosine similarity).

Related Investigations. It is appropriate to mention here some investigations containing data on Transformer-based solutions to the problems of semantic text similarity. The possibility of determining the semantic similarity of a text content within the

HIPSTO AI technology setup on the basis of sentence embeddings and the use of the first problem of the pre-trained multilingual BERT model (the problem of word masking) is described in [1]. In particular, it is found out that without fine-tuning, the best results can be obtained through the formation of word embeddings by concatenating the last 4 layers and then forming sentence embeddings using a special integrating layer.

The universal multilingual sentence encoder for semantic retrieval for 16 languages using the family of sentence embedding models of the universal sentence encoder (USE) was performed in [6, 7]. Those models appear to be implementations of CNN (Kim, 2014) and Transformers (Vaswani et al., 2017) architectures [8, 9].

A technique of multilingual similarity discovery based on bidirectional LSTM encoder using a LASER (Language-Agnostic Sentence Representations) preparation is proposed by Lee in Ref. [10, 11].

The experimental results obtained by different methods of fine-tuning of BERT model for text classification problems, including determining the semantic similarity of the text are reported by Chi Sun et al., 2019 [5]. The Sentence-BERT (SBERT) model which is a modification of the pre-trained BERT and uses Siamese and triplet neural network structures to obtain semantically sensible vector representations comparable by cosine similarity, is described by Nils Reimers et al. (2019) [12]. A semantics-oriented search system that uses BERT embeddings and additional neural networks for estimating similarity and subsequent arrangement of documents in the ascending – descending order of their importance is proposed by Manish Patel (2019) [13].

A search system using BERT embeddings and the cosine similarity to estimate a search query and document similarity is proposed in (X.Han, 2019) [14].

It is noteworthy that none of the above studies considers the problem of efficient processing of texts over 512 tokens long.

Generalized study design. The generalized research scheme of solving the problem of determining the semantic similarity of multilingual arbitrary-length text contents consists in performing the following main steps:

1) The problem is considered as a task of classifying pairs of text sequences using Transformer-based models (Strategy 1) or;

2) The problem is treated as a task of determining an arbitrary-length vector representation applying the sliding window principle and Transformer-based models with the subsequent formation and determination of the degree of similarity of the sentence embeddings (Strategy 2);

3) Conclusions are drawn through a generalized assessment of the results obtained and the formulation of further directions of the research.

The solution to the problem of determining the semantic similarity of a multilingual arbitrary-length text content. Strategy 1

Within strategy 1, the problem of determining the semantic similarity of a multilingual arbitrary-length text content is considered as a classification of pairs of text sequences. The applied research scheme includes the following main stages:

1) Specifying the architecture of the software module for determining the semantic similarity of multilingual arbitrary-length text contents (strategy 1);

2) Analysis of the possibilities of various types of multilingual Transformer-based models (BERT, DistilBERT and XLM) to determine the semantic similarity of multilingual text contents;

3) Evaluation of hyperparameters for fine-tuning of the model. Automatic tuning of the training rate for multilingual models taking into account the model type and the training data set;

4) Fine-tuning of multilingual Transformer-based models (BERT, DistilBERT, XLM) for solving the problem of classifying pairs of text sequences as well as the subsequent analysis of the fine-tuning results and selection of the most appropriate Transformer-based model for solving the problem of determining the semantic similarity of multilingual arbitrary-length text content within strategy 1;

5) Selection of the basic method of summarization of a text sequence over 512 tokens long to solve the problem of determining the semantic similarity of a multilingual arbitrary-length text content.

The program module for determining the semantic similarity of multilingual arbitrary-length text content within strategy 1 is shown schematically in Fig. 1. The software module consists of two main blocks:

- A block of summarization of a text sequence to no more than 254 tokens (for the BERT, DistilBERT and XLM models);
- A fine-tuned multilingual Transformer-based model for classification of pairs of sequences.

Three multilingual Transformer-based models (BERT, DistilBERT, and XLM) available in the Hugging Face library for TensorFlow [15] have been considered for determining the semantic similarity of a multilingual text content. The multilingual BERT (177,854,978 trainable params) in 104 languages is a bidirectional transformer pre-trained through combining a modeling target using a masked language (MLM) and the next sentence prediction (NSP) [2]. On training NSPs in BERT, a specialized token [CLS] was used as a sequence for evaluating the forecast results. In this study, this token (the first token in the sequence) is used to solve the problem of classifying pairs of text sequences in all Transformers. The multilingual DistilBERT model (135,326,210 trainable params) covering 104 languages is an enlightened version of BERT that operates 60% faster and retains more than 95% of the BERT characteristics measured in the GLUE (General Language Understanding Evaluation) test [3]. The multilingual XLM model (571,499,522 trainable params) in 100 languages is a pre-trained transformer aimed at the following objectives [4]: 1) Causal language modeling (CLM) (next token prediction); 2) Masked language modeling (MLM) (BERT-like); 3) Translation Language Modeling (TLM) an object (extension of BERT's MLM to multiple language inputs).

