On the Usage of Semantic Text-Similarity Metrics for Natural Language Processing in Russian

Mikhail Koroteev
Financial University under the government of the Russian Federation
Moscow, Russia
mvkoroteev@fa.ru

***Abstract*—This work is devoted to the analysis of the applicability of text processing methods in natural languages to the problem of assessing the similarity of text sequences of arbitrary length in Russian. This task is a particular problem that finds its application in many areas of computational linguistics, primarily in the analysis of the effectiveness of the computer-generated text. The analysis revealed the advantage of using the BERTScore method, based on the semantic similarity of text attachments, especially in terms of the reliability and sequence of detecting semantic similarity of sentences compared to traditional text metrics based on character-by-character analysis of texts.**

Keywords—computational linguistics, natural language analysis, semantic analysis, text attachments, machine learning.

I. INTRODUCTION

Central event of 2019 in the field of natural language processing was the introduction of a new pretrained model of text attachments BERT, which allows to achieve unprecedented results of accuracy in many automatic word processing problems. This model is likely to replace the widely known word2vec model in prevalence, becoming, in fact, the industry standard. Throughout 2019, almost all scientific articles devoted to the problem of word processing in natural languages, in one way or another, were a reaction to the release of this new model, the authors of which have become one of the most cited researchers in the field of machine learning.

Natural language processing tasks include a wide range of applications from conversational bots and machine translation to voice assistants and online speech translation. Over the past few years, this industry has experienced rapid growth, both quantitatively, in the volume of market applications and products, and qualitatively, in the effectiveness of the latest models and the proximity to the human level of language understanding.

Text Generation Quality Assessment The problem of text generation quality assessment arises in the process of solving many problems, for example, machine translation. It can be reduced to comparing a set of candidate proposals with a sample proposal in a given textual context. However, the most commonly used sentence similarity metrics, such as the one described in the 2002 BLEU (bilingual evaluation understudy), focus only on superficial similarity. The aforementioned BLEU metric, the most common in the development of machine translation systems, relies on the comparison of the intersection of n-grams of text. For all its simplicity, such metrics miss the lexical and semantic diversity of natural languages.

A vivid example of this problem, described in 2005 by B. Sataneev in [4]: a number of popular text metrics for this reference sentence “People love foreign cars” prefers the sentence “People like to travel abroad” in comparison with those that are obviously more semantically close to the original “ Customers prefer foreign cars ”. As a consequence, machine translation systems that use such metrics to assess translation quality will prefer syntactically and lexically similar constructs, which is suboptimal in the context of wide linguistic diversity.

The representation of the BERT system [5] allows it to be used as a basis for measuring the similarity of sentences in natural languages using the metric of the distance between text attachments of compared sentences.

In general, the text metric, or the metric of the quality of text generation, is a function , where is the vectorized representation of the sample proposal, and is the vectorized representation of the candidate proposal. A good metric should reflect the person's judgment as closely as possible, that is, show a high correlation with the person's assessment results. All existing metrics can be roughly classified into one of four categories: n-gram match, edit distance, attachment comparison, and trained functions.

A good metric should reflect the person's judgment as closely as possible, that is, show a high correlation with the person's assessment results. All existing metrics can be roughly classified into one of four categories: n-gram match, edit distance, attachment comparison, and trained functions.

The most common textual metrics are based on counting the number of n-grams found in both sentences. The larger the dimension n in an n-gram, the more the metric is able to capture the similarity of whole words and their order, but at the same time, the more this metric is limited and tied to a specific formulation and word forms. The already mentioned BLUE and MeTEOR belong to this category of metrics.

There are methods that calculate the proximity of proposals by the number of edits that translate a candidate into a reference. Methods such as TER and ITER take into account

978-1-7281-1094-3/20/$31.00 ©2020 IEEE

semantic proximity of words and normalization of grammatical forms. This also includes methods like PER and CDER, which take into account the permutation of text blocks. Some more modern methods (CharacTER, EED) have unexpectedly better results.

In recent years, metrics have begun to appear based on the use of dictionary embeddings, that is, trained dense word vectorizations (MEAN, YISI-1). The advantage of using BERT as a basis for constructing such metrics is that it takes into account the context of a word within its environment, which makes it possible to use attachments not at the level of individual words, but at the level of the entire sentence.

Since the criterion for the quality of a text metric is its correlation with human judgments, it is not surprising that machine learning is used to build trained metrics, where the target function is just such a coincidence. For example, BLEND uses a regression model to weight 29 known metrics. The disadvantage of this approach is the dependence on the presence of a corpus of pre-labeled data for the training set, as well as the risk of overfitting in a certain subject area and, as a consequence, the loss of generalizing ability and universality.