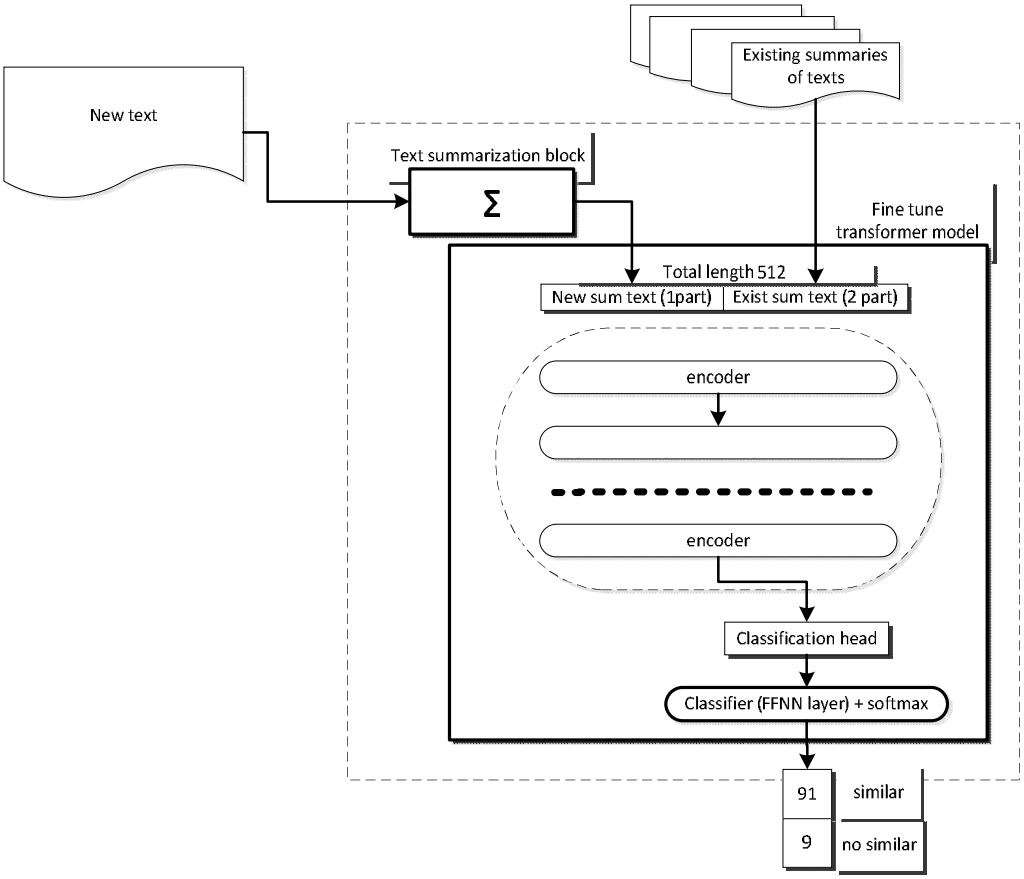


Fig. 1. Overall design of the module for determining the semantic similarity of multilingual arbitrary-length text content (strategy 1)

Thus, unlike BERT and DistilBERT, XLM is not trained specially for NSP. And yet, the input data format of the XLM model, as in BERT, affords encoding of two different sequences in the same input identifiers (token type ids). However, the input data format in DistilBERT does not have the token type ids and hence, does not indicate which token belongs to which segment of the text sequence. To solve this problem within DistilBERT, segments are simply separated using a special [SEP] token (as in BERT).

Neural networks contain a number of hyperparameters that must be set before training can be started. In practice, many of them have reasonable, by default, values, but some of them must be configured before training. The training rate is one of the most critical parameters since it determines the degree of adjusting the weights during training. In this study, we used the method of preliminary automatic tuning and adjustment of the learning rate to minimize more efficiently the losses in training [16]. The learning speed of a model was tuned automatically taking into account the type of multilingual model and the training data set. The training was carried out using Microsoft Research Paraphrase Corpus (MRPC) kit containing 5800 pairs of sentences [17]. The choice of this dataset is explained later in this paper when fine-tuning of the Transformer-based models is described.

The results of the automatic tuning of the learning rate of Transformers are presented in Fig. 2, 3. The learning rate values were chosen so that they correlate to still falling losses (before the losses diverge). The learning rates based on the analysis of the graphs in Fig. 4 are as follows:

1) learning_rate = 5e-5 for BERT;
2) learning_rate = 8e-5 for DistilBERT;
3) learning_rate = 5e-6 for XLM.

In this research, for the task of classifying pairs of text sequences, multilingual Transformer-based models take the final hidden state ***s***. As a function of activation of a fully connected classifier layer, *softmax* is used to predict the probability p for class label l [18]:

$$p(l \vee s) = softmax(W_S), \quad (1)$$

where W_S is the resulting tensor of the hidden state s.

Fine-tuning of multilingual Transformers (BERT, DistilBERT, XLM) aimed at solving the problem of classifying pairs of text sequences, was performed using the MRPC dataset. Detection of rephrasing is a task of examining two text objects and determining whether they have the same meaning. In the general case, to attain high accuracy in performing this task, thorough syntactic and semantic analyses of two text objects are required.

According to the style of rephrasing, paraphrases can be subdivided into five types [17]:

1) Trivial change of words in a phrase (sequence);
2) Replacement of a phrase;
3) Changing the order of phrases;
4) Separation and/or combination of phrases;
5) Complex paraphrase.

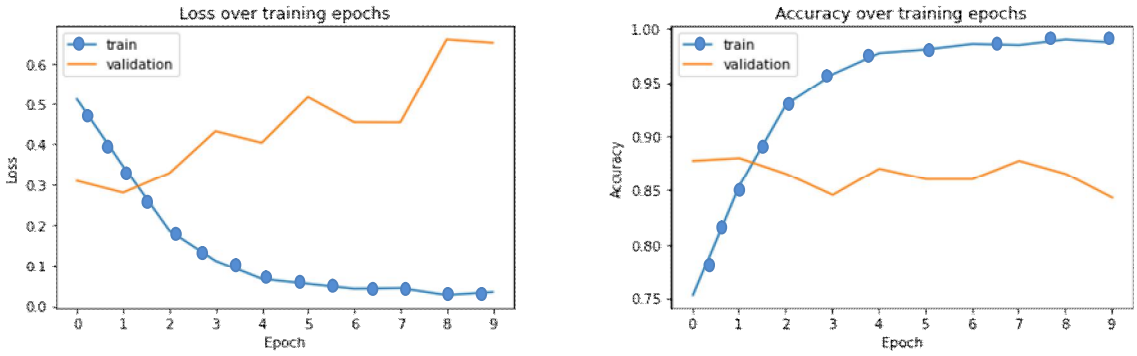


Fig. 2. Plots of error and accuracy of training for BERT

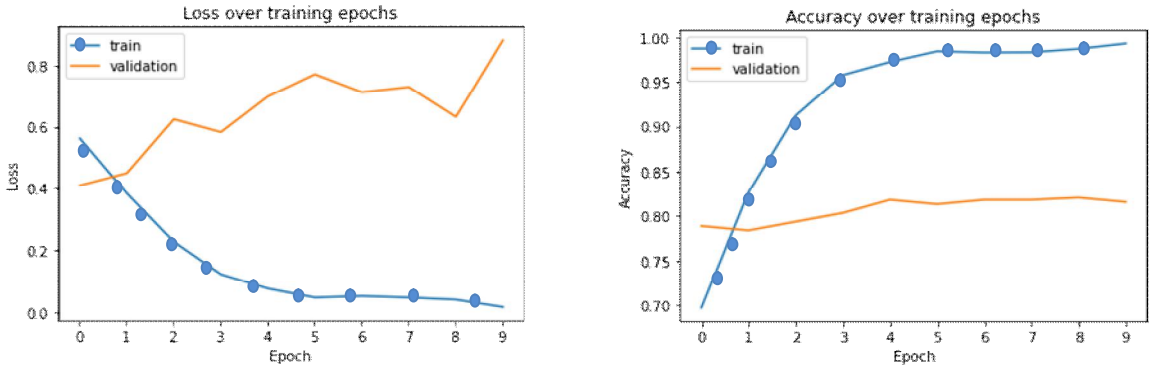
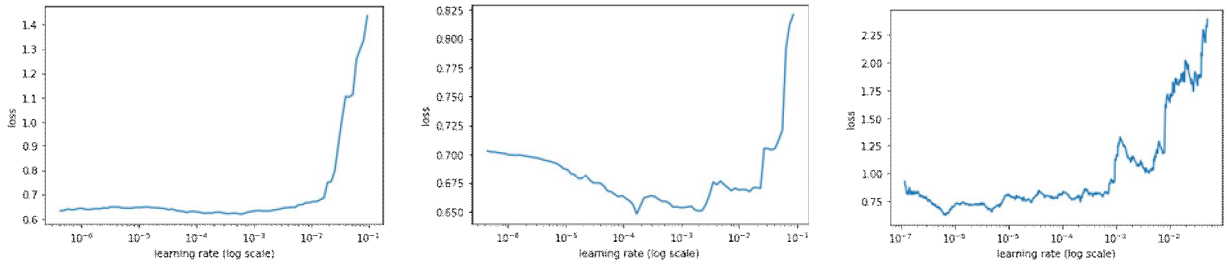


Fig. 3. Plots of error and accuracy of training for the DistilBERT model



a b c

Fig. 4. The results of automatic tuning of the training rate for: 1 - BERT, 2 - for DistilBERT, 3 - for XLM

In this study, using type #5 of "paraphrase" is of particular importance in determining the semantic similarity of the multilingual text content: on the one hand, the MRPC data set implements the corresponding type of paraphrases and, on the other hand, it allows checking the effectiveness of the results of fine-tuning the Transformers (BERT, DistilBERT, XLM), taking into account the NLP data available in literature [2 - 4].

The results of the fine-tuning of the multilingual Transformers (BERT, DistilBERT, XLM) are presented in Fig. 5–7 and Table 1. The analysis of the graphs in Fig. 3–5 shows that the values of loss validation and accuracy validation start moving apart even in the third epoch of training, which suggests that two epochs of training on an appropriate data set is quite sufficient for these models. The data in Table 1 also show that the multilingual BERT retrained on the MRPC dataset, exhibits the highest measures of accuracy among all multilingual Transformers fine-tuned to solve the problem of classifying pairs of text sequences of accuracy. At the same time, the accuracy measures for the DistilBERT and XLM models are very close even though the XLM model has four times more trained parameters than DistilBERT. Thus, with no limitation on hardware, the multilingual BERT is most efficient to solve the problem of determining the semantic similarity of multilingual arbitrary-length text content within strategy 1. If there are any restrictions in place, the multilingual DistiBERT is the best choice. The XLM model did not prove to be effective within strategy 1 as it was never pre-trained for a NSP task.

Generalization is a task of reducing the text to a shorter version, reducing the size of the source text and at the same time preserving the key information elements and the meaning of the content. In this study, the main task of summarization was to generalize a text sequence to one no longer than 254 tokens (for the BERT, DistilBERT, and XLM models) and use it then as one of the sequences of the input pair for the multilingual model of choice. On having analyzed several models of summarizers: LSA, KLS, LexRankS and TextRank were short-listed. The models were assessed on English texts, and their performance was compared using their ROUGE points. Based on the results, it was decided to use and further research the model of hidden semantic analysis (LSA) [19].

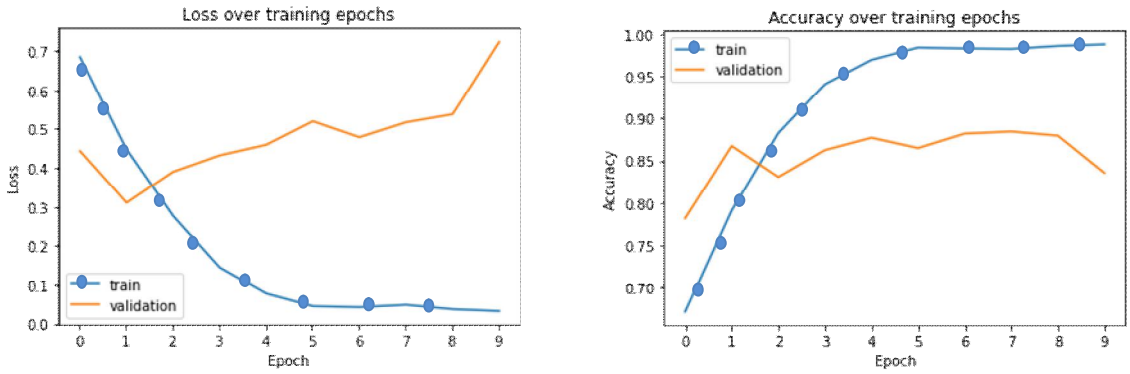


Fig. 5. Plots of error and accuracy of training for the XLM model

Table 1 – **Values of the accuracy of fine-tuned multilingual Transformer-based models**

Model		Precision	Recall	F1-score
BERT	no similar	0,78	0,86	0,82
	similar	0,94	0,89	0,92
	accuracy			0,88
DistilBERT	no similar	0,82	0,67	0,74
	similar	0,87	0,94	0,90
	accuracy			0,86
XLM	no similar	0,84	0,69	0,76
	similar	0,87	0,94	0,91
	accuracy			0,86