In February 2020, researchers from Cornwell University proposed a special mechanism for evaluating the effectiveness of text models based on BERT - BERTScore [2]. BERTScore is used to automatically assess the quality of natural language text generation. The authors of the proposal argue that BERTScore correlates better with human judgments about the quality of the text and, as a result, can form the basis for a more efficient and effective model selection process. Today it is the most progressive metric for evaluating the quality of text generation.

In a comparative analysis of text similarity metrics, metrics based on BERT show consistently higher results than classic text metrics. This means that they are statistically significantly closer to human estimates.

In addition, the original article introducing BERTScore focuses on performance issues. In this regard, BERTScore is, of course, often slower than classic models. The authors give an assessment of the comparison with the popular implementation of SacreBLEU, in which BERTScore is about three times slower. Since the pretrained BERT model is used to assess the similarity, the increase in the accuracy of the estimate is given at the expense of a decrease in speed that would not be expected from such a more complex model. Given the typical size of natural text processing datasets, the increase in computation time during model validation should not have a significant effect on the performance of the machine learning system as a whole.

Publications already beginning to appear [3] are that improve the original BERTScore algorithm by parallelizing computational operations. Variations of classic metrics using BERT are presented, showing improved correlation with human scores.

Undoubtedly, the direction of the future development of textual metrics will be the wider use of BERT as a basis for assessment, a kind of semantic engine. Also promising is the development of specific models that take into account the peculiarities of specific subject areas and increase the basic level of accuracy through specialization. With regard to BERTScore, another advantage is its differentiability, which will allow it to be integrated into the methodology for training text models in the future, which promises to further increase the performance and quality of machine learning models of natural text processing.

II. APPLICABILITY OF THE BERTSCORE FOR THE ANALYSIS OF RUSSIAN TEXTS

As you know, the analysis of natural texts has a strong language specificity. Within the framework of the study of the applicability of modern methods of machine learning for the problems of a controlled generation of natural texts, carried out by the authors and colleagues, it is necessary to solve the problem of assessing the efficiency of text generation. This task involves the use of the proximity metric of text sequences to assess the efficiency of generation systems.

Such a system, given the recent advances described in the previous chapters of this work, should be based on an assessment of semantic similarity, rather than mechanical analysis at the level of symbols or words.

The initial stage of the study involves assessing the applicability of modern methods, in particular, BERTScore, on a dataset of Russian-language text sequences. For this, we chose the NLPDatasets corpus of texts (https://github.com/Koziev/NLP_Datasets), namely the permutation paraphrasing dataset. These data will allow determining the applicability of the semantic distance metrics of text sequences to determine the similarity of sentences in meaning. This dataset is convenient in that it is divided into groups of sentences, each of which has an identical meaning, but grammatically and orderly they have a significant difference.

We used the implementation of the BERTScore method from the authors of the methodology in the form of a package for the Python programming language called bert_score (https://pypi.org/project/bert-score/0.1.1/).

The methodology for determining the integral quality indicator of the semantic similarity metric is as follows:

One sentence is selected from the original dataset and compared with all the others. At the same time, the metric that takes into account precisely the semantic similarity should show a high degree of closeness of the reference sentence with its paired one from the original set and a low one with all the others.

For a given proposal, a quality metric is constructed according to the following formula:, where is the assessment of the metric F1 of the similarity of the referenced proposal with its paired one, is the arithmetic mean of the metrics F1 of the similarity of this proposal with all the others. This indicator will approach 1 in the case of an ideal classification and zero in the absence of discrimination of the referenced proposal for the entire sample.

This procedure is repeated for all reference sentences from the initial dataset. The obtained indicators are averaged to

assess the average efficiency of this method for determining the

The above quality assessment is performed for three metrics of the distance between text sequences - BERTScore, normalized Levenshtein distance, and the SequenceMatcher metric expressed in fractions of one. The last two metrics are standard ways to programmatically determine the similarity of two text sequences of arbitrary length.

To reduce program execution time, which is most relevant for the BERTScore metric, which uses a deep neural network BERT to represent text embeddings, we used 20 sentences from the initial dataset in each test case for comparison with each reference sentence so that these 20 examples must be included in itself paired with the reference. Thus, 500 comparisons of different reference sentences for the BERTScore model and 11,000 for the character models were carried out. The comparison results can be seen in Table 1.