Solution to the problem of determining the semantic similarity of multilingual arbitrary-length text content. Strategy 2

The problem of determining the semantic similarity of multilingual arbitrary-length text content within strategy 2 is considered as a flow of: determining the representation vector of the text content (contextual word embeddings) using a sliding window approach by the "medium core" rule and a model based on Transformer architecture; calculating the degree of similarity of sentence embeddings using the selected metric (e.g., similarity cosine). In accordance with the tasks formulated, the research scheme of strategy 2 includes the following main steps:

1) Defining the architecture of the software module for determining the semantic similarity of multilingual arbitrary-length text content within strategy 2;

2) An analysis of the possibilities of various types of multilingual Transformer-based models (BERT, DistilBERT and XLM) to determine the semantic similarity of multilingual text content;

3) Suggesting a method of determining the vector representation of an arbitrary-length text content (contextual word embeddings) using the procedure of a sliding window by the "medium core" rule and the subsequent formation of sentence embeddings;

4) An analysis of the results delivered by the software module within strategy 2 as well as the selection of the most appropriate Transformer for solving the problem of determining the semantic similarity of multilingual arbitrary-length text content in the course of strategy 2.

The software module for determining the semantic similarity of multilingual arbitrary-length text contents within strategy 2 is presented schematically in Fig. 6. It includes four major blocks:

1) block 1 is intended to receive intersecting fragments of tokenized sequences of texts over 512 tokens long complying with "the medium core" rule;

2) block 2 forms contextual word embeddings for each fragment obtained at the previous stage, using the multilingual Transformers;

3) block 3 forms resulting sentence embeddings based on the word embeddings obtained for the entire sequence using the averaging or maximization operations;

4) block 4 implements the process of determining the degree of similarity of content by measuring the distances between their sentence embeddings.

All the multilingual Transformer-based models (BERT, DistilBERT, and XLM) reviewed in this paper were pre-trained through combining a modeling target using a masked language (MLM), which in turn determines their efficiency in feature extraction in the context of determining the semantic similarity of multilingual arbitrary-length text content in strategy 2.

The method of determining the vector representation of an arbitrary-length text content is based on the sliding window technique following "the medium core" rule, and it is accomplished through the following steps (blocks 1-3) (Fig. 7):

1. Block 1

1.1. After tokenization, a list of main fragments of 510 tokens long is formed with the remaining fragment of arbitrary length $ABF = \{BF_i\}, i = \underline{1,k}$ (the special ending and beginning characters are added for each sequence). Then the whole sequence can be presented as

$$WST = \sum_{i=1}^{k} BF_i .$$

1.2. A list is made up, which includes the medium fragments of token sequences (the “medium core”) $AMF = \{MF_j\}, j = \underline{1,k-1}$ formed at the joint of the left and right main fragments. The medium core is made of a half of right fragment and a tail of left fragment having the length same as those of the part of the right one used, the two additional special characters being added to mark the end and the beginning of each sequence. The length of the middle fragment is, as a result, equal to the length of the right main fragment

$$length(MF_j) = \frac{length(BF_{i+1})}{2} .$$

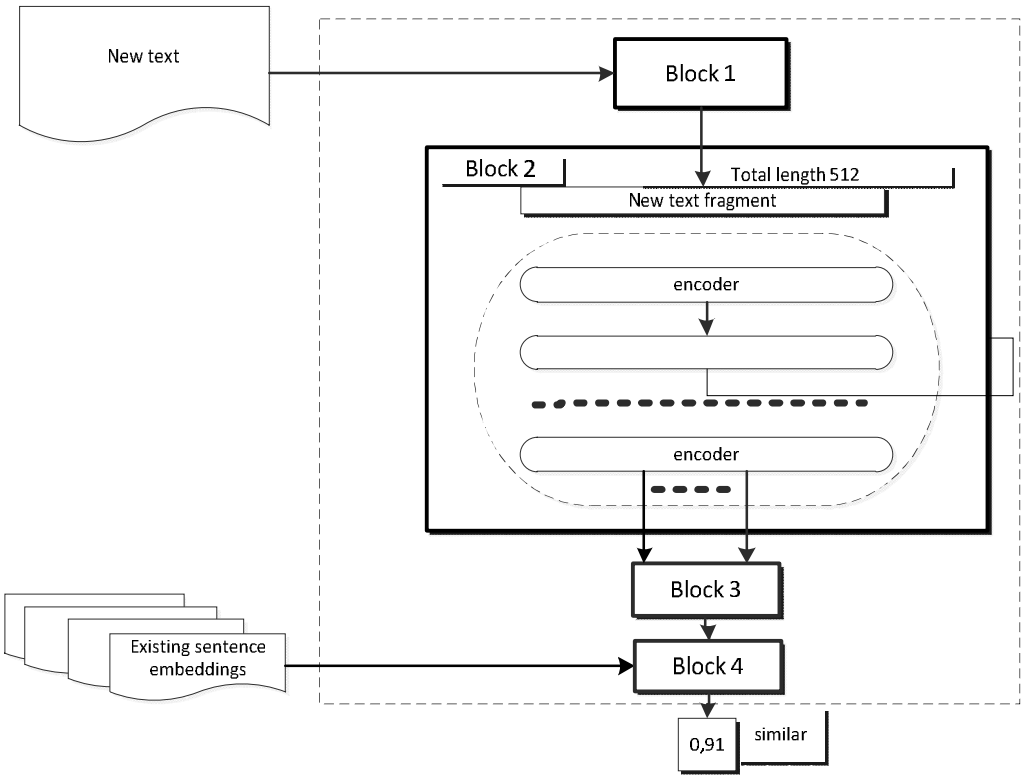


Fig. 6. Overall design of the module for determining the semantic similarity of multilingual arbitrary-length text content within Strategy 2

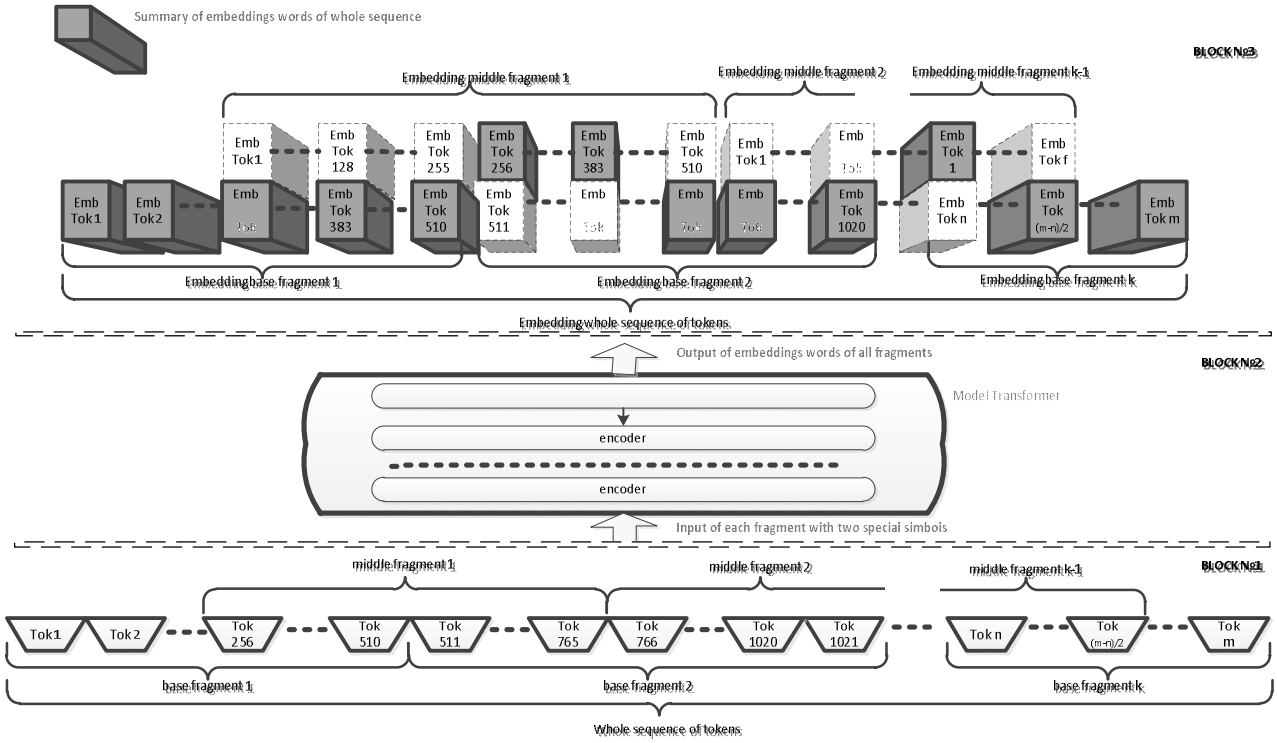


Fig. 7. High-level presentation of the method for determining the vector representation of text content of arbitrary length based on a sliding window according to the "medium core" rule

2. Block 2

2.1. The contextual word embeddings of the main and medium fragments are formed from the output of the multilingual Transformer-based model.

Corresponding lists are made up which can be formalized as follows:

$$AEBF = \{EBF_i\} \to \{BF_i\}, i = \underline{1,k,} \quad (2)$$

$$AEMF = \{EMF_j\} \to \{MF_j\}, j = \underline{1,k-1}. \quad (3)$$

3. Block 3

3.1. The contextual word embeddings of the main fragments are joined together, and the basic word

embeddings of the entire sequence are formed as $WES = \sum_{i=1}^{k} EBF_i$.

3.2. Word embeddings of medium fragments are half-truncated from the left and the right sides (to maximize the preservation of contextual dependence within the main sequence ("truncated medium core")) and the corresponding fragments in the main sequence are replaced with the resulting word embeddings

$$\left(EBF_i + EBF_{i+1}\right) = \left(EMF_j / 2\right),$$

$$i = \underline{1,k}, \quad j = \underline{1,k-1}.$$

3.3. The resulting sentence embeddings are formed according to the word embeddings of the entire sequence using the averaging or maximization strategy. In this research, Cosine Similarity was used as the main metric for determining the degree of similarity to determine the similarity of sentence embeddings of two text content items, after which the results of determining Cosine Similarity were converted into angular distance [20]

$$\left(u,v\right) = \left(1 - arccos\left(uv/\left(u+v\right)\right)/p\right), \qquad (4)$$

where u,v are sentence embeddings obtained from the results of the multilingual Transformers model.

Moreover, in accordance with the recommendations given in [1, 14], a relative, rather than an absolute assessment of the results of Cosine Similarity determination is performed

$$IF\ \left(u,v\right) > \left(u,c\right) THEN\ uis\ more\ v\ THAN\ c\ . \qquad (5)$$

The experiment was carried out using pre-trained multilingual BERT, DistilBERT and XLM. When determining the sentence vector, the average value was used (operation MEAN). The experiments have shown that using the MAX operation greatly raises the lower bound of the estimate. The extreme values of the estimate for the cosine similarity (expression (4)) and the Manhattan distance (considered as an additional metric for determining the degree of similarity) practically coincided; In this case, the relative similarity score (5) was taken into account.

Table 2 shows the average results of using the sliding window method according to the "medium core" rule in comparison with the truncation method (selecting the first 512 token fragments of a sequence) and the hierarchical method (combining word embeddings of all sequence fragments) with the following comparison using the cosine similarity (expression (4)) and Manhattan distance.

At the same time, the results of using mean pooling and max pooling were also analyzed for each of the methods for forming embedding of the entire sequence longer than 512 tokens.

As the experiment condition, a sequence of 512 tokens was used as a reference, and sequences of different lengths no more than 50% of the tokens from the reference were used as fragments.

The analysis of the averaged results (Table 2) proves the effectiveness of the method described in Section 3.2 in comparison with both the truncation and hierarchical methods. Also, it shows that an increase in the fragment length gives an increase in the accuracy of the sliding "medium core" window method. Therefore, these results can be extrapolated to processing sequences longer than 512 tokens.

In future, it is necessary to conduct additional research to determine the optimal size of truncated word embeddings of medium fragments for more efficient use of the sliding window method according to the "medium core" rule for processing sequences longer than 512 tokens.