TABLE I. RESULTS OF COMPARING THE EFFECTIVENESS OF DIFFERENT METHODS FOR ASSESSING THE DISTANCE BETWEEN TEXT SEQUENCES

	BERT	***Levenstein***	***SM***
Average	0.24995	0,01146	0,37585
Variance	0.00082	0,00007	0,01320
Minimum	0.18985	-0,00858	0,10730
Maximum	0.33478	0,03738	0,68553
time for 100 examples, sec	0.56182	0,21429	

It should be noted that a model based on the pre-trained neural network BERT Multilingual was used to analyze texts in Russian. Additional training on the corpus of Russian texts was not carried out, since the purpose of the study was primarily to assess the native performance of this model, without increasing its performance through specialized additional training on the corpus of texts from the target subject area.

The most indicative is the indicator of the average deviation of the distance of the reference sentence and its paraphrase and the average distance of the reference sentence from 20 sentences from the initial dataset. In this case, the ability of the distance metric to determine the paraphrase of a sentence against the background of others that have nothing in common with it, both in meaning and in form, is compared. The higher this indicator, the more predictive power the metric has.

We also estimated the variance of this indicator on the array of evaluated reference proposals. This value evaluates the reliability and consistency of the metric, its independence from the form and meaning of the proposal. The lower this indicator, the more reliable the metric.

The minimum and maximum values of the deviation characterize the spread of values. Here, of interest are those cases when the minimum value was less than zero. This means that this metric, at least in one case, completely failed the task of determining the paraphrase of a sentence, that is, it estimated its distance from the reference one higher than from some other one that does not belong to it.

semantic similarity of text sequences.

Besides, for practical application of these methods, the comparison execution time, which characterizes the performance and speed of the metric, is of interest. Here it must be said that comparisons using the BERTScore method were made on the Google Colab cloud service, data processing was performed on the CPU. The calculation of metrics 2 and 3, that is, based on character-by-character analysis of sequences, was carried out on a computer with an AMD Ryzen 5 1600, 16GB DDR4 processor.

III. CONCLUSIONS

When analyzing the results of a numerical experiment, some significant conclusions can be drawn about the applicability of methods for assessing the distance between text sequences using grammatical and semantic approaches. First of all, it is necessary to note the significantly higher computational complexity of the semantic approach. As already mentioned, any models and methods for analyzing natural texts using the BERT model require significantly more serious computing power, preferably including the ability to calculate on a GPU or TPU.

As for the immediate efficiency, it can be seen that the metric based on the calculation of the Levenstein distance does not give the proper level of discrimination of paraphrases, which is evident from the significantly lower values of the mean deviation. Moreover, in some cases, this metric generally gives negative deviations, which indicates the possible issuance of lower distance values for sentences that are similar in meaning and different in form than for sentences that are similar in form and different in meaning.

As for method 3 (SequenceMatcher), on average it gives a clearer differentiation of the paraphrase of the reference sentence than even the semantic method BERTScore. However, in comparison with it, it shows a much higher variance of the results, which indicates a serious dependence on the reliability of this method on the form and grammatical similarity of the compared sentences. On the contrary, the BERTScore method, with a more conservative average result, has a very low variance, due to the analysis of just text attachments, and not the symbolic similarity of text sequences.

Immediately after its inception, the BERT model received an intense reaction from the scientific community and is now used in almost all word processing tasks. Almost immediately, the proposals for improving the model considered in this work appeared, which led to an improvement in the results of its application in all subsequent problems. With all that said, we can confidently assert that BERT represented a quantum leap in the field of intelligent natural language processing and consolidated the superiority of using pre-trained text representation models on huge data sets as a universal basis for building intelligent algorithms for solving specific problems.

Undoubtedly, we will see many more new scientific results based on the application and adaptation of the BERT model to various problems of word processing in natural languages. Further improvement of the neural network architecture, coupled with fine-tuning of the training procedure and parameters, will inevitably lead to a significant improvement in

many computer NLP algorithms, from text classification and annotation to machine translation and question-answer systems.

As for the current research, undoubtedly we need more empirical evidence for the applicability and restrictions of the semantic text embeddings for the Russian language texts. Investigating different datasets, generated corpora, including transfer learning on the problem-specific data may help to enhance performance dramatically.

ACKNOWLEDGMENT

The author is grateful to the staff of the Department of Data Analysis and Machine Learning of the Financial University under the Government of the Russian Federation for the consultative support of the study.

REFERENCES

[1] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. Retrieved March 20, 2020 from http://arxiv.org/abs/1907.11932

[2] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. Retrieved March 20, 2020 from http://arxiv.org/abs/1904.09675

[3] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). DOI: https: //doi.org/10.18653 / v1 / d19-1053

[4] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and / or Summarization, 65–72. Retrieved March 20, 2020 from https://www.aclweb.org/anthology/W05-0909.pdf

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Retrieved March 20, 2020 from http://arxiv.org/abs/1810.04805

[6] Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. Retrieved March 20, 2020 from http://arxiv.org/abs/1907.07355