Table 2 – **The results of using word processing methods for the input sequence longer than the maximum allowed for BERT**

Processing Method	Cosine similarity		Manhattan distance	
	mean pooling	max pooling	mean pooling	max pooling
Sliding window method with the "medium core" rule	87,77	97,96	108,73	193,61
Hierarchical method	87,4	97,91	109,01	194,09
Truncation method	84,45	97,14	124,62	251,59

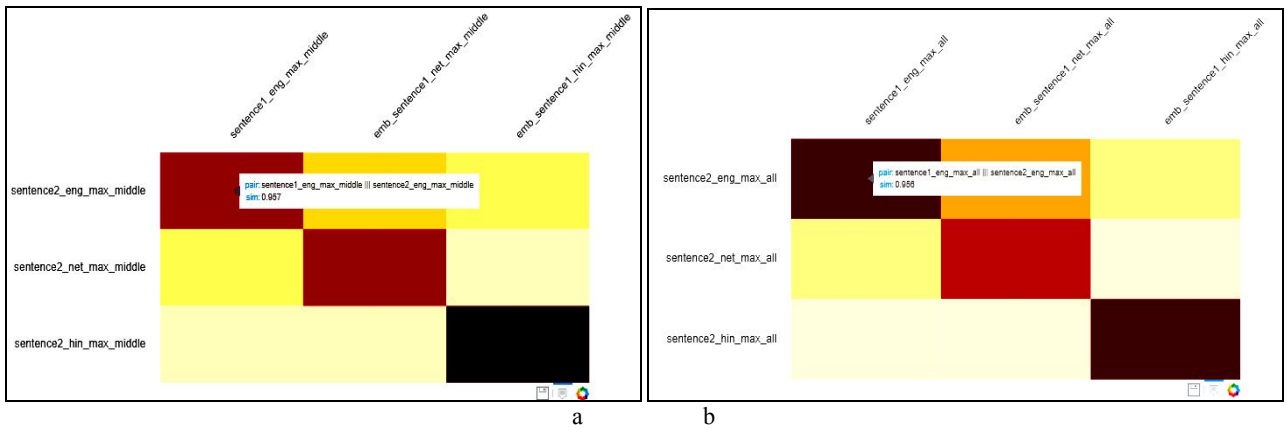
Moreover, with the maximum sequence length being just slightly over the maximum allowed for the Transformer-based model, the results of using the sliding window method and the hierarchical method for sequences up to 1024 tokens in length are almost the same. While the longer the sequence (e.g. more than twice the maximum input sequence of the model), the more effective the sliding window method compared to the hierarchical one.

Diagrams 1 and 2 of Fig. 8 show the results of Cosine Similarity for evaluating BERT outputs when solving problems of determining cross-language semantic similarity using the sliding window method according to the "medium core" rule and the hierarchical method respectively (formation of embedding sequences using the maximization method). The data set included pairs of two semantically similar sequences, but in different languages (English, Dutch, Hindi). The maximum length was 895 tokens for the sequence in Hindi, the minimum - 550 tokens in English.

In Fig. 9–12 there are the examples of scatter diagrams and dimension values of vector elements in solving problems of determining cross-language semantic similarity for semantically equivalent pairs of sentences ("English" - "English", "English" - "Dutch") for the sliding window method (Fig. 9, 10) and the hierarchical method (Fig. 11, 12).

An analysis of the experimental results showed that in the absence of hardware limitations to solve the problem of determining the semantic similarity of multilingual text content of arbitrary length using the second approach, the most effective is the use of "heavy" multilingual models such as XLM, as well as BERT.

If there are restrictions, the use of the multilanguage DistilBERT model will be most optimal.



a b

Fig. 8. The results of the Cosine Similarity

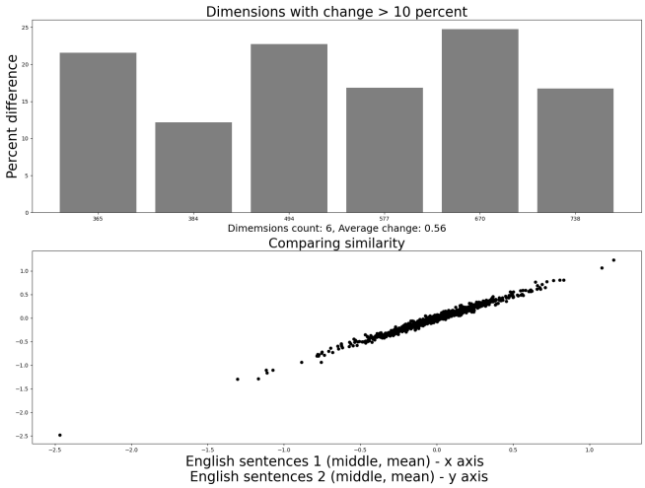


Fig. 9. Scatter diagrams and dimension values for equivalent pairs of sentences "English" - "English" for the sliding window method

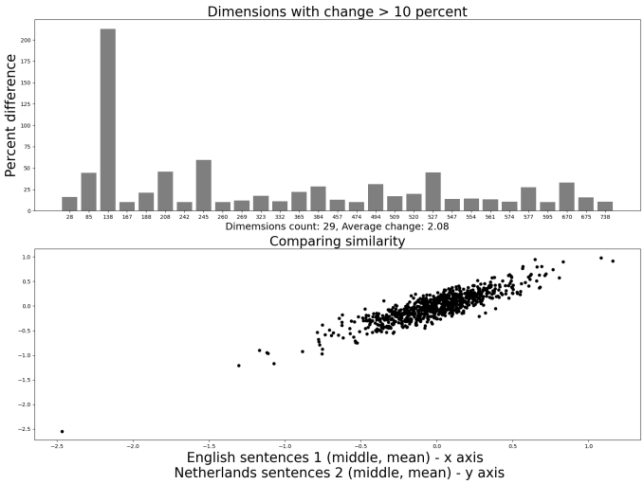


Fig. 10. Scatter diagrams and dimension values or equivalent pairs of sentences "English" - "Dutch" for the sliding window method

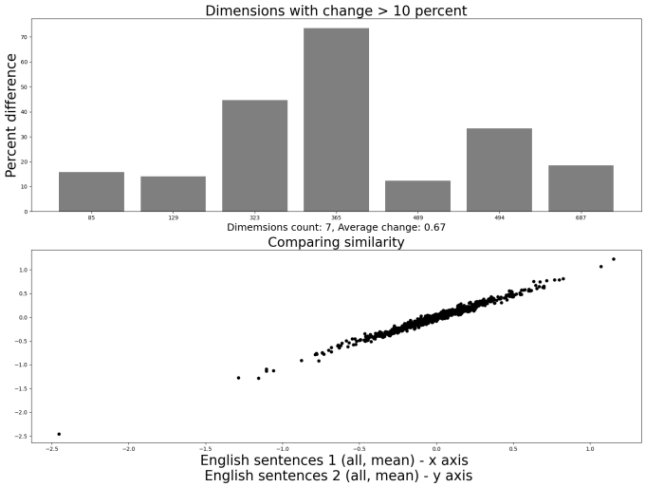


Fig. 11. Scatter diagrams and dimension values for equivalent pairs of sentences “English” - “English” for the hierarchical method

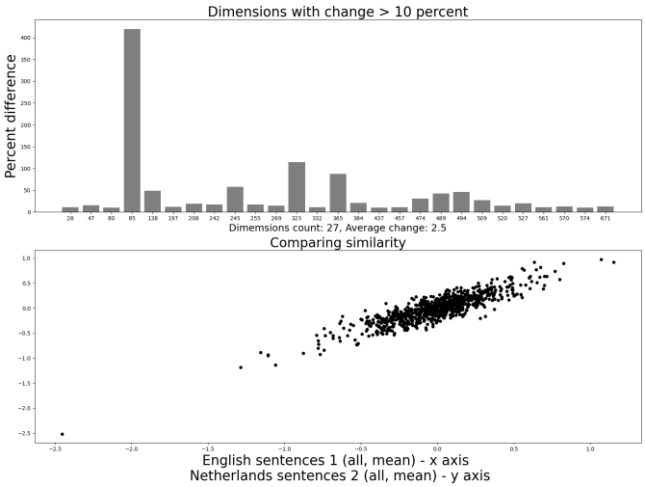


Fig. 12. Scatter diagrams and dimension values for equivalent pairs of sentences “English” - “Dutch” for the hierarchical method

Conclusions

This research has been carried out as a part of the HIPSTO Open AI Information Discovery Platform. The research shows the ways of using multilingual models based on Transformer architecture (BERT, DistilBERT, XLM) to determine the semantic similarity of text content of variable length.

There are two new approaches proposed to overcome the limitation on 512 input tokens for the considered Transformer-based models:strategy 1 is based on attacking the problem of determining the semantic

similarity of multilingual text content of variable length as a task of classification of pairs of text sequences using Transformer-based models (currently being validated in the HIPSTO AI technology setup); strategy 2 solves the problem as the task of getting a vector representation of variable length based on a sliding window using models based on Transformer architecture and followed by the determination of the degree of similarity of sentence embeddings (currently implemented within the HIPSTO AI technology setup).

The key point of strategy 1 implementation is the preliminary processing of text content longer than 512 tokens using an intelligent summarization and the subsequent application of the fine-tuned multilingual Transformer-based model for classifying pairs of text sequences.

Our research has demonstrated that BERT is the best option to use for this approach in case there are no strict computation power limitations. However, if the limitations are in place, the use of the multilanguage DistilBERT will be most optimal and acceptable in terms of effectiveness.

The main advantage of strategy 1 is the ability to overcome the restrictions on the length of the input sequence when determining the semantic similarity of text content in combination with all the advantages of the fine-tuned Transformer-based model pre-trained for Next Sentence Prediction (NSP).

The key feature of strategy 2 implementation is the initial processing of text content longer than 512 tokens with a sliding window based on the middle core rule which ensures the maximum preservation of the contextual dependence of the elements of the entire sequence. This step is followed by pushing the sliding window output through the Transformer-based multilingual model to form sentence embeddings.

The experiments have shown that both XLM and BERT are good options to use for this approach. However, if the limitations are in place, the multilanguage DistilBERT will still work as an acceptably effective substitution.

The advantage of strategy 2 is the ability to overcome the restrictions on the input sequence length in combination with all the advantages of the most powerful Transformer-based models in the generation of contextual word embeddings and the corresponding sentence embeddings (for example, such as the XLM model).

The continuation of our research is planned in:

1) an analysis of the possibilities of using PAWS (Paraphrase Adversaries from Word Scrambling) dataset in English and its extension PAWS-X (up to six typologically different types of languages: French, Spanish, Russian, Chinese, Japanese and Korean) for fine-tuning the multilingual BERT to determine the semantic similarity of text content in the course of strategy 1;

2) an analysis of the possibilities of using more complex extractive and abstract models of generalization and summarization of texts, including those based on deep neural networks.

Thus, this paper presents the results of the research of two effective and unique strategies of determining the semantic similarity of text content of arbitrary length using multilingual Transformer-based models. Strategy 2, which is based on the use of the Transformer-based base model and a sliding window according to the "medium core" rule, is currently implemented in the HIPSTO AI technology setup. Strategy 1, which is based on pre-summarization of text and the use of a fine-tuned Transformer-based model, seems more promising for further research and use. However, it is more resource "expensive" in terms of implementation, as it requires more extensive datasets for Transformer fine-tuning.

The HIPSTO AI technology setup is growing its dataset organically using the implementation of strategy 2 in production. In addition, the research and development of a neural network-based summarizer (which can also be trained with the datasets already collected by HIPSTO) should improve the accuracy and quality of strategy 1 implementation outcome.

References

1. Olizarenko, S. and Argunov, V. (2019), *Research into the possibilities of the multilingual BERT model for determining semantic similarities of text content*, available at: https://hipsto.global/BERT-Application-Research-for-HIPSTO-Related-News-Detection.pdf
2. Devlin, J., Ming-Wei Chang, Lee, Ke.and Toutanova, K. (2019), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805v2 [cs.CL] 24 May 2019.
3. Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2020), *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*, arXiv:1910.01108v4 [cs.CL] 1 Mar 2020.
4. Guillaume, Lample and Alexis, Conneau (2019), *Cross-lingual Language Model Pretraining*, arXiv:1901.07291v1 [cs.CL] 22 Jan 2019.
5. Sun, C., Qiu, X., Xu, Y. and Huang X. (2020), *How to Fine-Tune BERT for Text Classification*, arXiv:1905.05583v3 [cs.CL] 5 Feb 2020.
6. Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y., Strope, B. and Kurzweil R. (2019), *Multilingual Universal Sentence Encoder for Semantic Retrieval*, arXiv:1907.04307v1 [cs.CL] 9 Jul 2019.
7. Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.St., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B. and Kurzweil, R. (2018), "Universal sentence encoder for English", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169–174.
8. Yoon, Kim (2014), "Convolutional neural networks for sentence classification", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751.
9. Ashish, Vaswani, Noam, Shazeer, Niki, Parmar, Jakob, Uszkoreit, Llion, Jones, Aidan, Gomez, Łukasz, Kaiser, and Illia, Polosukhin (2017), "Attention is all you need", *Proceedings of NIPS*, pp. 6000–6010.

10. (2020), *Multilingual Similarity Search Using Pretrained Bidirectional LSTM Encoder. Evaluating LASER (Language-Agnostic SEntence Representations)*, available at: https://medium.com/the-artificial-impostor/multilingual-similarity-search-using-pretrained-bidirectional-lstm-encoder-e34fac5958b0.
11. (2019), *Zero-shot transfer across 93 languages: Open-sourcing enhanced LASER library*, POSTED ON JAN 22, 2019 TO AI RESEARCH, available at: https://engineering.fb.com/ai-research/laser-multilingual-sentence-embeddings.
12. Reimers, N. and Gurevych I. (2019), *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, arXiv:1908.10084v1 [cs.CL] 27 Aug 2019.
13. Patel, M. (2019), *TinySearch - Semantics-based Search Engine using Bert Embeddings*, available at: https://arxiv.org/ftp/arxiv/papers/1908/1908.02451.pdf.
14. Han, X. (2020), *Bert-as-service*, available at: https://github.com/hanxiao/bert-as-service.
15. (2020), *State of the art Natural Language Processing for Pytorch and TensorFlow 2.0*, available at: https://huggingface.co/transformers/index.html.
16. Arun, S. Maiya (2020), *Ktrain: A Low-Code Library for Augmented Machine Learning*, available at: https://arxiv.org/pdf/2004.10703v2.pdf.
17. Dolan, B. and Brockett, C. (2005), "Automatically Constructing a Corpus of Sentential Paraphrases*", Proceedings of the 3rd International Workshop on Paraphrasing* (IWP 2005), Jeju Island, pp. 9–16.
18. Goodfellow, I., Bengio, Y. and Courville, A. (2018), *Softmax Units for Multinoulli Output Distributions. Deep Learning*, MIT Press. pp. 180–184, ISBN 978-0-26203561-3.
19. Markovsky, I. (2012), *Low-Rank Approximation: Algorithms, Implementation, Applications,* Springer, ISBN 978-1-4471-2226-5.
20. Daniel, Cer, Yinfei, Yang, Sheng-yi, Kong, Nan Hua, Nicole, Limtiaco, Rhomni, St. John, Noah, Constant, Mario, Guajardo-Cespedes, Steve, Yuan, Chris, Tar, Brian, Strope, and Ray, Kurzweil (2018), "Universal sentence encoder for English", *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169–174.

Received (надійшла) 12.05.2020
Accepted for publication (прийнята до друку) 24.06.2020

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Олізаренко Сергій Анатолійович – доктор технічних наук, старший науковий співробітник, професор кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Serhii Olizarenko – Doctor of Technical Sciences, senior researcher, professor of the electronic computers department, Kharkiv National University of Radio Electronics University, Kharkiv, Ukraine;
e-mail: sergejolizarenko5@gmail.com; ORCID ID: https://orcid.org/0000-0002-7762-6541.

Аргунов Володимир Володимирович – керівник відділу досліджень штучного інтелекту, HIPSTO, Харків, Україна;
Vladimir Argunov – Head of AI Research, HIPSTO, Kharkiv, Ukraine;
e-mail: warg.silencer@gmail.com; ORCID ID: https://orcid.org/0000-0002-2505-1969.

Дослідження особливостей визначення семантичної подібності текстового контенту довільної довжини з використанням багатомовних моделей на основі Transformer

С. А. Олізаренко, В. В. Аргунов

Анотація. В роботі досліджені можливості визначення семантичноъ подібності багатомовного текстового контенту довільної довжини на основі їх векторних уявлень, отриманих з використанням різних багатомовних моделей на основі архітектури Transformer. Проведено порівняльний аналіз моделей Transformer для вибору найбільш ефективної моделі для вирішення даного класу задач. Запропоновано два нових унікальних підходи до визначення семантичної подібності багатомовного текстового контенту для використання в платформі HIPSTO Open AI Information Discovery з подоланням проблеми використання тексту довільної довжини. Аналізуються експериментальні дані, отримані при реалізації нових підходів для вирішення завдання семантичного подібності текстового контенту довільної довжини.

Ключові слова: обробка природної мови; BERT; семантична подібність; новинний конвент; глибоке навчання; багатомовний текстовий конвент; векторне подання; трансформери; тонке налаштування.

Исследование особенностей определения семантического сходства текстового контента произвольной длины с использованием многоязычных моделей на основе Transformer

С. А. Олизаренко, В. В. Аргунов

Аннотация. В работе исследованы возможности определения семантического сходства многоязычного текстового контента произвольной длины на основе их векторных представлений, полученных с использованием различных многоязычных моделей на основе архитектуры Transformer. Проведен сравнительный анализ моделей Transformer для выбора наиболее эффективной модели для решения данного класса задач. Предложены два новых уникальных подхода к определению семантического сходства многоязычного текстового контента для использования в платформе HIPSTO Open AI Information Discovery с преодолением проблемы использования текста произвольной длины. Анализируются экспериментальные данные, полученные при реализации новых подходов для решения задачи семантического сходства текстового контента произвольной длины.

Ключевые слова: обработка естественного языка; BERT; семантическое сходство; новостной контент; глубокое обучение; многоязычный текстовый контент; векторная подача, трансформеры; тонкая настройка.